Universidad del Norte
Systems Engineering Department
Doctorate in Systems Engineering

# Privacy Protection in Location Based Services

Mayra Zurbarán Nucci
Adviser: Pedro M. Wightman Rojas, Ph.D.

Thesis submitted to the Universidad del Norte
for the degree of Doctor of Philosophy

June 2018

# Abstract

This thesis takes a multidisciplinary approach to understanding the characteristics of Location Based Services (LBS) and the protection of location information in these transactions. This thesis reviews the state of the art and theoretical approaches in Regulations, Geographic Information Science, and Computer Science.

Motivated by the importance of location privacy in the current age of mobile devices, this thesis argues that failure to ensure privacy protection under this context is a violation to human rights and poses a detriment to the freedom of users as individuals. Since location information has unique characteristics, existing methods for protecting other type of information are not suitable for geographical transactions. This thesis demonstrates methods that safeguard location information in location based services and that enable geospatial analysis.

Through a taxonomy, the characteristics of LBS and privacy techniques are examined and contrasted. Moreover, mechanisms for privacy protection in LBS are presented and the resulting data is tested with different geospatial analysis tools to verify the possibility of conducting these analyses even with protected location information.

By discussing the results and conclusions of these studies, this thesis provides an agenda for the understanding of obfuscated geospatial data usability and the feasibility to implement the proposed mechanisms in privacy concerning LBS, as well as for releasing crowdsourced geographic information to third-parties.

# Acknowledgements

While this thesis is my own work, it would not have been possible without the help, guidance and support of my colleagues, peers and friends.

My adviser took a memorable role in mentoring me from the beginning, engaging me on this research path; during my very first steps up until now. The opportunities I have had while going through my PhD have enriched me both personally and professionally. It would not have been possible for me to conduct this research without the economic support from organisations like Universidad del Norte and Politecnico di Milano in cooperation with the Sustain-T project by Erasmus Mundus.

Emotional support I owe to many: quite dearly to my father, Fermín Zurbarán for teaching me tenacity and hard work throughout his life and the necessary skills to cope with life itself during his presence on earth; to Jota, the kindest person I have known and whose faith in me continues to push me forward; to my grandmother who's my biggest fan and let me know what unconditional love means; to my mom, for her patience and understanding; to Laika, for I could not acknowledge enough her efforts for accompanying me in numerous all-nighters and for showing me what loyalty really means. To the De Castro Aguilar family for becoming my own and their support in all possible ways. To Francis for showing me what it is like to have my feet on the ground.

Thanks to Rosita and friends made in Italy! Very dearly to Ariana, Eylül and Luca. To my co-advisor that I incidentally met in Frascati, Italy as well: Dr. Iliffe for his consistent advice on how to prioritise and get things done.

The Geolab team and my Italian Prof were key in showing me the other side of the coin; introducing me to GIS concepts and to the FOSS community. Without them, this work could not have been completed.

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

## 1.1 Introduction

During the past few decades, the market associated with mobile technology has grown at an impressive rate, becoming attractive to all actors involved: manufacturers, operators, governments, and research centres. The massive adoption of this technology, the improved computing power of new devices, commercialisation opportunities for traditional voice services, now allow high-level interaction between users and access to information on the Internet. This evolution of mobile communication has turned cell-phones into the essential way people communicate daily, and there is an increasing demand for new applications that suit their needs.

Location aware devices such as smartphones, use Global Navigation Satellite Systems (GNSS)[1] or technologies like WiFi or cell towers to calculate location information. These devices feature portability, connectivity, processing capability, and a personal use nature; favouring the demand for context aware services, specifically Location Based Services (LBS). The advantages that these services provide include tracking goods, enabling requests of points of interest (PoIs), and shortest path calculation even with multimodal transport, among many others.

LBS are part of Location Based Information Systems (LBIS). Labrador et al. (2010) defines it as *"applications that provide users with information based on their geographical position, which could be obtained from the mobile device they are accessing the service, or using a manually defined location"*. The origin of LBS dates to the introduction of the E911 (Enhanced 911) system in the United States in 1996, which required the mobile operators to locate the callers of the emergency line with prescribed accuracy according to Bellavista et al. (2008). In order for the LBS to provide the requested information properly, sensitive data about the subject's location is required. While this sensitive information is sent from the mobile device and stored

[1] Like Global Positioning System (GPS), Galileo, GLONASS, BeiDou, etc.

unprotected, it is in danger of being intercepted and misused by untrusted parties and even by the LBS provider itself.

Location privacy violation can lead to the identification of the victims and their physical surroundings. This breach enables attacks like: stalking, physical assaults, and targeted advertising, among others. Attacks are not limited to the use of real-time location information of an individual. Rather, historical records can be used to estimate a user's routine; revealing where a person is going to be based on past information. Additionally, location information can expose personal preferences, there are places where the mere recordings of someone's permanence there gives away sensitive information subject to moral discrimination, e.g. hospitals, churches, political centres, etc.

In some cases, governments may obtain telephone records and location information of persons involved in judicial acts; however, if this information is accessed unrestrained, it could be used improperly and put at risk the freedom of citizens. This concern is not oblivious to governments, which have included lawful protocols to handle the communications of location information as stated by the European GNSS Agency (2010). At the time of writing, some of the current related regulations are:

- In the EU, guidelines for public administrations on location privacy by the European Commission (2016), the Directive 2002/58/EC on privacy and electronic communications by the EU (2002), and Regulation 2016/679: General Data Protection Regulation (GDPR) by the EU (2016). These aim to provide european citizens with the right to e-privacy and specifically to location privacy protection;
- In Colombia, there is the law Ley 1581 by Congreso de Colombia (2012), that mandates entities that handle personal data to notify their users to what extent their data will be stored or manipulated; this does not address specific policies for location information;
- In the U.S the Location Privacy Act has been introduced some times in Congress, but not yet passed, the latest attempt being by Franken (2015).

Privacy is a complex subject. One of the first definitions to this term came earlier on the 19th century in the U.S by Warren and Brandeis (1890) and is quite simple: *"The right to be let alone"*. Brandeis argued that privacy was the most cherished of freedoms in a democracy, and was concerned that it should be reflected in the Constitution. The connotation of this has evolved along with communications and technology, making it a day-by-day challenge to maintain the balance within the developed technology and available privacy measures.

The Universal Declaration of Human Rights by the United Nations (1948), Article 12 refers to the right to privacy: *"No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honour and reputation. Everyone has the right to the protection of the law against such interference or attacks."*

In the U.K, The Human Rights Act of 1998 by the ECHR (1998) further states:

1. *"Everyone has the right to respect for his private and family life, his home and his correspondence."*
2. *"There shall be no interference by a public authority with the exercise of this right except such as is in accordance with the law and is necessary in a democratic society in the interests of national security, public safety or the economic well-being of the country, for the prevention of disorder or crime, for the protection of health or morals, or for the protection of the rights and freedoms of others."*

These explicit rights aim to preserve privacy, acknowledging the need for such right in a democratic society and for the best interest of the individual. It is for these reasons that a holistic approach for information handling in this era is critical to preserve the basic rights that everyone should be entitled to.

The need for identifying specifically location privacy led to its definition in Perilla et al. (2006) as: *"A special type of information privacy which concerns the claim of individuals to determine for themselves when, how, and to what extent location information about them is communicated to others"*. This still retains the ideology of Article 12 of the UN Declaration of Human Rights to avoid arbitrary inferences, but within the more limited frame of location information. This definition will be characteristic in the followings of this work, since many questions arise from it on the 'how-to' accomplish the aforementioned right to privacy under the location information frame.

The practical measures to assure this fundamental right are discussed in the first chapters of this work, reviewing many mechanisms to provide location privacy. Formally, these mechanisms will be addressed as Location Privacy Protection Mechanisms (LPPMs). The LPPMs available in the literature are very diverse in their application, this is explained by the broad features offered by LBS that demand tailored privacy measures. In spite of this, there is a lack of implementation of LPPMs in commercial applications.

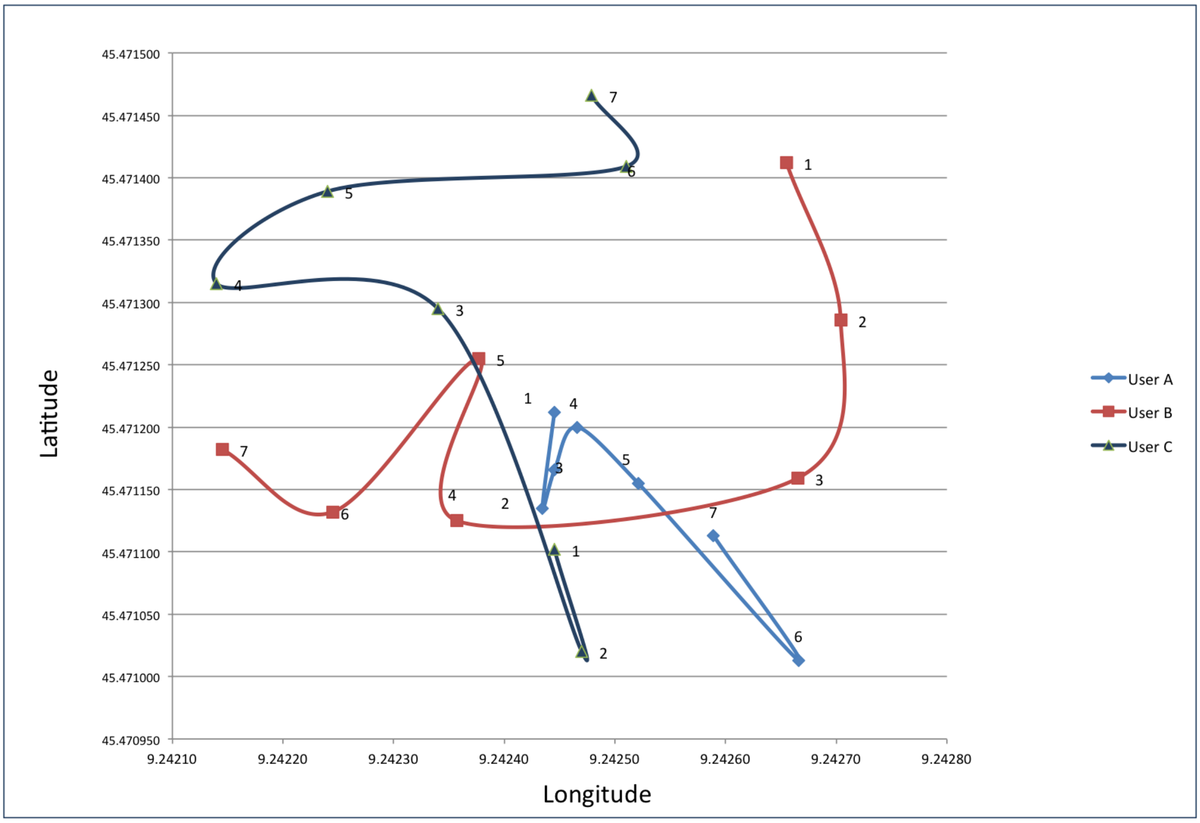
The surplus in LPPMs from the academic community evidences the need for researchers to work hand in hand with the industry, to develop solutions available to the public in the benefit of an individual's right to privacy, as well as to generate confidence in the service providers for introducing practices on how to handle user's information with privacy requirements. This would result in increased transparency and user participation. Xu and Gupta (2009) found that user's privacy concerns translate into fear of losing control of personal information, leading to cause stress and anxiety. With the adoption of privacy protective measures, the perception of privacy and user's trust on the providing companies of a service can increase, leading to more compliance to adopt an LBS.

To provide privacy, it is important to understand the components of the information that requires this protection. There are three aspects to location information: *identity, location* and *time*. If an adversary is able to link between them, location privacy is broken. These components form an instance of location information; a sequence of such instances gives historical location information, which allows to establish behaviour patterns and the possibility of identifying the user's home, work, and routines.

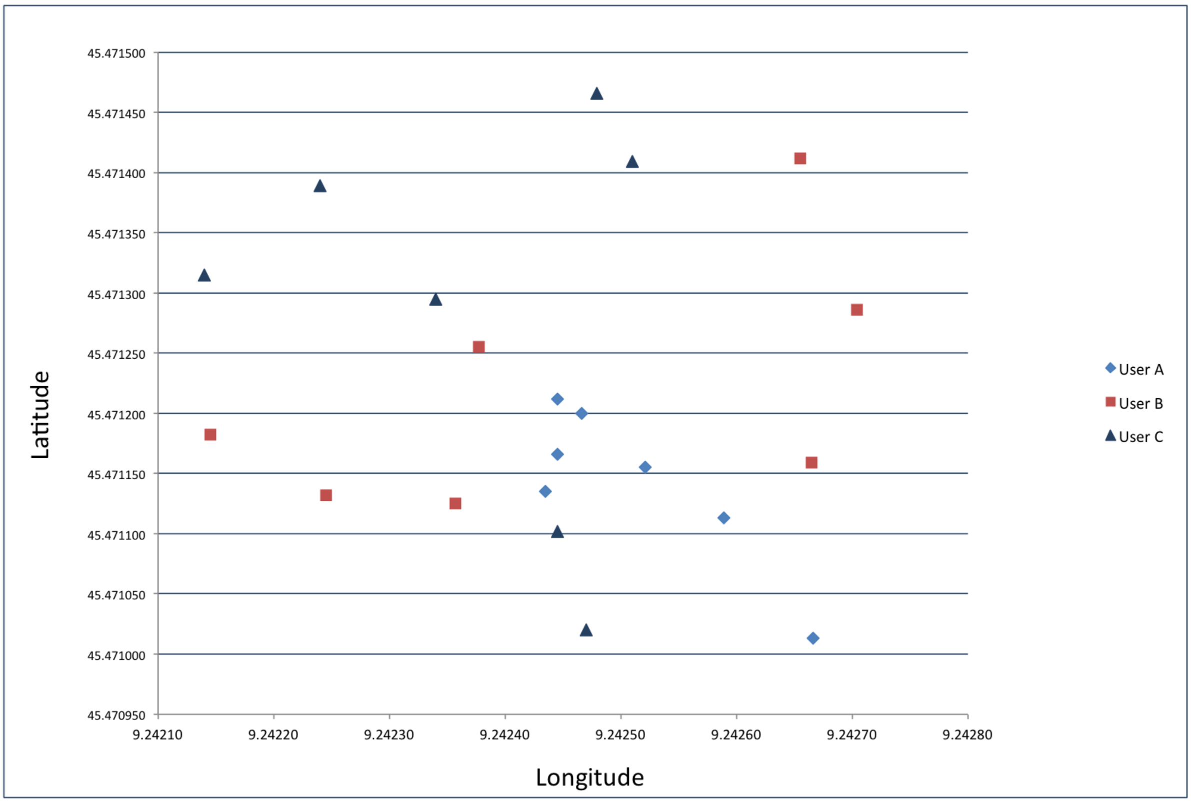
For this work, the Idem Identity or the Diachronic Meaning of Identity is preferred instead of Ipse Identity[2] to explain the identity component of location information. This is expressed by the following quote of Beller and Leerssen: *"Identity becomes to mean being identifiable, and is closely linked to the idea of permanence through time"* found in Rannenberg et al. (2009a). This notion is relevant in the way that not only the revelation of real world identifiers such as ID number or name is considered a break in anonymity, but also pseudonymous if correctly linked can lead to the identification of an individual on a dataset.

There are diverse approaches that have been proposed to satisfy location privacy requirements; some of them are designed to protect user's identity while issuing queries, others focus on protecting the user's location and some even offer protocols to obfuscate query predicates as well.

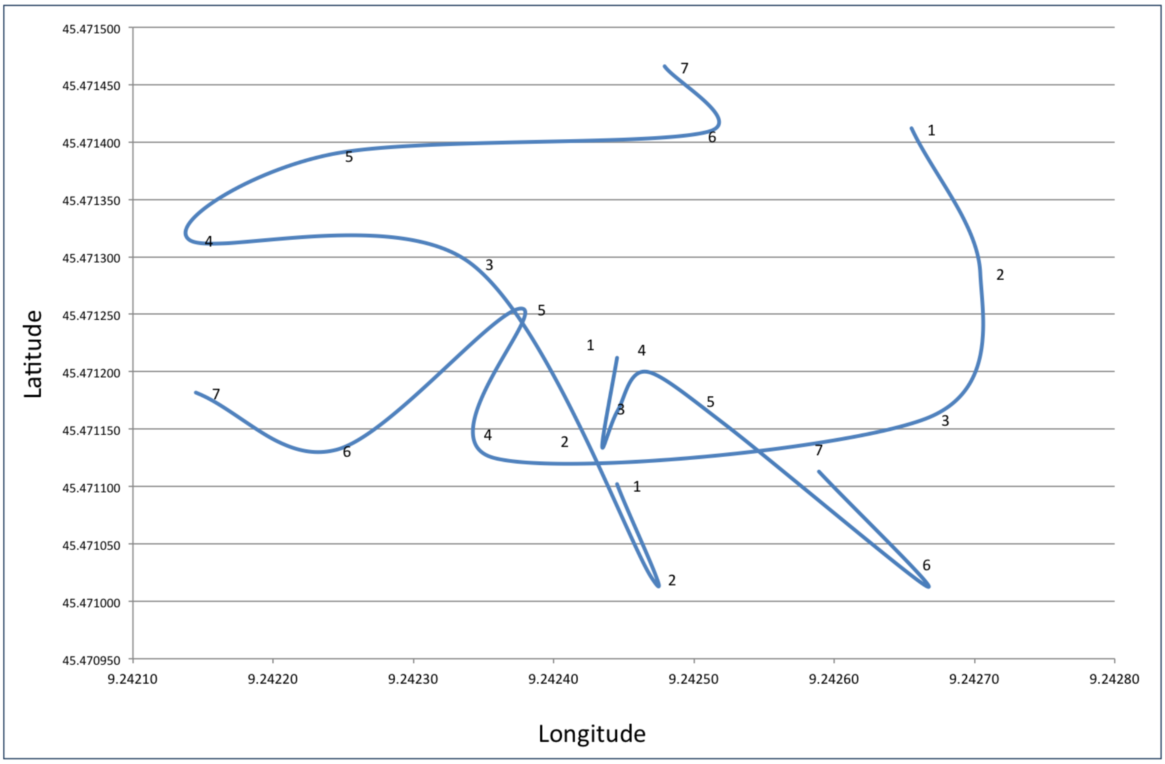
[2] "The synchronic meaning of identity refers to the 'unique sense of self' that a person has about his own being" from (Beller and Leerssen, 2007, p. 4)



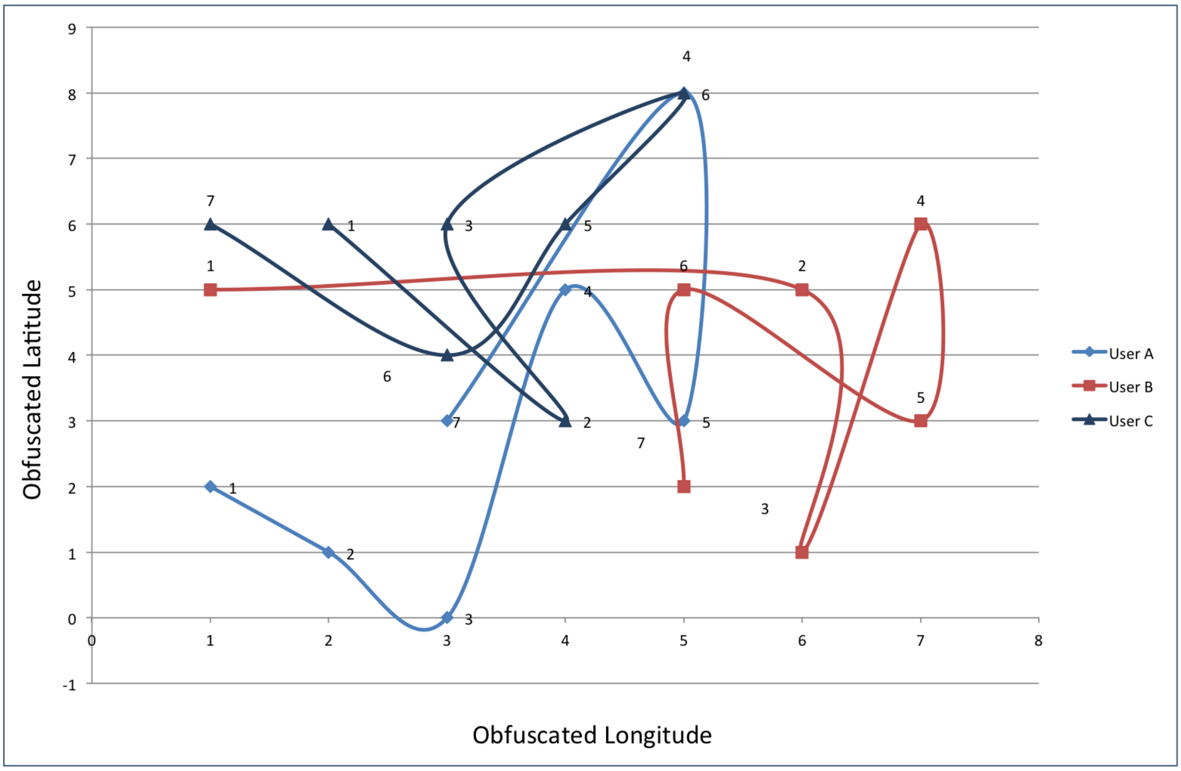
(a) Sample of location information with no alterations



(b) Sample of location information with no timestamps



(c) Sample of anonymous location information



(d) Sample of location information with obfuscated location

Figure 1.1: Behaviour when altering components of location information

Figures 1.1 show how one component can damage the quality of location information, providing perspective of what would be available under different scenarios. For a no protection scheme, an attacker would be able to construct the user's full path as in Figure 1.1a. Now, suppose an attacker gains access to identity and time but has no clear knowledge of what places the user visited, since the location component is obfuscated as in Figure 1.1d. In this case, very little can be inferred given that the context is highly altered and data loses usability; this specific scenario corresponds to *location privacy*, the main focus of this work.

Another solution is the implementation of identity privacy or anonymity. In Figure 1.1c, different pathways can be seen, but no information on the identity of the subjects is revealed. However, travel information alone could serve to infer the identity of the person by matching records on a phonebook as in the experiments conducted by Krumm (2007).

Figure 1.1b describes the case when location and identity are specified, but the time component is missing. In this case, the resulting information lacks context and pathways may not be reconstructed accurately. However, implementing a model in which this occurs is not likely, since requests and LBS responses might lose relevance if delayed arbitrarily.

Some of the mechanisms designed for location privacy protection are based on location obfuscation, which is explained in Duckham and Kulik (2005) as *“the means of deliberately degrading the quality of information about an individual’s location in order to protect that individual’s location privacy”*. However, there is still an open discussion on how much degradation of the location information is sufficient for providing location privacy to users. Moreover, if the resulting obfuscated information remains useful to access an LBS, obtain relevant responses, and enable data analysis. Some of the contributions of this work are based on Location Obfuscation, this technique will be examined further in Chapter 2.

An ideal approach for location privacy would not only allow the LBS to function as close as possible as it would under a no privacy scheme, but would also provide statistics on the LBS users’ behaviour while protecting each user’s location information. There are challenges that this pose, particularly on how to handle location information; the way it is transferred and stored is critical against potential attacks. Geospatial analysis is also a key component for the LBS and data stakeholders in order to ensure improvements to the service as well as for research purposes e.g. citizen science.

This work will focus on creating LPPMs with location privacy techniques that do not alter the geographic coordinate structure of location information in order to enable geospatial analysis. To verify that the proposed LPPMs enable geospatial analysis, common GIS analysis tools are used under different experimental scenarios to quantify the variations that occur when compared to a no privacy scheme.

## 1.2 Thesis Motivation

> *Geoprivacy* are the individual rights to prevent disclosure of the location of one’s home, workplace, daily activities, or trips. The purpose of protecting geoprivacy is to prevent individuals from being identified through locational information.
> Kwan et al. (2004)

Location information is an intrinsic attribute to contextual data of users in available services nowadays. Given that this information reveals sensitive information that leads to a user’s physical surroundings and everyday routines, it is essential to guarantee location privacy protection. However, regular privacy protection algorithms for other kind of data are not suitable to maintain LBS functionality or to provide the possibility of statistical inferences from spatial data due to unique characteristics of this kind of information. Consequently, novel implementations are required to provide geoprivacy, leading to the following question and research objectives that will shape the contents of this thesis.

*How to provide location privacy with mechanisms that are scalable, efficient, non-intrusive and with low impact on geospatial analysis?*

This will be achieved through the design of LPPMs and by experimenting with different parameters and datasets to identify the best settings that comply with the stablished requirements, ultimately creating a family of LPPMs that are suitable for different LBS and their specific purpose. This thesis is guided by a series of research objectives, namely:

**Research Objective 1** Create a taxonomy for LPPMs based on the purpose of LBS;

**Research Objective 2** Design LPPMs that are reliable, scalable, efficient, non-intrusive and that allow geospatial analysis;

**Research Objective 3** Asses the impact of LPPMs based on obfuscation and aggregation on geospatial analysis.

## 1.3 Thesis Contributions

Taking into account the previously described status of the emerging technologies and the standing aside measures to provide location privacy on LBS, this thesis aims to provide an understanding of the current state of LPPMs, examine mechanisms developed by the author throughout this work and their specific purposes. For this, a taxonomy of LPPMs is presented, highlighting the benefits and drawbacks of geoprivacy protection mechanisms in the literature and a categorisation of LBS. This understanding is critical to reach a consensus for integrating the available mechanisms into existing LBS, that should be intrinsically intertwined.

Furthermore, openly released location information through APIs of popular services (e.g., Twitter, Foursquare, etc.), poses an opportunity for researchers to conduct geographical inference studies. This thesis will explore how LPPMs affect such inferences and propose solutions to decrease this impact, the proposed solutions attempt to preserve geospatial analysis results obtained using unprotected geographical data, but with obfuscated location information.

A number of papers, book chapters, and a book have been published for the work of this thesis:

**Zurbarán, M.**, Wightman, P., Oxoli, D., Brovelli, M., Iliffe, M., Jimeno, M., & Salazar, A. NRand-K: Minimizing the Impact of Location Obfuscation in Spatial Analysis. *Transactions in GIS (accepted for publication)*

**Zurbarán, M.**, Wightman, P., Paolo, C., Mather, S. V., Kraft, T. J., & Park, B. (2018). PostGIS Cookbook (Second Edition). PACKT PUBLISHING LIMITED.

Wightman, P., & **Zurbarán, M.** (2018). An Initial Evaluation of the Impact of Location Obfuscation Mechanisms on Geospatial Analysis. In S. V Ukkusuri & C. Yang (Eds.), *Transportation*

*Analytics in the Era of Big Data* (p. 28). Springer International Publishing.

**Zurbarán, M.**, & Wightman, P. (2017). VoKA: Voronoi K-aggregation mechanism for privacy in location-based information systems. In 2017 International Carnahan Conference on Security Technology (ICCST) (pp. 1-6). IEEE.

Brovelli, M. A., Minghini, M., Kilsedar, C. E., **Zurbarán, M.**, Aiello, M., & Gianinetto, M. (2017). Migrate: A foss web mapping application for educating and raising awareness about migration flows in Europe. In *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives* (Vol. 42, pp. 51-55).

Oxoli, D., Prestifilippo, G., Bertocchi, D., & **Zurbarán, M.** (2017). Enabling spatial autocorrelation mapping in QGIS: The hotspot analysis Plugin. *Geoingegneria Ambientale E Mineraria*, 151(2), 45-50.

Brovelli, M. A., Oxoli, D., & **Zurbarán, M. A.** (2016). Sensing slow mobility and interesting locations for lombardy region (Italy): A case study using pointwise geolocated open data. In *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives* (Vol. 41, pp. 603-607).

**Zurbarán, M.**, Avila, K., Wightman, P., & Fernandez, M. (2015). Near-Rand: Noise-based location obfuscation based on random neighboring points. In IEEE Latin America Transactions (Vol. 13, pp. 3661-3667).

Wightman, P., **Zurbarán, M.**, & Santander, A. (2014). High variability geographical obfuscation for location privacy. In Proceedings - International Carnahan Conference on Security Technology (pp. 1-6).

**Zurbarán, M.**, Gonzalez, L., Wightman, P., & Labrador, M. (2014). A Survey on Privacy in Location-Based Services. Ingeniería Y Desarrollo, 32(2), 314-343.

Gonzalez, L., **Zurbarán, M.**, Wightman, P., Jabba, D., Jimeno, M., & Zurek, E. (2013). Sensitivity analysis and countermeasures for transformation-based location obfuscation. In Proceedings - International Carnahan Conference on Security Technology (Vol. 2014-Oct).

Wightman, P., **Zurbarán, M.**, Zurek, E., Salazar, A., Jabba, D., & Jimeno, M. (2013). $\theta$-Rand: Random noise-based location obfuscation based on circle sectors. In ISIEA 2013 - 2013 IEEE Symposium on Industrial Electronics and Applications.

Wightman, P. M., **Zurbarán, M.**, Rodríguez, M., & Labrador, M. A. (2013). MaPIR: Mapping-based private information retrieval for location privacy in LBISs. In Proceedings - Conference on Local Computer Networks, LCN (pp. 964-971).

## 1.4 Thesis Structure

The remainder of this thesis is structured as follows:
**Chapter 2** reviews the relevant literature, including the different LBS types and characteristics, leading to a taxonomy.
**Chapter 3** presents the mechanisms developed in this work, designed for real-time location reporting.
**Chapter 4** describes the mechanisms developed in this work, designed for the time when the LBS releases information as open data or to third parties.
**Chapter 5** shows an initial analysis on the impact of LPPMs on geospatial analysis, emphasising on Exploratory Spatial Data Analysis (ESDA).
**Chapter 6** discusses open issues in the research area, like the lack of commercial implementation and the need for regulations to ensure privacy protection.
**Chapter 7** provides concluding remarks and proposes future research directions, against which work can be evaluated and enhanced.

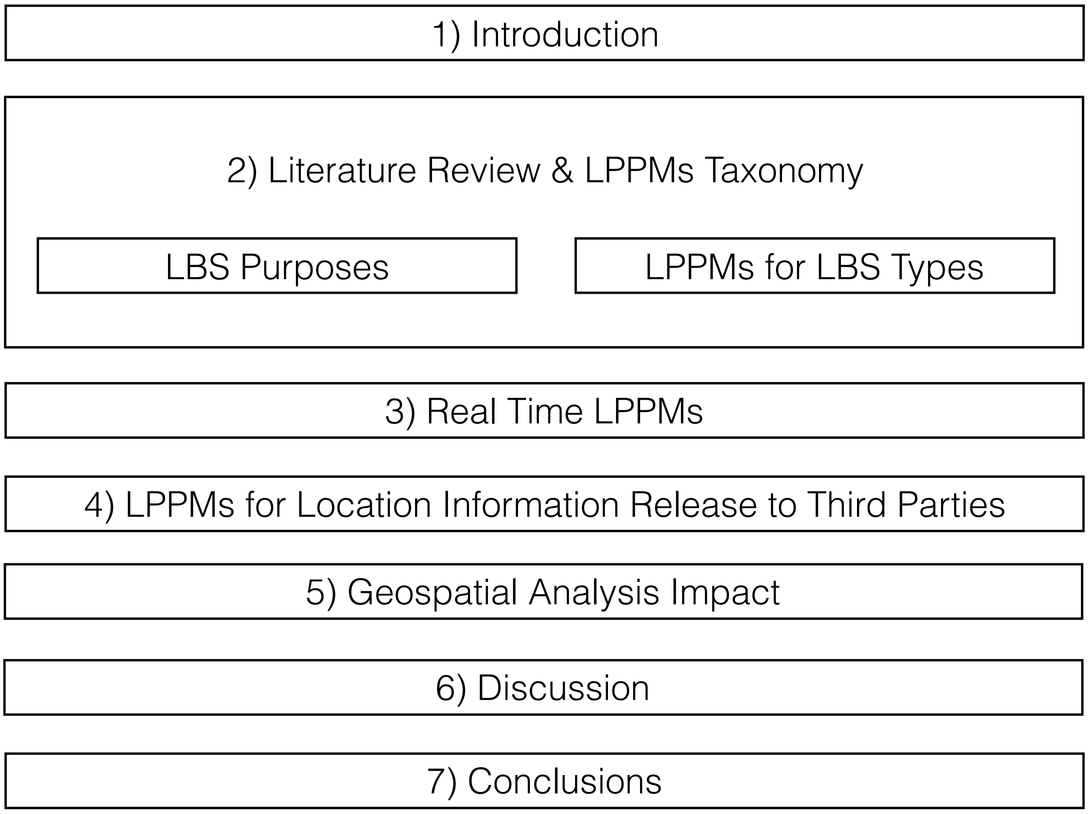

1) Introduction
2) Literature Review & LPPMs Taxonomy
LBS Purposes
LPPMs for LBS Types
3) Real Time LPPMs
4) LPPMs for Location Information Release to Third Parties
5) Geospatial Analysis Impact
6) Discussion
7) Conclusions

Figure 1.2: Thesis Diagram

Figure 1.2 shows the main structure of the thesis.

## 1.5 Chapter Summary

This chapter has laid the foundations for this thesis by introducing the background and relevance of this line of research, the motivation of the research and contributions have been stated as well. The structure of this thesis according to the aforementioned objectives and research question are illustrated through an overview diagram in Figure 1.2, that follows the details described in Section 1.4.

# Chapter 2

# Literature Review

## 2.1 Introduction

This chapter provides a review of the relevant literature, including: a background on geospatial positioning, the description of the basic architecture of an LBS and their different purposes; the use of crowdsourcing for the collection of geographic information, emphasising in the distinction between volunteered and contributed geographic information; and a taxonomy for location privacy techniques. At last, the chapter concludes with a discussion comparing different privacy mechanisms and a chapter summary. This lays the ground for understanding the gaps and existing problems in the field of geoprivacy, giving support to the following chapters and to the research question.

## 2.2 Positioning Systems: GPS, A-GPS & WiFi GPS

In order to explore the concept of LBS, it is important to know the technologies that contributed to its birth; Global Navigation Satellite Systems (GNSS) such as the *United States Global Positioning System (GPS)*, the European Union's Galileo or the Russian GLONASS are some key developments that made possible the emergence of LBS, being GPS the most popular nowadays. Duckham and Kulik (2006) classifies positioning systems into *client-based*, *network-based*, and *network-assisted* based on the resources involved for calculating the receiver position.

When GPS was first launched in 1978, these systems were not conceived to work indoors or for civilian purposes. Rather, they were meant for the military; to guide soldiers, sailors or aircrafts in open field and with a clear view of the sky as stated by the Department Of Defense (2008). The GPS receivers were designed to calculate their position using the observed Doppler shift function, the receivers would get the inputs by searching and acquiring the signal for each satellite and then decoding the satellite's data. This process would take a start-up time of 1 minute and then continue to provide continuous positioning according to van Diggelen (2009). The wait for a first fix is due to the long distance and how weak the signals from the satellites are at the time they reach the GPS receiver. Autonomous GPS receivers used in actuality have a median error of 5 m according to the Department Of Defense (2008), while Zandbergen (2009) argues that this error is of 8 m.

Positioning systems have since been improved by other technologies, bringing in *Assisted-GPS (A-GPS)*; where cellular towers assist the receiver in the way it looks for signals, the time for the initial fix would reduce to 1 second as the signals come from a closer source and are not as weak as the ones from satellites. This scheme that produces stronger signals also helps in reducing the communications and energy cost for the receiving device, allowing for positioning calculations in urban infrastructures. The median error is increased, ranging from 20 m outdoor and 74 m indoor according to Zandbergen (2009), which is higher than dedicated GPS receivers under ideal outdoor conditions, but this is still acceptable for most LBS application.

In addition to cellular networks, bluetooth and WiFi networks were introduced to improve calculation times and accuracy of A-GPS, this new method called *WiFi GPS* or *WPS* use fingerprints and signal strength measures from access points to ensure a higher accuracy and promptness for geo-spatial positioning. Even if no GPS signal is available, location database vendors would still provide acceptable positioning of 10 to 20 m of accuracy including indoor settings (Skyhook, Skyhook) thanks to a crowd-sourced database of WiFi hotspots and cell tower locations.

## 2.3 Location Based Services

As previously mentioned in Section 1.1, LBS are part of LBIS, and form a specific type of services that differentiate from others in that they use geographic information from users in order to provide their functionalities. KuÌĹpper (2005) complies with this definition citing the *GSM Consortium* stating that *"LBS use the location of the target for adding value to the service, where the target is the 'entity' to be located (and this entity is not necessarily also the user of the service)."* Furthermore, he also proposes a general characterisation of actors involved in LBS by discriminating by operational and non-operational roles.

In the following subsections, LBS architecture and a classification of the different types will be examined, further on, a discussion on privacy techniques.

### 2.3.1 LBS Architecture

The communication architecture for an LBS involves basic components as shown in Figure 2.1. The steps an LBS follows to provide its functionality are:

1. *The target* –or namely, the user– uses her mobile device's GPS receiver that calculates its position from satellites, cellular towers and access point signals;
2. The target accesses the cloud through Wi-Fi or a *mobile network operator (MNO)* to use the LBS through a web or a native mobile application;
3. The user's device sends its position while making a request to the *LBS provider*;
4. LBS processes the request with the received position, during this process, the LBS may use a Geographic Information System (GIS) with a database that provides the necessary content in order to send a response to the user;
5. The target's device receives the response and displays it in the application.

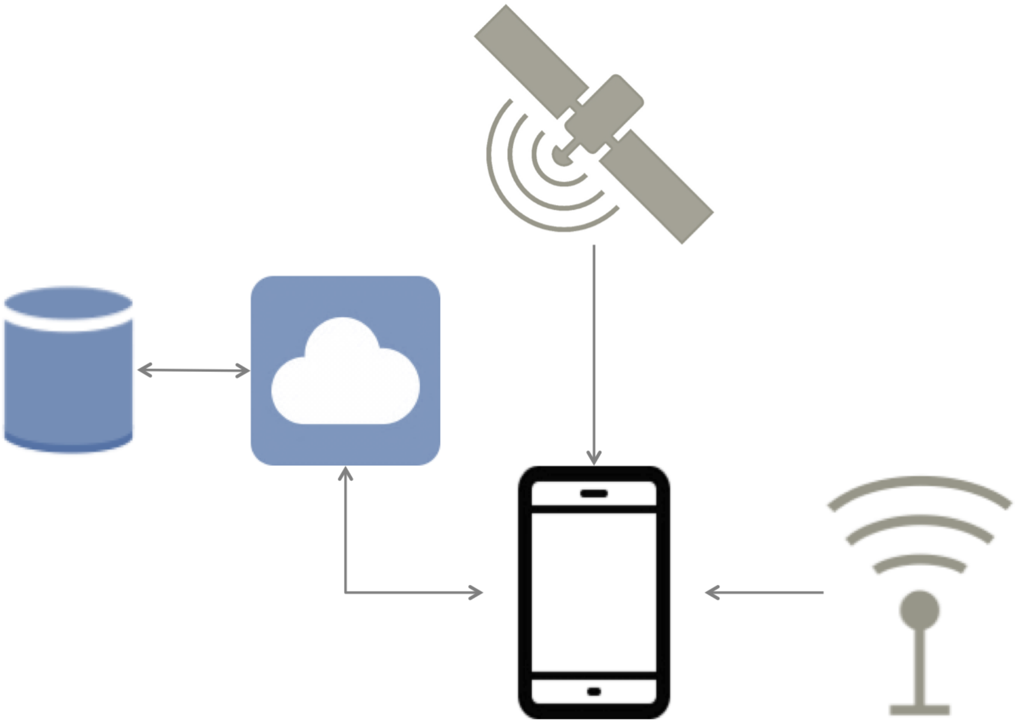


Figure 2.1: Architecture of an LBS

### 2.3.2 Types of LBS

LBS serve for many purposes: traffic navigation, tracking of goods, social networks, among many others. From a location reporting viewpoint, there is however a differentiation that helps in their study: LBS can require continuous or discrete location reporting according to the offered functionalities. Continued reporting is understood in the frame of coordinates seconds apart from each other, producing trajectories of the subject, while discrete reporting occurs every once in a while and only when the user queries the service, producing a more spread distribution of the collected location information and not a trajectory. This classification is useful to discern specific privacy requirements for the type of LBS, this differentiation has been used as well in (Seidl et al., 2015) and further explained in (Labrador et al., 2010).

### 2.3.3 Continuous Location Reporting or Proactive Applications

Continuous reporting or *'proactive applications'* (Labrador et al., 2010), are the ones that require intensive location reports from the user's side. They usually provide real time information with a prescribed accuracy according to a predefined set of conditions. Hence, there is a need to use the most up to date information from the user when these conditions are met. Examples for such applications are friend proximity alerts, traffic navigation, or the use of *geofencing*, i.e. monitoring users within a predefined boundary. In this kind of services, the actors that take part and their roles are identified as follows:

1. *Monitoring User* is the LBS user that requires to trace locations of a tracked device; many monitoring users could track one device;
2. *Tracked device* is the one that is constantly reporting its location to the LBS to be observed by approved monitoring users;
3. *Server* is what provides a platform for communicating between users and tracked devices, also stores logs and historical traces and provides content related to the offered service.

### 2.3.4 Discrete Location Reporting or Reactive Applications

These LBS applications act upon user's request: the user sends a query containing their spatial location and asks for information that depends on their geographic context. The LBS processes this request and sends a response that satisfies the query. A common application of this kind of application is *nearest-neighbor* (NN) search, in this case the user *"searches for data objects that minimize a distance-based function with reference to one or more query objects* (Khoshgozaran and Shahabi, 2007); when the user specifies or is able to retrieve more than one neighbour, then it is referred as *k-nearest-neighbour* (KNN) search. Unlike proactive services, these applications are of the request/response type as mentioned in (Labrador et al., 2010). Some examples of this type of LBS applications are NN searches, consulting local news, requesting a transportation service, among others. The identified actors for reactive applications are:

1. *Requesting User* is the one who issues the request based on their current location calculated by their mobile device;
2. *Server*; as in the proactive scheme, provides the platform for communicating with users and serves as a geospatial database to fulfil user's requests;
3. *Predicate* of the request, which contains specifications to delimit the server response to match the user's needs. This predicate may be in the form of a category for PoIs, time frames for consulted information or a delimited geographic area. The use of predicates was introduced by Lu and Jensen (2008).

## 2.4 Crowdsourced Spatial Data Collection

There are two approaches for collecting crowdsourced geographic information, the differentiation is based on the awareness that the providers of the collected data may have about the purposes of the provided information. In this way, geographic information can be: *'Volunteer Geographic Information' (VGI)* or *'Contributed Geographic Information' (CGI)*, both methods involve laymen or amateur users in the collection of spatial data and form a rich volume of the available geographic data, specifically the kind of data that this work aims to protect.

Harvey (2013) stresses the distinction between *volunteered* and *contributed* geographic data, stating that where they differ is in the kind of agreement under which the collection happens. The user agreements can be *opt-in* or *opt-out*: the first one allows volunteered or *active* participation, where the user has control over the collected data and over data reuse, while the second one is contributed or *passive* participation and the user does not have enough control, if any, over the frequency and amount of data collected or to what happens to the data after collection.

On the uses of crowdsourced spatial data, Harvey (2013) affirms that *"crowdsourced geographic information already can play an important role in the ability of companies and government agencies to know and predict people's activities"*[1]. Goodchild (2007) explains the risks and opportunities that abound in geographic information generated by amateurs and how it impacts cartography nowadays.

Harvey (2013) establishes differences between CGI and VGI in his work. He also acknowledges that data collected as VGI may be reused as CGI. This situation is depicted in the following example: *"when access and reuse are controlled by the person, but the company operating the*

[1] (Chapter. 3 Harvey, 2013, p. 33)

*site uses geographic information about the location to profile users and sell aggregated data to mobile advertisers"* [2]. This is not uncommon in search engines or social media platforms, where the user's data is in many cases reused to conduct statistical analyses to target audience with specific promotions relevant to their context and spatial surroundings. In other words, this is referred as *'location analytics'* and according to a study by Google (2014) *"four in five consumers want search ads to be customised to their city, zip code or immediate surroundings"*. This provides a profitable ground for indiscriminate location data collection.

The possibilities of VGI took researchers focus to *"understanding its characteristics, discussing approaches of data collection, presentation, validation, and community building with reference to VGI supporting environmental monitoring"*, as well as for *"investigating the characteristics of VGI stakeholders"*[3].

There are successful experiences made possible by the use of VGI, some of them include response for disaster relief, like the case of the 2010 Haiti earthquake, In (Koukoletsos, 2012, Chapter 7), the author examines the geospatial quality of the contributions of crowdsourced data for this disaster; finding that when compared to other data sources, such as those from the UN Stabilisation Mission for Haiti ("MINUSTAH") and from Google Map Maker (GMM), OpenStreetMap (OSM) was the richest dataset. Koukoletsos (2012) also concludes that in general, crowdsourced data is more up to date than authoritative data.

With the Volunteered Geographic Information (VGI) model, services have been created by using spatial information from conscious users, this information may be the user's geographic knowledge or geographic location reports, that are then made publicly available through the service to provide a specific functionality. Whether this information is obtained from volunteers, should not imply a resignation to a minimum privacy level. It can be argued that VGI systems could benefit from attracting privacy concerned users (or volunteers) by implementing protection mechanisms as mentioned by Xu and Gupta (2009).

## 2.5 Privacy Techniques in LBS

Privacy in location based services focuses on reaching a desirable trade-off between performance and user's privacy; the more privacy provided, less likely it is for the service to function as proper as it would without any privacy protection considerations. Privacy techniques should adjust to enable the existing features in LBS; however, the provided features are evolving continuously. LBS change in order to keep a balance between: meeting users' requirements, using the latest technology, and maximising the application lifetime value. This poses a challenge for creating fitting LPPMs that are reliable, efficient, scalable, and non-intrusive.

[2] Found in (Chapter. 3 Harvey, 2013, p. 34)

[3] (Chapter. 2 Iliffe, 2017, p. 22)

When applied to proactive services, LPPMs ought to keep location information about the *tracked devices* undecipherable to anyone different from the allowed *monitoring users*, even to the LBS itself; due to the privacy breach this implies for a user considering data reusability by the service providers. Generally, proactive services require a high level of accuracy on the tracked subject's location, relying mostly on how good is the approximation of the positioning system. The challenge for LPPMs intended for proactive applications is to maintain the level of accuracy that the positioning component in the device can deliver, while assuring that the location information provided does not get disclosed. On the other hand, mechanisms intended for reactive services often do not require critical accuracy, and therefore it is tolerable to alter the subject's position in order to provide location privacy; however, it is not a straightforward task to determine how much alteration is enough or too litle for every service.

Among the existing LPPMs, some require special implementation on the server side architecture or database structure, in some cases even requiring intermediaries, such as third parties or the use of proxies. Other alterations of the basic LBS architecture presented in Figure 2.1, include a peer to peer approach between LBS users, namely the community, to cooperate in order to mask their location. In Figure 2.2 are depicted different schemes that will be found in proposed mechanisms included in this review.

The additional actors present in Figure 2.2 that are introduced for privacy protection and were not previously described in section 2.3.1 LBS Architecture are defined as follows:

1. *Community* includes all the users of the LBS that may intervene in the functionality of the service, such is the case of applications used to monitor traffic or location-based social networks. Community members could participate in methods for providing location privacy as well.
2. *Proxy* is a service that provides security at network level to protect user's communications. These services could be distributed such as The Onion Router (TOR) (The Tor Project, 2016) or centralised like Virtual Private Networks (VPNs).
3. *Third party* is an external relation that intervenes to provide location privacy in conjunction with the LPPM. Third parties relations act as proxy-like servers at application level that centralises the architecture, Rannenberg et al. (2009b) defines it as: *"A subjective, dynamic, context-dependent, non-transitive, non-reflexive, non-monotone, and non-additive relation between a trustor and a trustee"*.

Pragmatically, an architecture scheme that does not modifies the existing model of the LBS is desirable; for example, in Figure 2.2, *A* represents the straightforward approach, while *C* presents a more complex and expensive alternative in terms of infrastructure, communication and energy consumption. In general, a simpler model should be more convenient while assessing LPPMs; in this work modifications required by an LPPM in infrastructure are considered a drawback for potential adoption by service providers.

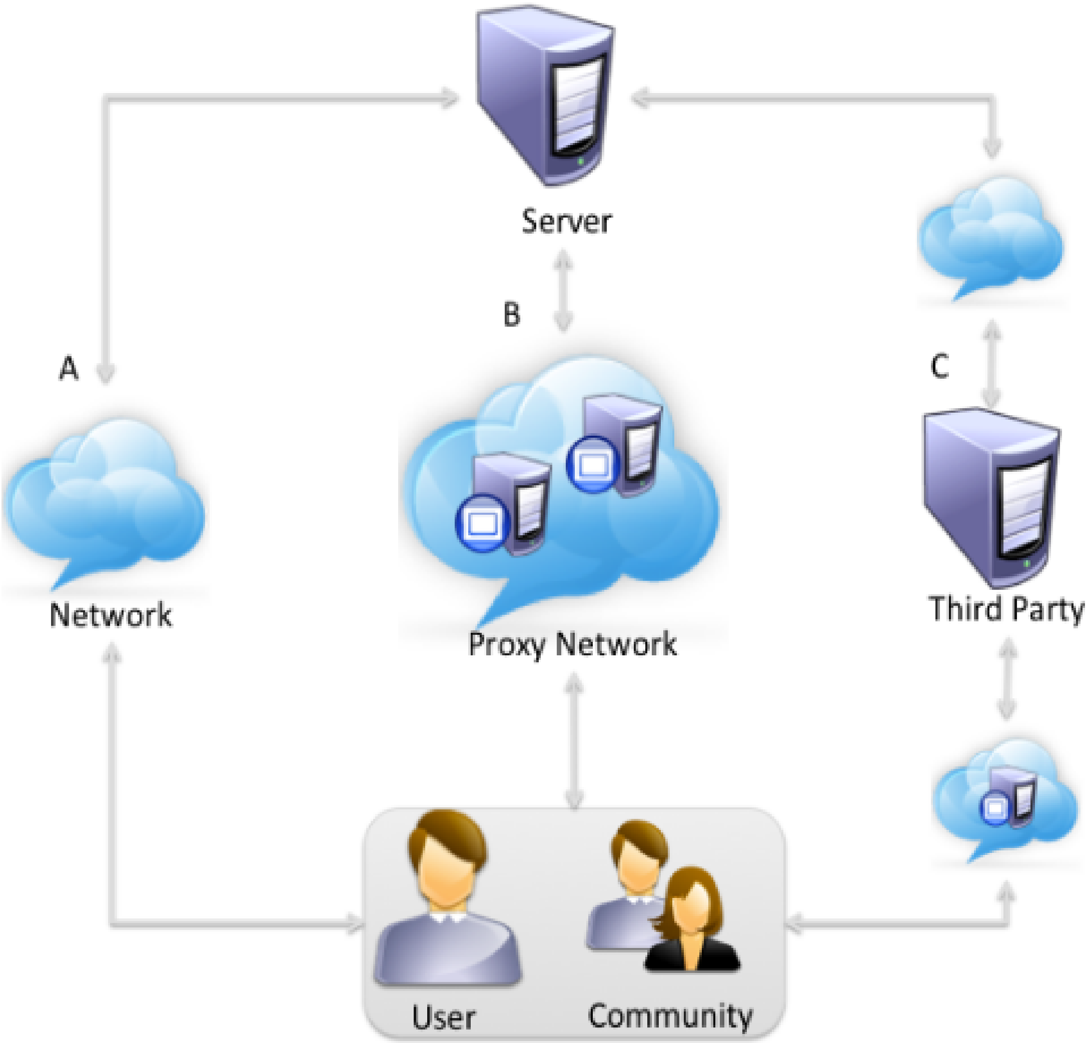


Figure 2.2: Architecture schemes for LPPMs

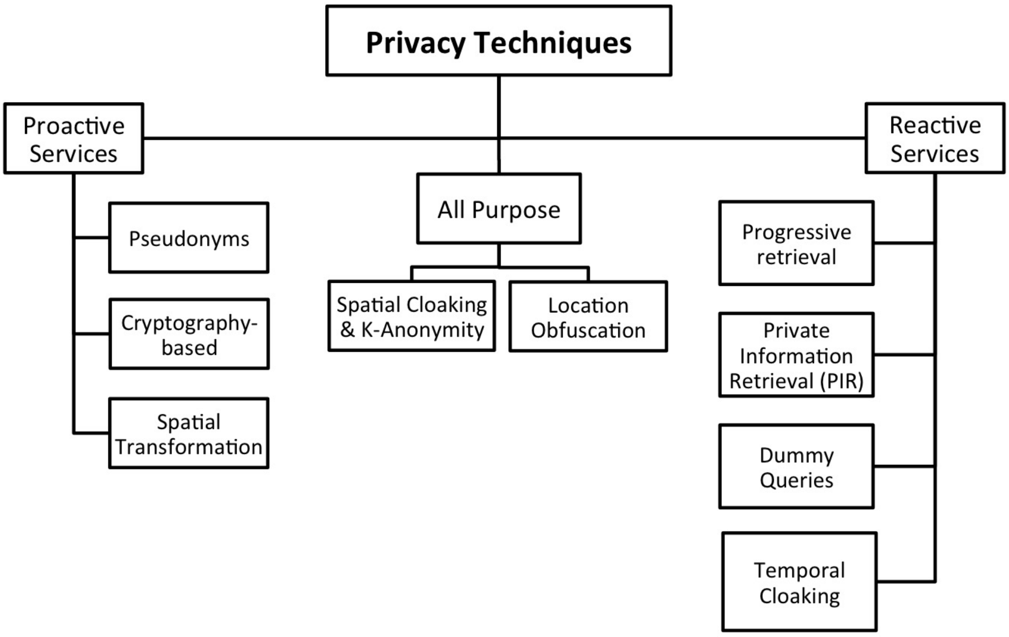


Figure 2.3: Privacy techniques for LBS

In the literature, there are many techniques for preserving location privacy, the techniques explored in this work include:

- Spatial Cloaking & K-Anonymity
- Location Obfuscation
- Pseudonyms
- Cryptography-based
- Spatial Transformation
- Progressive Retrieval
- Dummy Queries
- Temporal Cloaking
- Private Information Retrieval (PIR)

A Location Privacy technique is a general approach that LPPMs use in order to achieve location privacy, some LPPMs may use a combination of privacy techniques in their configuration. In order to study the existing techniques, a taxonomy was designed based on the types of LBS and their purpose as described in Section 2.3.2. The diagram in Figure 2.3 shows techniques that are suitable for continuous, reactive or both kinds of LBS. This diagram serves to guide in identifying among the broad spectrum of LPPMs available, each technique will be presented with mechanisms that use in this review to assess the state of the art in privacy mechanisms using a fit for purpose approach to classify them.

## 2.5.1 LPPMs for All Purpose

### 2.5.1.1 Spatial Cloaking & K-Anonymity

This category encompasses mechanisms that aim to anonymise the reported location of an LBS user among other *k-1* users. In this technique, it is usually required a trusted third party for anonymising or a decentralised architecture (P2P), where the *community* share their location between them before issuing queries. The anonymiser, either centralised or through the community, is assumed to know the location of all users. According to Gong et al. (2010), k-anonymity algorithms work as follows: a user sends its location, query and a *k* value to the anonymiser. The anonymiser removes the ID of the user, then cloaks the user location including at least *k-1* other users. The cloaked region is sent with the query to the LBS sever, which calculates the candidate results and sends them back to the anonymizer. Kalnis et al. (2007) refer to this cloaked region as *k-anonymising spatial region* (K-ASR). Finally, the anonymiser calculates the actual results for the exact user location and sends them to the issuing user. It is worth noting in these techniques that spatial density impacts the K-ASR. Regions with higher density of users, result in smaller K-ASR, whereas regions with low population may require larger K-ASR. Both extremes derive in providing little protection.

It has been argued that k-anonymity in relational databases for a high-dimensional relation of any kind of data leads to unacceptable information loss due to the *'dimensionality curse'* Aggarwal (2005).

Algorithms used for spatial cloaking are often based in k-anonymity for calculating cloaked regions. Spatial Cloaking is where *"an individual's device or a third party cloaks the individual's location before giving it to the provider of a location-based service"* [4]. Chow et al. (2009) propose the Casper query processing framework, that enables the use of cloaking regions for querying public (restaurants, gas stations, etc.) and private (nearby friends) location information. This mechanism requires an anonymising third party and can be used for proactive or reactive LBS.

Kalnis et al. (2007) present Nearest Neighbour Cloak based on quadtrees to map the query space and the Hilbert Cloak, which uses Hilbert Curves to cloak location information. The user locations are mapped into a Hilbert Curve and split into k-sized blocks or *'k-buckets'*, these blocks are then used instead of single locations to guarantee k-anonymity.

Cheng et al. (2006) present a cloaking approach based on probability, this mechanism does not employ k-anonymity, it rather uses uncertainty regions to produce an imprecise location. The authors do not specify a geometry type for the uncertainty regions, but it must be a closed polygon. The LBS has to implement a way for interpreting the imprecise queries.

Monreale et al. (2013) aim to provide differential privacy, which is a strong privacy model independent from the knowledge an adversary may have. For this, they propose the use of generalisation and aggregation over trajectory data based on space partitioning. After generalisation, the user issues the protected location information. For aggregating the trajectory data, Monreale et al. (2013) argue that *"administrative districts or regular grids, do not reflect the spatial distribution of the data"*. Consequently, the use of historical data is proposed to identify spatial clusters to generate a Voronoi tessellation of the territory. For estimating the provided privacy, they introduce a utility evaluation where the error induced with the spatial transformations is analytically measured and compared in different experiments.

#### 2.5.1.2 Location Obfuscation

This category includes methods that expand the LBS's assumption about the actual query location to a wider sub space of the spatial domain, called obfuscation region. Location obfuscation is a technique to provide privacy that, according to (Duckham and Kulik, 2005, p. 155), is defined as *"the means of deliberately degrading the quality of information about an individual's location in order to protect that individual's location privacy"*.

[4] (Zhong et al., 2007, p. 2)

Geographical masking or simply masking was first coined by Armstrong et al. (1999). Masking was initially designed for geographic health data, because of the critical impact of disclosure. In their words: *"The purpose of this paper, therefore, is to describe and evaluate alternative approaches to encoding the geography of health records which we call geographic masking that protect the confidentiality of individuals but which also ensure the possibility of valid geographical analyses of the data"*(Armstrong et al., 1999, p. 498). Even though these terms come from different areas of research, they both share similarities and are often used interchangeably. For this work, location obfuscation will be preferred.

Noise-based location obfuscation mechanisms have very little requirements for their implementation; these involve minimal or no alteration on the LBS architecture while preserving a geographic coordinates' data structure, which is critical for geospatial analysis.

Ardagna et al. (2007) refer to *'relevance'* as a measurement of the provided privacy by location obfuscation. This is achieved by enlarging the inaccuracy of the already inaccurate reported location provided by the MNO. In another work, Ardagna et al. (2011) present various obfuscation operators that serve to decrease location accuracy while still providing coordinates information. These operators are *inaccuracy*, which concerns a lack of correspondence between information and reality; *imprecision* which is a lack of specificity in information; and *vagueness*, that concerns the existence of boundary cases in information.

As it happens with other techniques, there is still an open discussion with noise-based location obfuscation on how much degradation of the location information is sufficient for providing location privacy to users. Seidl et al. (2015) use maximum noise thresholds of 30, 50, and 250m; Kwan et al. (2004) experimented with 1032m; and Krumm (2007) demonstrates how inferences can still be made with an induced noise of 1000m, but are less likely than when 500m is used. This suggest that there is no magic number for this value. Rather, the induced obfuscation depends on the LBS application requirements: 500m of data distortion in an accuracy critical application like Waze would be impossible to deal with, while an app for finding a drugstore, e.g., yellowpages, could manage to function with 200m of accuracy loss.

Location obfuscation is not a robust way to ensure privacy protection because these mechanisms still reveal the whereabouts of the user. However, obfuscation does provide a lightweight, decentralised, non-reversible and customisable alternative to hide the exact location of the user, while still being able to access the service.

Wightman et al. (2011) present the *N-Rand* algorithm, consisting on the generation of N uniformly distributed random points within a circular domain centred on the original coordinates. The farthest generated point from the original location is then reported to the service.

## 2.5.2 LPPMs for Proactive LBS

### 2.5.2.1 Pseudonyms

Pseudonyms are an alternative to provide identity privacy in location based services; however, the use of pseudonyms alone is not sufficient to provide location privacy in a LPPM, given that if a pseudonym stays the same over time, it will eventually lead to the identification of the user. Krumm (2007) shows that phone book inference attacks are possible from location records if the location is not altered sufficiently, location information itself can be linked to other data for the identification of subjects.

A Mix Zones method is presented by Beresford and Stajano (2004), where users are given pseudonyms that serve to communicate with the LBS, but these pseudonyms are changed every time a user enters a mix zone. This is a zone where users cannot be tracked and swapping pseudonyms can safely occur, providing periods of anonymity so the server cannot build usable traces for each user. In order to control the registry of mix zones, a trusted middleware or third party is required to handle communications between users and the LBS.

### 2.5.2.2 Cryptography-based

Common offered functionalities in LBS involve geofencing, friend finder applications are an example of this; monitoring *targets* within a delimited boundary. Location privacy for these kind of services should aim to protect the target's location from the LBS and alert only the *monitoring users* when pertinent (i.e., when the target is within the specified boundaries). Zhong et al. (2007) propose three LPPMs named: Louis, Lester and Pierre for solving the nearby-friend problem with privacy concerns. For this, the author consider a privacy scenario as *"an instance of a secure multiparty computation problem, where multiple parties jointly compute the output of a function without learning each other's inputs"*. The proposed mechanisms fit in the cryptography-based technique by using homomorphic encryption, which Micciancio (2010) defines as *"a special kind of encryption that allows operating on cypher-texts without decrypting them; in fact, without even knowing the decryption key"*. Among these LPPMs, the Louis protocol requires a semi-trusted third party; however this one does not learn any location information from *targets*. In the Lester protocol is not necessary to include a third party, but has the drawback that a user might be able to learn a friend's location even if the friend is in an area that is no longer considered nearby. The Pierre protocol, being the most privacy preserving, does not have the disadvantage of Lester, but is not able to tell the user the precise distance to a nearby friend.

Mascetti et al. (2011) focus on friend finder applications as well and define a privacy scheme where the service provider should have as little information as possible and the user's friends should know the proximity but not the exact position, besides, any eavesdropper in the network should not be able to filter any location information about the users. To accomplish this, it is used a *'Minimal Uncertainty Region'* (MUR) where *"the user accepts that the adversary knows she/he is located in a MUR R, but no information should be disclosed about her position within R"*. The concept of MUR is comparable to that of *relevance* for location obfuscation by Ardagna et al. (2007).

In order to obtain these uncertainty regions, they use spatial granularity, which is understood in many LPPMs as *"a subdivision of the spatial domain into a discrete number of non-overlapping regions, called granules"* (Mascetti et al., 2011). The two protocols presented are C-Hide & Seek and C-Hide & Hash, both adopt symmetric encryption techniques where each user poses a unique key that is shared with their friends and vice versa. The key exchange is performed through a secure communication before executing the protocols. In this scheme each user has to report their location to the service provider, this is done by discretising time in update intervals. The success of these protocols lies in the fact that for every update of a user, a different key is used. This is possible due to the generation of a *'keystream'* based on the initially exchanged key; each buddy will be able to generate the key corresponding to the current update interval of their approved buddies and therefore decrypt the identification of the granule where the user is located. The protocols C-Hide & Seek and C-Hide & Hash differ in the provided privacy. The first one lets the buddies know the granule where the user is located, while the second manages to provide full privacy without disclosing the granule unless the buddies are in proximity. C-Hide & Hash use more computational resources than C-Hide & Seek.

#### 2.5.2.3 Spatial Transformation

Wightman et al. (2012) use homomorphic encryption for proactive LBS in the Matlock mechanism, which is based on a matrix obfuscation technique, where an array $\mathrm{M}_{(1,3)}$ is used, containing latitude, longitude and the time when the coordinates were obtained. An additional squared $\mathrm{N}_{(3,3)}$ array is required to perform the obfuscation operations, this is, multiplying the array M by N; N could be arbitrarily generated with the only restriction that it must be invertible. The resulting array $\mathrm{Q}_{(1,3)}$ contains the location information encrypted and undecipherable to anyone without knowledge of N, the key. In order to decrypt the information, the inverse of the N matrix ($\mathrm{N}^{-1}$) is multiplied by Q, resulting in the original M array with the location information; N is used as shared key between the monitoring users and target devices, using a secure protocol for exchanging it. Each user has one N matrix key for all reported locations. The method allows to recover the unaltered reported position, while providing location information for the LBS. Matlock can be used for most proactive applications. This mechanism uses key exchange, which is typical of cryptography-based mechanisms, but also transforms the space and is able to perform proximity operations even with this alteration. In this way, both techniques, cryptography and spatial transformation are used to achieve geoprivacy.

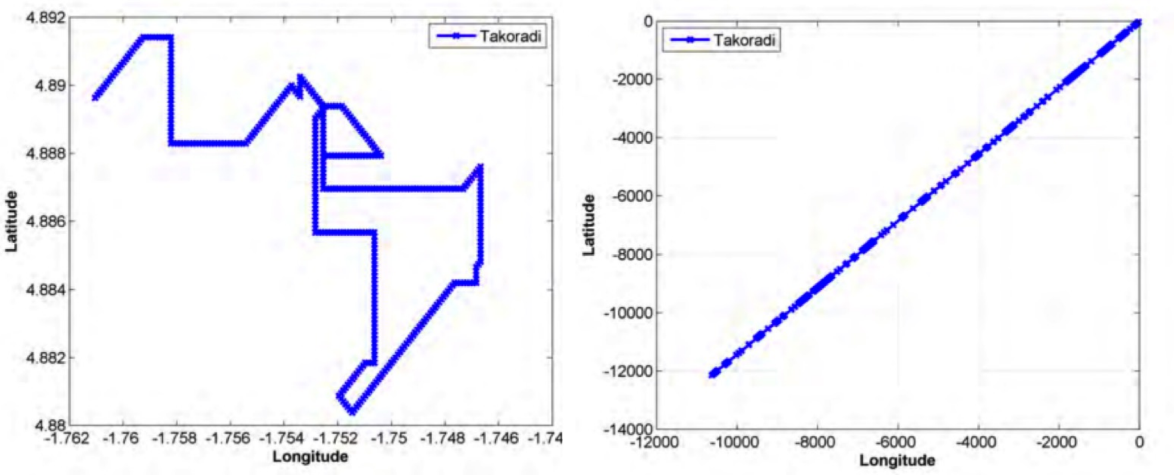


Figure 2.4: Original path vs the Matlock transformation [5]

### 2.5.3 LPPMs for Reactive LBS

#### 2.5.3.1 Progressive Retrieval

Methods based on Progressive Retrieval (PR) perform many requests for a single user interaction, these methods use *'Incremental Nearest Neighbour'* queries (INN) presented in Hjaltason and Samet (1999). This technique aims to reveal as least location information as possible to obtain the desired service. A PR mechanism is presented by Huang et al. (2008) named SpaceTwist, this mechanism has been modified into other versions to introduce privacy improvements; AnonTwist by Wang and Wang (2009), ASkNNA by Gong et al. (2010) and PuPPeT by Riboni et al. (2011). All algorithms based on SpaceTwist use an *'anchor'*, which is a fake location that is contained within a circular area of radius *P* from the original location; a *'demand space'*, which is centred in the original location and serves to evaluate if a PoI is near

[5] Taken from (Wightman et al., 2012, p. 1832) with permissions from the authors.

enough to be considered as relevant; and a *'supply space'*, which initially is just the anchor, but then increases the circular region the nearest PoI retrieved in each iteration. The SpaceTwist algorithm finishes when the demand space is fully contained within the supply space, guaranteeing that the nearest PoI for the real location is available to the client without revealing the exact location to the server.

Gong et al. (2010) introduces cloaking algorithms to determine a k-anonymous anchor, while AnonTwist (Wang and Wang, 2009) uses density maps avoiding the need for a trusted third party for anonymising. Further on Riboni et al. (2011) present an added privacy feature to AnonTwist by guaranteeing *'absence privacy'* as well; this is defined as allowing the user to specify a *'puppet location'* where she does not want an attacker to infer that she is not present, i.e., avoid disclosing when the target is not at home. In order to do so, there is a constrain that makes the maximum distance between a puppet location and the user's real location to be half the initial radius, this is *P/2*. With a *puppet location* already specified, the algorithm keeps requesting PoIs until the puppet location is contained within the candidate area, at the cost of making this technique less efficient.

#### 2.5.3.2 Temporal Cloaking

Gruteser and Grunwald (2003) developed a temporal cloaking approach, arguing that temporal resolution of location information can be reduced to guarantee k-anonymity. The authors use the U.S geological survey (USGS) to create automotive traffic simulation since no testbed was available when the mechanism was introduced. The authors conclude that an accuracy comparable to that of E-911 requirements could be provided and improved with temporal cloaking, ranging from 30 to 250 m of median accuracy. This mechanism is not suitable however, for interactive or proactive applications due to the increase in waiting time to guarantee k-anonymity.

#### 2.5.3.3 Dummy Queries

The use of dummy queries is a popular approach to provide location privacy in PoI search services or NN searches, it consists on sending *n* fake requests along with the real one in order to disguise the user's true location, this technique poses downsides as it requires the server to process *n-1* additional queries to the one relevant to the user, this incurs in computation overhead and communication costs, however there are some techniques developed based on dummy queries that manage to decrease such costs.

Kido et al. (2005) developed the first mechanism based on dummy queries, this first approach proposes the generation of *n-1* fake locations to be sent to an LBS to provide location privacy. The scheme assumes users that are constantly reporting its location to an LBS and therefore would not be located too distant from the immediately previous reported location. The dummy locations are generated in a way that form feasible traces of a regular user; for a first query of a user, there are generated n-1 random fake location and sent to the LBS along with the real one; for the following requests, the method bases the generation of *dummies* in the ones previously reported in order to build *n* possible traces for that user.

Quercia et al. (2011) introduces the *SpotME* mechanism, that aims to work with a large scale amount of users to count people in certain areas in order to provide information related to traffic, crowd analysis, and other crowd-sensing applications. SpotME requires the geographical space to be divided in discrete locations, these locations are then chosen to say whether or not a user is present in such location, each statement with 50% of probability and the users are not forced to answer truthfully, hence producing sometimes dummy reports. This data is later manipulated by an algorithm that estimates the real proportions of all the data received by the LBS, showing an accurate result of people concentration at certain locations without identifying between users. SpotME is also able to indicate if people are entering or exiting a location, which results useful to estimate population flows. The vulnerability of this mechanism resides in the ability of an attacker to collect more than one map of the location sent by a user and while comparing and intersecting these maps it is possible to reveal location information and break anonymity.

Lu and Jensen (2008) propose two techniques to generate dummy requests. Both techniques send a single message to the service, producing lower communication costs in the request than in previous mechanisms; however it requires a light transformation on the server side for processing the requests. The requests are formed by *n* positions and a type of interest or predicate. The first proposed mechanism generates a grid of dummies with a dummy position on each vertex, while the second one generates the dummies based on a virtual circle that contains the user's real location. After receiving the request, the server processes all the locations with the same predicate and retrieves the responses for each location, this is later filtered by the client leaving only the pertinent results for the real user location.

#### 2.5.3.4 Private Information Retrieval (PIR)

Private Information Retrieval is a widely used approach for providing Location Privacy on NN searches, formally, it was first defined by Chor et al. (1998) as *"schemes that enable a user to access k replicated copies of a database with k$\geq$2 and privately retrieve information stored in the database. This means that each individual database gets no information on the identity of the item retrieved by the user"*, but this approach was not initially intended to be used on a single database, it required replication on at least two databases with communication restriction between them, since it aimed to provide information theoretic privacy, which demands an adversary with no knowledge of the information requested and assumes unlimited computational resources for the attacks. It was not until 1997 when Kushilevitz et al. (1997); Chor and Gilboa

(1997) presented a computational PIR (cPIR) technique with a single database, this scheme assumes an attacker limited to probabilistic polynomial-time computations. PIR techniques are challenged to provide solutions with reasonable computation and communication costs since these implementations usually require special processing in the server side of the LBS.

Khoshgozaran and Shahabi (2007) proposed a one-way transformation for the 2-D space, being able to *blindly evaluate*, or in other terms privately retrieve results for KNN queries with a complexity of O(K). For this, they use space filling curves, more specifically, *Hilbert Curve* that pose the property of keeping the proximity and neighbouring aspects of the data, in this way this mechanism uses spatial transformation for PIR of location information. For processing the queries, the authors propose tamper-proof devices and a trusted third party in the architecture to distribute *Space Decryption Keys* (SDK); these keys are composed by the Hilbert Curve construction parameters, hence serving to decrypt the retrieved results.The privacy scheme the authors use to evaluate the proposed mechanism introduces the concepts of *u-anonymity*; the issuing user is indistinguishable of any other user, *a-anonymity*; the location of the query point is not revealed, and *result set anonymity*; the retrieved PoIs are secret to the LBS. Under this scheme, a KNN query is blindly evaluated if the three constraints are satisfied.

Ghinita et al. (2008) presents PIR methods with reasonable computation and communication costs, AproxNN uses Khoshgozaran and Shahabi (2007) Hilbert Curve transformation to represent PoIs in a 1-D space. The PoIs are queried by the user and an approximate nearest neighbour is retrieved using binary search. For the query processing, it uses a B+ tree that contains the PoIs in ascending order; with this approach it is not required a trusted third party and provides cPIR. In this work is also introduced ExactNN, a mechanism that maps PoIs using a *voronoi tessellation* in a way that each voronoi cell contains exactly one PoI, a regular granularity squared grid is superposed and is privately queried by the user retrieving the PoIs contained in the voronoi cells that intersect the grid cell. For grid cells that are fully contained within a voronoi cell, it generates fake PoIs to match the number of bytes retrieved by cells with maximum intersected voronoi cells. This approach guarantees that the nearest neighbour is retrieved, however neither of the presented mechanism allows for KNN queries, rather a single relevant PoI is retrieved.

Olumofin et al. (2010) present a cryptography-based cPIR protocol designed to be used with any PIR technique according to the authors. This method adds spatial cloaking to reduce the database domain to be searched. The protocol consists on discretising the space in the form of a space granularity based on a Hilbert curve of the concentration of PoIs in an area, this is done in a way that each resulting granule will contain the same amount of PoIs. The number of PoIs is specified at setup and cannot be changed later without altering the database. The biggest area granule resulting from the calculation is set as the size of a cloaking region, a user can chose a bigger region consisting of more than one cell, which is later consulted with the chosen PIR mechanism to retrieve only the PoIs in the user Hilbert Cell. Olumofin et al. (2010) argues that this mechanism differs from *AproxNN* presented by Ghinita et al. (2008), in that this approach is specifically based on the 1997 computational PIR scheme by Kushilevitz et al. (1997).

Khoshgozaran et al. (2008) distinguish between *cryptography-based* and *hardware-based PIR*, where the first one utilises cryptographic transformation for the query and/or database structure, while the second requires special hardware architecture with a Secure Coprocessor (SC), that acts as a securely protected space where the retrieval of information takes places in a way that the LBS cannot decipher; the SC can be seen as a 'black box' third party that is embedded into the LBS server. The SPIRAL mechanism by Khoshgozaran et al. (2008) is hardware-based and uses random permutation of the database items with a mapping that is only stored in the SC, the SC also caches the items retrieved to a user to ensure that each item in the database is queried at most once and to avoid inferences from attackers or the LBS itself; when the cache in the SC becomes full a reshuffling of the entire database is performed, for this, they propose to generate offline reshuffled databases to avoid increasing computational costs, a downside of this method is that it does not support KNN search, but rather retrieves the i$^{th}$ item requested by users.

Papadopoulos et al. (2010) propose a more specific classification for PIR mechanisms than that of Khoshgozaran et al. (2008), stating that *"they can be grouped into: information theoretic, computational and secure hardware."* In their work is presented a hardware-based PIR which the authors call the AHG mechanism for KNN queries that aims to provide *'strong location privacy'*, where *"an adversary cannot distinguish the query location from any other location in the data space"*. Referring to a previously introduced concept, this aim translates into providing location privacy with a MUR of the entire spatial domain. For the AHG mechanism an architecture with *m* databases is proposed, where $m \geq 1$; this set of disjoint databases containing fixed-sized blocks, where each block is privately queried by the client using the hardware PIR protocol of Williams and Sion (2008); this architecture scheme echoes what Khoshgozaran et al. (2011) present as *'index structures'* for efficiently scanning a subset of records of an otherwise large and with prohibitive processing times database. The AHG approach ensures a processing time of 1 second even for large databases. For processing the private retrieval this mechanisms uses a superposed Hilbert Curve grid $G_h$ (as originally introduced by Khoshgozaran and Shahabi (2007)) over a regular granularity grid $G$, so that each cell in $G$ is mapped to a unique Hilbert value. Papadopoulos et al. (2010) argue that the mechanism presented by Khoshgozaran et al. (2011) guarantees that each query retrieval is completely private; however, the cardinality of the PIR requests per KNN queries may reveal information violating the scheme of strong location privacy.

#### 2.5.3.5 Comparison of LPPMs

Kounadi and Leitner (2015) propose a comparison table, which categorises *'geographical masks'* under groups, providing definitions for these categories and a masking degree that can be constant or variable. Zurbarán et al. (2014) proposes a comparison of mechanisms from various techniques included in this review.

For this thesis, the following table will serve to compare the presented LPPMs within the identified privacy techniques. The columns of the table describe the characteristics that are relevant to the author and are evaluated as follows: Type refers to the kind of LBS that the mechanism is applicable to; techniques are the privacy techniques described in this section; *Protected Region* assesses the scale of protection for each mechanism, it can take the values of: Entire Spatial Domain (ESD), city, neighbourhood or it could be variable; Third Party is understood as stated while describing the architectures for LPPMs (Item 3). The last column, indicates whether or not the LPPM requires modification on the LBS Architecture. This table differs from the one presented by Zurbarán et al. (2014) in the mechanisms included, privacy techniques covered and evaluated characteristics; however, it still follows the initial concept.

| LPPM | Type | Techniques | Protected Region | Third Party | Alters LBS |
|---|---|---|---|---|---|
| Chow et al. (2009) | All | K-Anonymity & Spatial Cloaking | Neighbourhood / Variable | Yes | Yes |
| Kalnis et al. (2007) | All | K-Anonymity & Spatial Cloaking | Neighbourhood | Yes | Yes |
| Cheng et al. (2006) | All | Spatial Cloaking | Neighbourhood | No | Yes |
| Monreale et al. (2013) | All | K-Anonymity & Spatial Cloaking | Neighbourhood | Yes | Yes |
| Ardagna et al. (2011) | All | Obfuscation | Neighbourhood / Variable | No | No |
| Wightman et al. (2011) | All | Obfuscation | Neighbourhood / Variable | No | No |
| Beresford and Stajano (2004) | Proactive | Pseudonyms | Variable | Yes | Yes |
| Zhong et al. (2007) Louis | Proactive | Cryptography-based | ESD | Yes | Yes |
| Lester | | | City | No | Yes |
| Pierre | | | ESD | No | Yes |
| Mascetti et al. (2011) C-Hide & Seek | Proactive | Cryptography-based | ESD | No | Yes |
| C-Hide & Hash | | | ESD | No | Yes |
| Wightman et al. (2012) | Proactive | Cryptography-based & Spatial Transformation | ESD | No | Yes |

| LPPM | Type | Techniques | Protected Region | Third Party | Alters LBS |
|---|---|---|---|---|---|
| Huang et al. (2008) | Reactive | Progressive Retrieval | Neighbourhood | No | No |
| Gong et al. (2010) | Reactive | Progressive Retrieval & K-Anonymity | Neighbourhood | Yes | No |
| Wang and Wang (2009) | Reactive | Progressive Retrieval & K-Anonymity | Neighbourhood | No | No |
| Riboni et al. (2011) | Reactive | Progressive Retrieval & K-Anonymity | Neighbourhood | No | No |
| Gruteser and Grunwald (2003) | Reactive | Temporal Cloaking & K-Anonymity | Neighbourhood | Yes | No |
| Kido et al. (2005) | Reactive | Dummy Queries | ESD | No | No |
| Quercia et al. (2011) | Reactive | Dummy Queries | City / Variable | No | Yes |
| Lu and Jensen (2008) | Reactive | Dummy Queries | City / Variable | No | Yes |
| Khoshgozaran and Shahabi (2007) | Reactive | PIR & Spatial Transformation | ESD | Yes | Yes |
| Ghinita et al. (2008) | Reactive | PIR & Spatial Transformation | ESD | No | Yes |
| Olumofin et al. (2010) | Reactive | PIR, Spatial Transformation & Spatial Cloaking | ESD | No | Yes |
| Khoshgozaran et al. (2008) | Reactive | PIR & Cryptography-based | ESD | Yes | Yes |
| Papadopoulos et al. (2010) | Reactive | PIR & Spatial Transformation | ESD | Yes | Yes |

Table 2.1: Comparison of LPPMs

## 2.6 Discussion

The Comparison of LPPMs table comprises the LPPMs included in the Privacy Techniques in LBS survey for this literature review. The characteristics evaluated for each LPPM have a desirable value from the author's perspective; ideally these would be as follows:

1. Type: All
2. Techniques: Any
3. Protected Region: ESD
4. Third Party: No
5. Alters LBS Architecture: No

Analysing the values on the table we find no match for such ideal model; however, there are mechanisms that do comply with most of them. Among these characteristics, avoiding the need for a third party or altering the LBS are critical for adoption, hence these will be prioritised. The mechanisms that are worth highlighting with the previous considerations are:

- Ardagna et al. (2011); Wightman et al. (2011), based on obfuscation for these can be applied to all types of LBS and do not require third parties or alterations;
- Huang et al. (2008); Wang and Wang (2009); Riboni et al. (2011), which use progressive retrieval with no alterations on the LBS or need of third parties. Wang and Wang (2009); Riboni et al. (2011) ensuring K-anonymity as well;
- Kido et al. (2005) based on dummy queries, providing a protected region of the ESD, but at the cost of too many requests from each user.

There are others where the Protected Region is of the ESD, but do require modification on the LBS. Mechanisms that allow this are presented in: (Zhong et al., 2007; Mascetti et al., 2011; Wightman et al., 2012) cryptography-based and (Ghinita et al., 2008; Olumofin et al., 2010) based on PIR.

With this canon in mind, the next chapter is dedicated to LPPMs developed by the author.

## 2.7 Summary

This chapter surveyed existing geoprivacy protection techniques and mechanisms; identifying their purposes aided in creating a taxonomy and to uncover the existing common characteristics to provide a ground for comparison between them.

# Chapter 3

# Real Time Location Privacy Protection Mechanisms

## 3.1 Introduction

As pointed out in the following quote:

> One of the main challenges for geoprivacy is to balance the benefit for an individual of participating on a geolocated application with the privacy risks he incurs by doing so.
> (Gambs et al., 2010, p. 34)

This chapter introduces LPPMs developed in this work to contribute to the thesis question and objectives. Specifically this chapter relates to Research Objective 2 in the Thesis Motivation: *Design LPPMs that are reliable, scalable, efficient, non-intrusive and that allow geospatial analysis*. The chapter has two main sections: one dedicated to noise-base location obfuscation, where a noise categorisation is explored, and the LPPMs developed using this technique. The other section is dedicated to a cPIR mechanism for reactive LBS.

The Noise-based section is partially based on the book chapter by Wightman and Zurbarán (2018) titled: *An Initial Evaluation of the Impact of Location Obfuscation Mechanisms on Geospatial Analysis*.

## 3.2 Noise-based Location Obfuscation

Noise-based location privacy techniques are one of the simplest ways to protect the exact location of the users. Its main characteristic is the induction of random noise to the original location obtained provided by a positioning system, in order to alter it permanently. Due to the randomness of the noise, it is very difficult to recover the original location, which becomes an useful attribute in terms of security.

It is a priority that a mechanism is able to provide location privacy without third parties or altering the LBS as seen in the Discussion of the previous chapter 2. Noise-based mechanisms pose important advantages: among them is the possibility to enable geospatial analysis, since the reported information is still in the form of geographic coordinates and they can be calculated on the user's device via simple calculations, thus computational complexity is low. In addition, the user can customise these mechanisms by defining the maximum amount of noise that will be applied to the locations; this range can go from 1 meter to a few kilometres, depending on the application's need for accuracy, and the perception of security that users have about their decision.

A naive approach to these techniques would be that a larger noise area would be beneficial in terms of security; however, it will degrade the quality of service for some applications. One example of the impact of the amount of noise are geomarketing services: If a user has a noise level of 3 kilometres, it may be reporting its location *nearby* a store in which he or she is not really close to; thus a geolocated offer could be lost if it is a time restricted one.

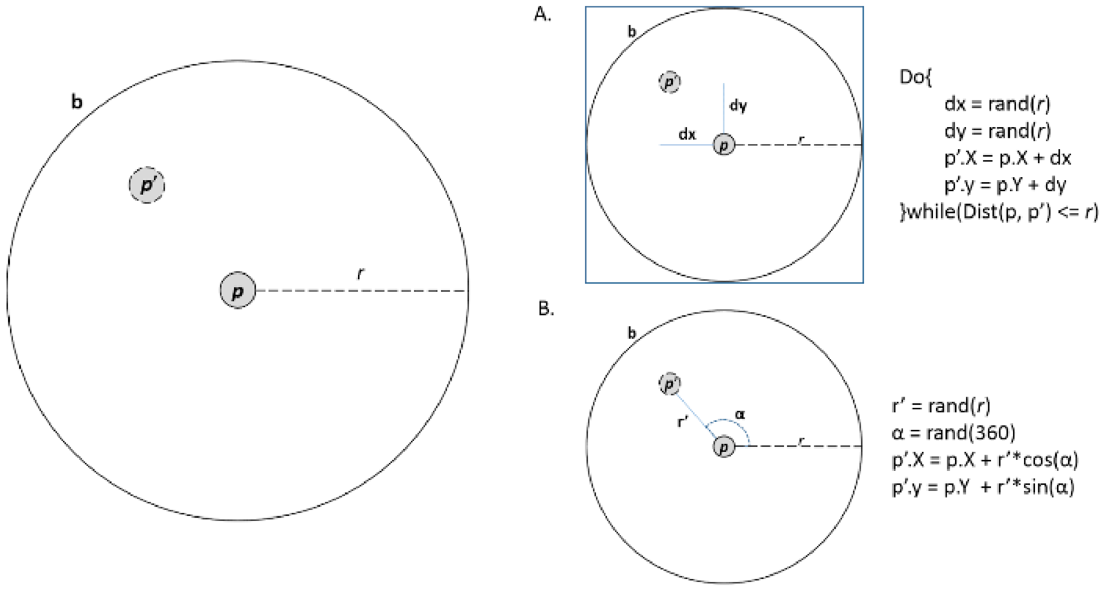
Several statistical random distributions can generate the values needed for inducing noise: Gaussian, Uniform, etc. The noise generation algorithm is the core element in these techniques. Three LPPMs will be presented in this section, all of them based on a uniform distribution; this is because it guarantees the highest variability on the data, which facilitates the generation of far away points from the centre compared to a Gaussian distribution.

### 3.2.1 Noise categorisation

Polar-generated points are distributed more evenly than cartesian-generated along the radius. Wightman and Zurbarán (2018) explore the generation of a single point inside an open ball in two different ways as shown in Figure 3.1. The first one is the cartesian method, in which a random number is generated for each component of the coordinate, added up and then the new point is verified to fall inside the circle because it can happen that it falls on the external areas on the square area of $(2r)^2$.

The second one is the polar method, in which a distance and an angle are generated, transformed into cartesian coordinates and added to the original point p. This technique does not require validation because the random distance is validated to be between 0 and r; this ensures that it cannot fall outside the circle.

Despite the fact that both generate a single point, the distribution of the random points shows



```
Do{
    dx = rand(r)
    dy = rand(r)
    p'.X = p.X + dx
    p'.y = p.Y + dy
}while(Dist(p, p') <= r)
```

```
r' = rand(r)
α = rand(360)
p'.X = p.X + r'*cos(α)
p'.y = p.Y + r'*sin(α)
```

Figure 3.1: Random generation of a point

differences between these techniques which may affect the impact of the algorithm. Figure 3.2 show the graphical distribution of 500 points using the cartesian and the polar approaches.

Even though cartesian-generated points look more evenly distributed along the area, without a concentration of points, this impression is false: Polar-generated points are distributed more evenly along the radius. The fact that there seems to be more points closer to the centre is due to the symbol used to represent points, giving the impression of closeness, when in fact is a more uniform distribution, as it can be appreciated in Figure 3.3; where the distance distribution (in metres) is biased on the cartesian approach.

The main advantage of this technique is that it is very simple to implement. However, it can generate values very close to the centre, which is not entirely desirable because the protection for that specific point with no difference from the obfuscated one will be practically null. An improvement that includes the generation of N random points, guarantees an increase in the resulting induced noise, see (Wightman et al., 2011) for more details of this improvement (N-Rand).

### 3.2.2 $\theta-$Rand

This mechanism is inspired on the N-Rand mechanism presented by Wightman et al. (2011) and included in the literature review under Location Obfuscation. In $\theta-$Rand, each reported location is treated individually and independently from the others. The following parameters must be predefined by the user in order to apply the algorithm: $r_{max}$ defines the maximum

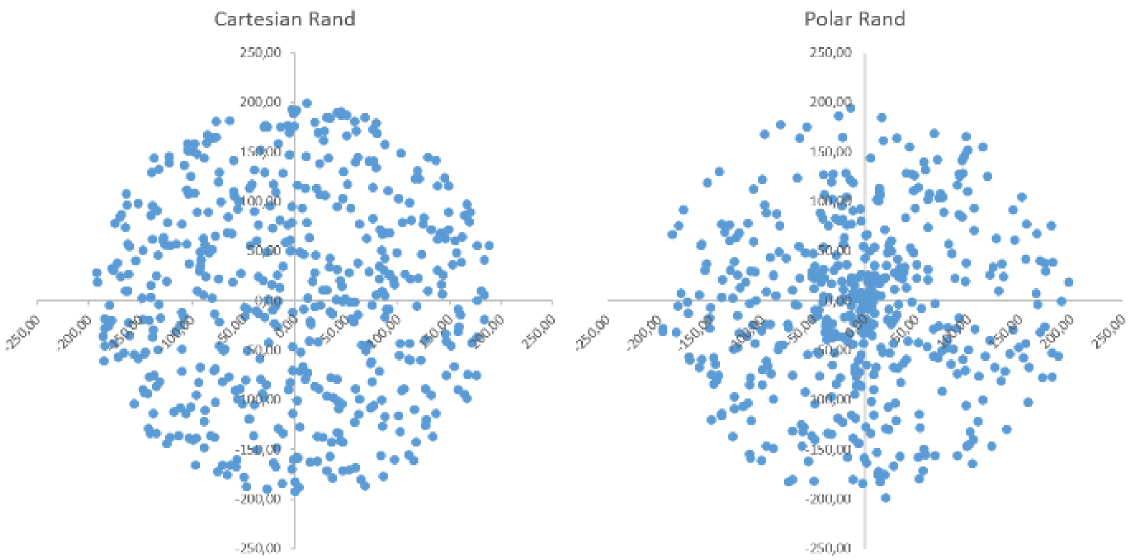


Figure 3.2: Cartesian and polar generation of 500 points

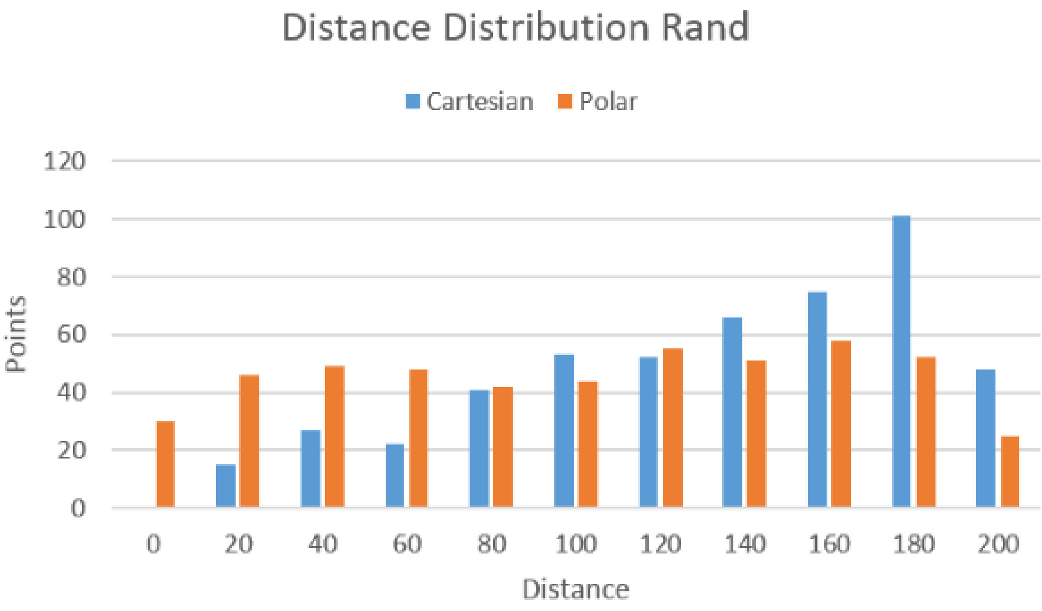


Figure 3.3: Distance distribution of cartesian and polar generation of 500 points

distance that an altered point can be from the reported original location, and *n*, which defines how many trials will be performed per obfuscation in order to increase the average distance of the output points from the real one. According to Wightman et al. (2011), a good rule-of-thumb number is to define *n* as 4, in order to maintain some variability in the average distance of the obfuscated points. This value for *n* will be considered for future noise-based obfuscation algorithms.

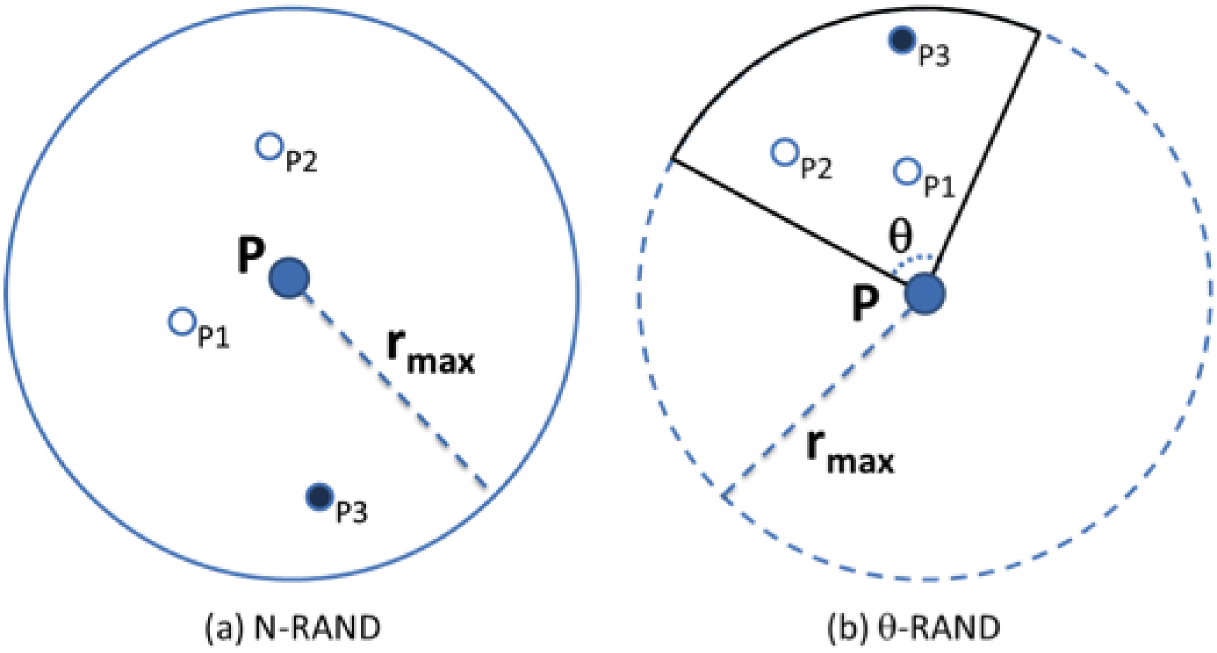


Figure 3.4: Graphical example of the $\theta$-Rand algorithm

In order to apply the $\theta-$Rand mechanism, the following steps should be followed:

1. Create a random circle sector $s$, by generating a $\theta$ angle and using the $r_{max}$ radius.
2. The $\theta$ angle is calculated by the difference between 2 angles: $\theta_i$ between $[0, 180°]$, and $\theta_i + \theta_f$, where $\theta_f$ is also between $[0, 180°]$.
3. Generate $n$ points inside the circle sector $s$. In order to optimise computational time, a point can be generated in terms of polar coordinates, generating a random distance less than $r_{max}$, centred on the original location and an angle between $\theta_i$ and $\theta_f$.
4. Select the most distant point from the original point P, calculate the respective cartesian coordinates, and report the obfuscated point instead of the original one.

In order to evaluate the resistance against filtering attacks, the Exponential Moving Average filtering approach is proposed[1]; this technique is used in time series analysis in order to smooth noise data with an underlying trend. When compared to N-Rand using synthetically generated paths; $\theta$-Rand proved to be more resistant to filtering, even though the induced noise by N-Rand was greater.

Due to its non symmetrical dominion for random point generation, this LPPM can withstand noise filtering attacks from exponential moving average techniques. This suggests that a uniform distribution is not always desirable.

Next subsection will present a mechanism inspired in a pinwheel form, providing a different distribution of noise.

[1] (Wightman et al., 2011, p. 102)

### 3.2.3 Pinwheel

Pinwheel was first introduced in Wightman et al. (2014), it is a mechanism inspired in the form of pinwheel vanes; it uses a function to determine the added noise producing a distribution pinwheel-like.

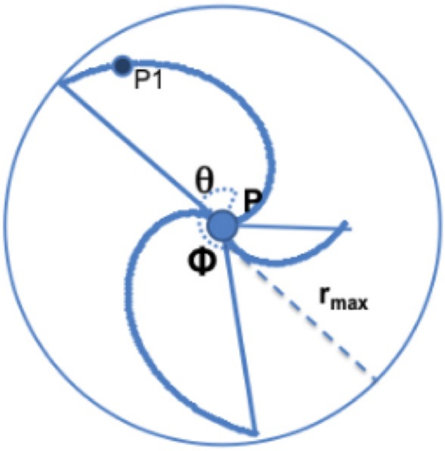


Figure 3.5: Graphical example of the Pinwheel algorithm

Figure 3.5 presents a graphical representation of how the Pinwheel algorithm works; where P is the is the original location point at the centre of the circumference, $r_{max}$ defines the radius and $\phi$ defines the period for the repetition of the vanes that form the pinwheel. Each of which at a given $\theta$ outputs the corresponding radius.

$$r(\theta) = (\theta mod \phi)/\phi + r_{max} \tag{3.1}$$

The noise added is calculated by defining a maximum radius, which is the maximum acceptable noise to induce. The maximum radius serves to limit the dominium of the resulting obfuscated location, but this dominium is also determined by the pinwheel formula in Equation 3.1; where given a $\theta$ value, defines a specific radius for each $\theta$ within the circle described by the circumference. This makes the selection of the radius a deterministic process in the generation of random points with low processing cost for doing on the fly transformations even on mobile devices, the formula used to calculate $r(\theta)$ is applied using polar coordinates.

The $\phi$ value has a great impact on how the random points are distributed. Figure 3.6 show the dominium of the random point generator with $\phi = 12^\circ$ and $\phi = 110^\circ$, respectively. In the first case, high periodicity and a symmetric distribution shows a distribution very similar to a uniform one, with the polar generation approach. The second case shows a much lower periodicity and very restricted areas for the generation of points, which is also asymmetrical, increasing the probability of generating points towards a sector of the circumference, which proved to perform better against filtering attacks in (Wightman et al., 2014).

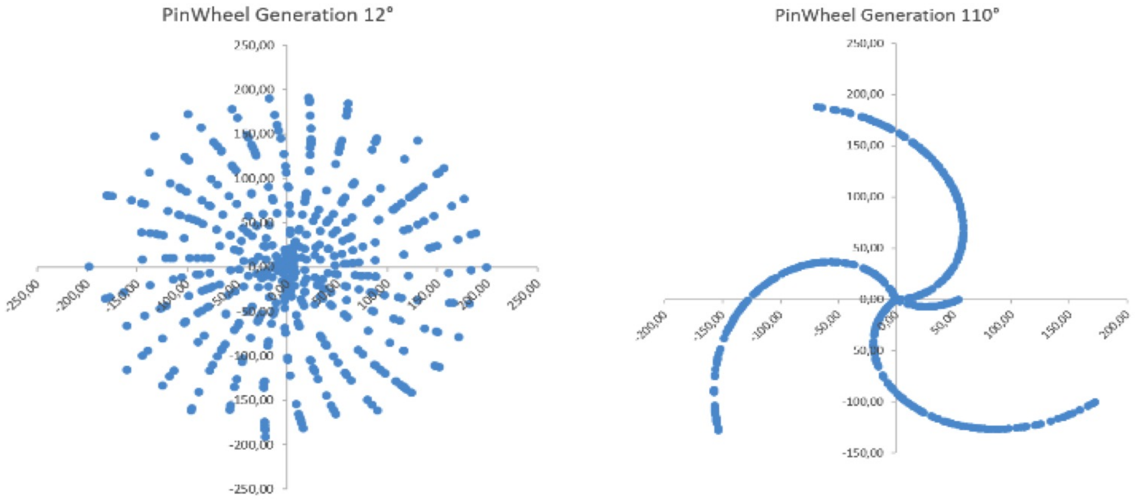


Figure 3.6: Pinwheel outputs with different $\phi$ angles

### 3.2.4 Near-Rand

Near-Rand is a noise-based algorithm for location obfuscation by Zurbaran et al. (2014). This mechanism obfuscates the reported location by producing a point near the original location, differing from its predecessors in that this mechanism is not limited to a circular dominium. Rather, the obfuscated location is calculated using the mean coordinates of $k$ nearest randomly generated points or *neighbours*; these neighbours are uniformly distributed within a *coverage area* delimited by the user, i.e. the city where the user is at. Figure 3.7 depicts a sample neighbour set over a squared area of 100 $km^2$ in the city of Barranquilla.

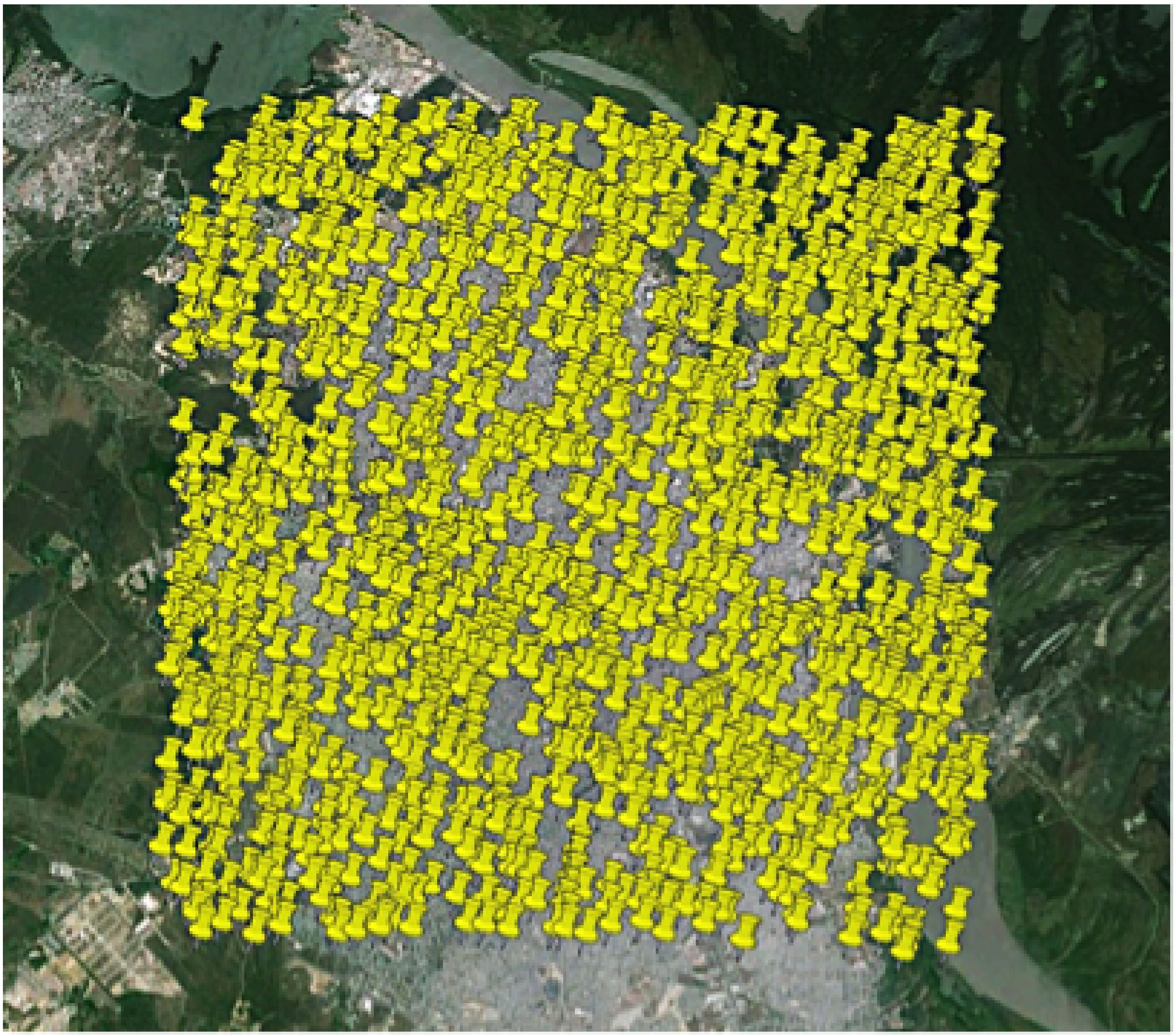


Figure 3.7: Uniformly distributed points for Near-Rand algorithm

The neighbour set is calculated in the user device and is locally stored, if a user exits the coverage area i.e., travels to another city, a new one will have to be generated.

To ensure an average distance between the generated points, it is used the Equation 3.2 from (Labrador and M. Wightman, 2009, p. 75), which is useful in wireless sensor networks to calculate the Critical Transmission Range (CTR); in the equation $r$ is the minimum homogeneous transmission range of every node. For the purpose of this algorithm, $r$, represents the desired distance between a pair of points in the specified coverage area. A small $r$ will produce a high density point set, which will generate closer neighbours to a reported location; on the other hand, a lower density generates neighbours far from the original location.

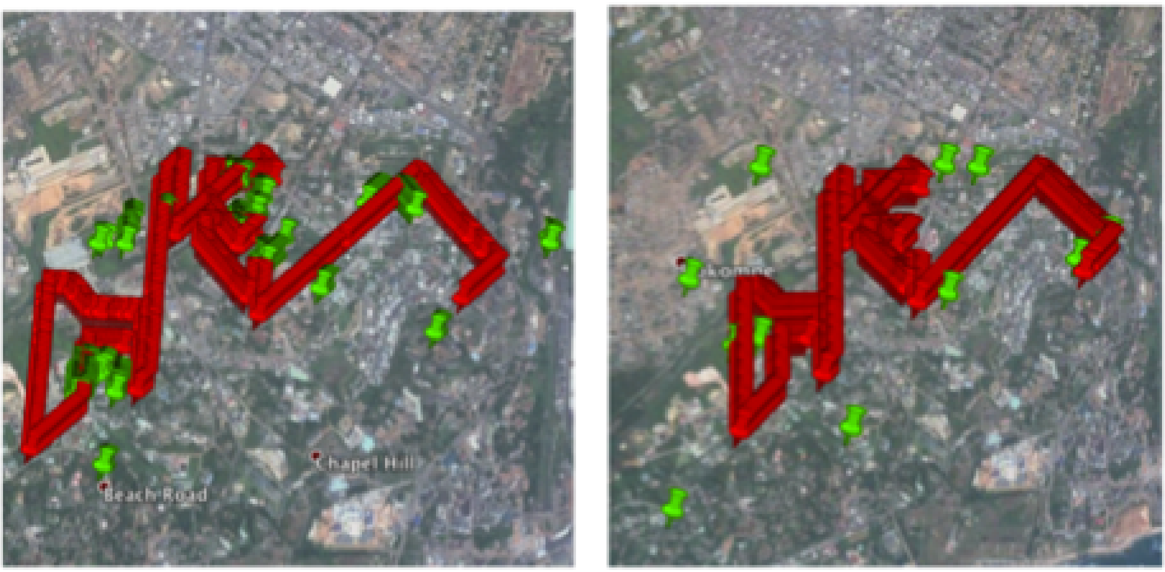
$$r = \frac{\sqrt{(log(n) + f(n))}}{n * \pi} \tag{3.2}$$

Where $f(n) = log(log(n))$, hence:

$$r = \frac{\sqrt{(log(n) + log(log(n)))}}{n * \pi} \tag{3.3}$$

The following are the algorithm steps for obfuscating a location with Near-Rand:

1. Fix an approximate value for $r$ in Equation 3.3, then iteratively evaluate the function to obtain the corresponding $n$ to ensure the desired distance $r$.
2. Set $k$ as the number of neighbours to include in the calculation of the mean coordinates; a large $k$ value increases the induced noise, especially with a low point density.
3. Calculate the mean coordinates of the $k$ nearest neighbours from the reported location.



a) Near-Rand including the original location b) Near-Rand without the original location.

Figure 3.8: Obfuscated synthetic path with the Near-Rand algorithm variations

Figure 3.8 shows a comparison of two possible variations of Near-Rand; when the original user location is included in the mean coordinates calculation, and when only the synthetically generated neighbours are included. The original path in both cases is shown in red with the obfuscated locations in green. The variation that does not include the original user location proved to induce more noise, hence more location privacy.

This mechanism can include a non uniform distribution for the synthetically generated points, in a way that the user can specify sensitive zones to have lower point density i.e., a higher $r$, which results in amplified added noise and more protection in those zones. An approach of how the generated neighbours for Near-Rand would behave in such case is shown in Figure 3.9.

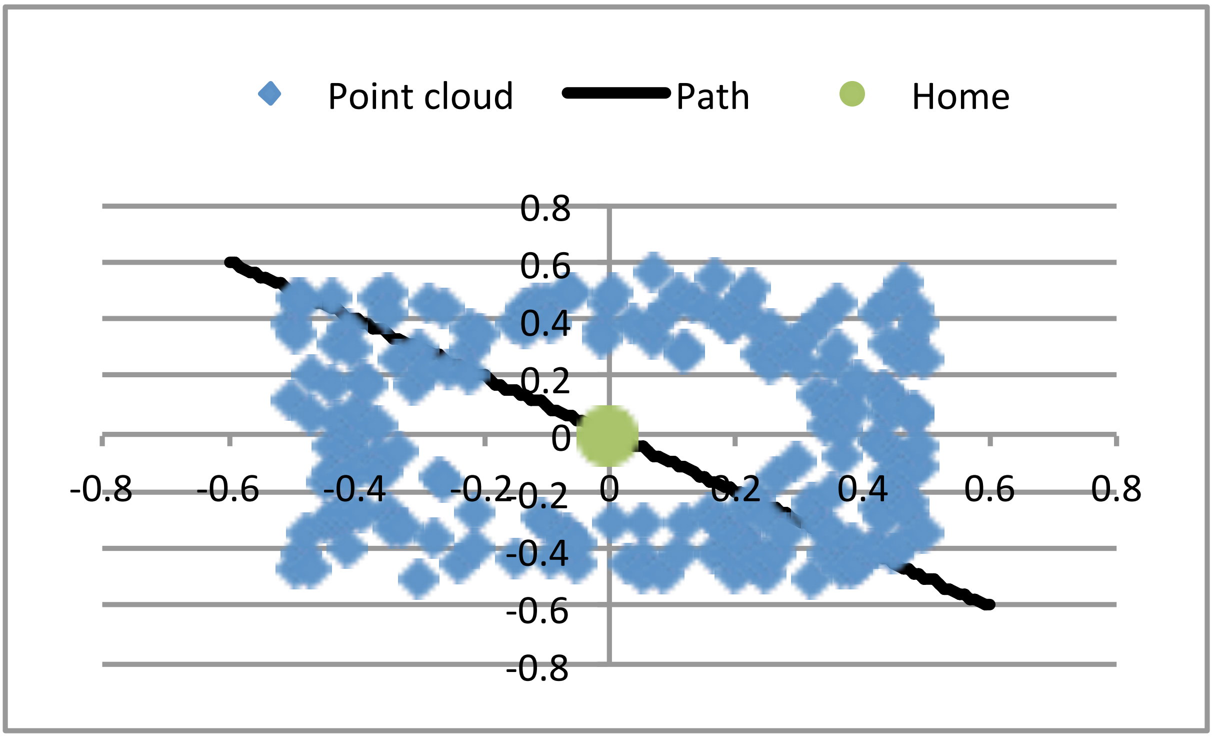


Figure 3.9: Non uniform neighbour distribution for differential privacy on sensitive locations

The use of a Poisson distribution with different $\lambda$ values is explored, for this, the coverage area is divided into cells from which the user can select the ones that require more privacy. The $\lambda$ value is fixed according to the desired privacy in a cell using Equation 3.4 for the distribution. Figure 3.10 shows a sensitive cell delimited by the pink pins and a location reported within it in green.

$$Pr(X = k) = (\lambda^k e^{-\lambda})/k! \tag{3.4}$$

Near-Rand is a flexible algorithm that can provide obfuscated locations with differential privacy according to a user preferences using different settings. The induced noise depends on the neighbour set and the specified $k$; in the performed experiments a value of 3 was used, producing approximately 200 m of induced noise with a uniform distribution. See Zurbaran et al. (2014) for more details on the experiments.

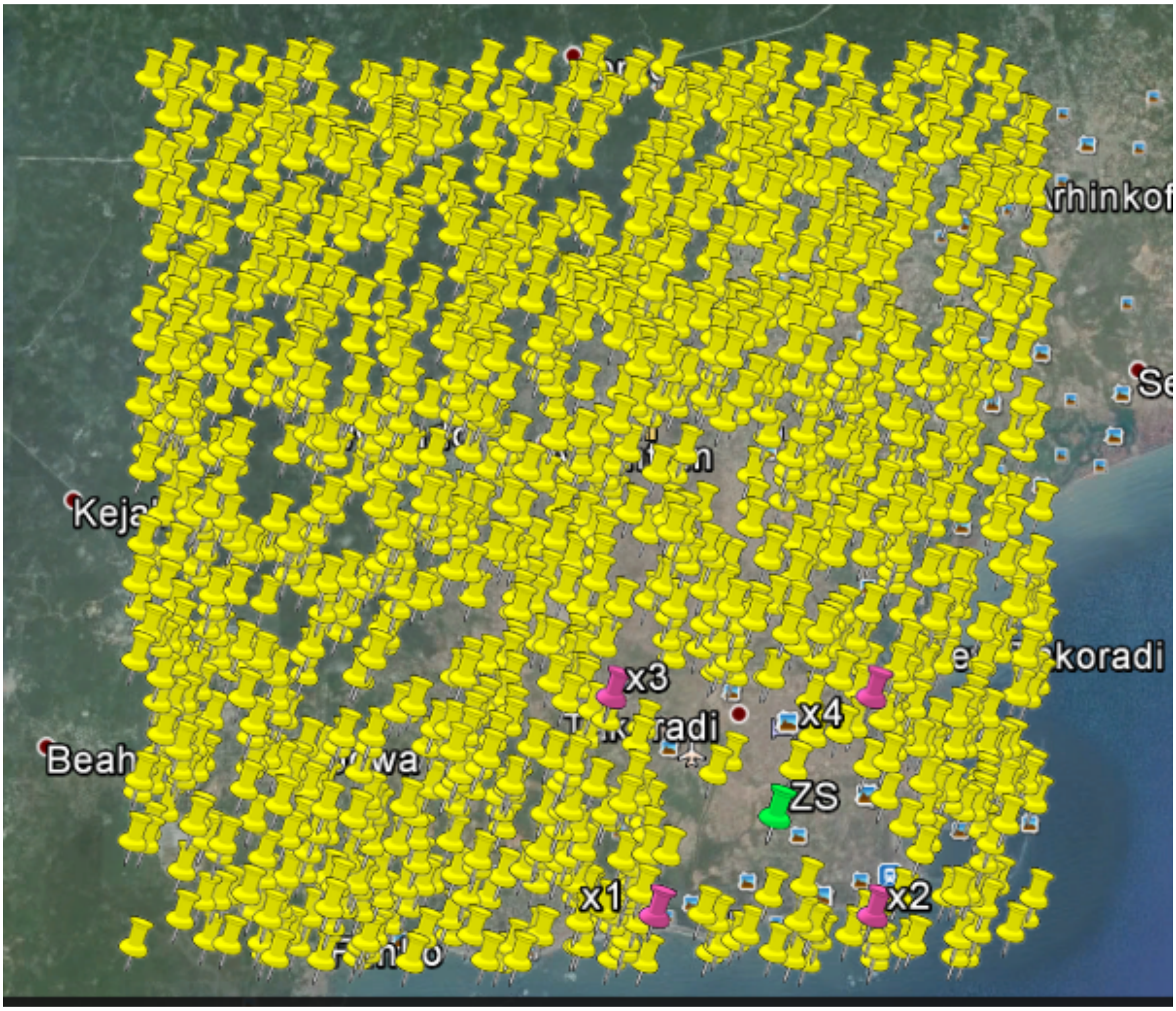


Figure 3.10: Non uniform neighbour distribution for sensitive locations

## 3.3 MaPIR: Mapping-Based Private Information Retrieval

MaPIR is a cPIR LPPM (see subsection 2.5.3.4 Private Information Retrieval (PIR)) that uses cryptography. This algorithm was presented by Wightman et al. (2013) and this section is partly based on this article.

This mechanism is based on a squared grid granularity with redundant identification for the cells. The IDs can be calculated with simple arithmetic operations on both, the mobile device and server; these are a reduced transformation of geographic coordinates that have values in the range from 1 to 10, so the ID alone does not reveal any location information.

The mapping function for MaPIR works in two phases: *Location Reduction* and *ID calculation*. The Location Reduction phase takes as parameter geographic coordinates and applies a technique to transform latitude and longitude into numbers that still contain information about location, without revealing it. In addition, this reduction technique should also preserve the

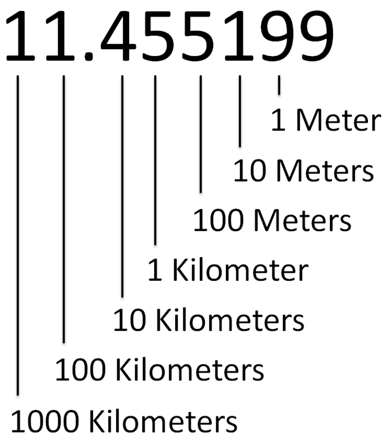


Figure 3.11: Reduction based on scale

geographical relation between nearby PoIs and should be able to manage different scales, e.g., hundreds of meters, kilometres, tens of kilometres, etc., in order to adapt the searching area to the user's privacy requirements while enabling NN search.

The second phase is the ID calculation, where the reduced location numbers serve to generate the ID of the cell containing the original location. The function for ID calculation must provide a non-unique identification of the grid cells, such that they are sparsely and randomly distributed.

The proposed Location Reduction function in this work is the following:

1. *Kth Digit Location Reduction*: This is based on a simple idea; approximately, $9e-6$ degrees is equivalent to a meter in the Equator (360 degrees by 40,400 kilometres). Thus, each significant digit on the location represents a scale of the distance, like shown in Figure 3.11. For example, the following two latitudes 11.455199 and 11.455891 produce the same reduction number when $k = 2$, this is 5, thus they are approximately within a kilometre from each other. Formally, the proposed function is the use of the $k$th least significant digit from the decimal portion of the location information, in order to reduce this information. The main advantages of this technique are: It is very simple to calculate; the reduced values are always in the same range [0-9]; and it can be adapted to different search scales, without increasing the complexity.
2. *Pseudo-random ID Calculation*: Once both latitude and longitude have been reduced, these values are used to define the ID of the grid cell in which the point is located. Let the reduced values of latitude and longitude be $i$ and $j$, and let $p$ be the next prime value after the maximum value that the reduction function can generate. Then, the ID of the grid cell where the reduced point is located is calculated based on Equation 3.5.

$$ID_{(i,j)} = \lceil (i+1)*(j+1) mod p \rceil \tag{3.5}$$

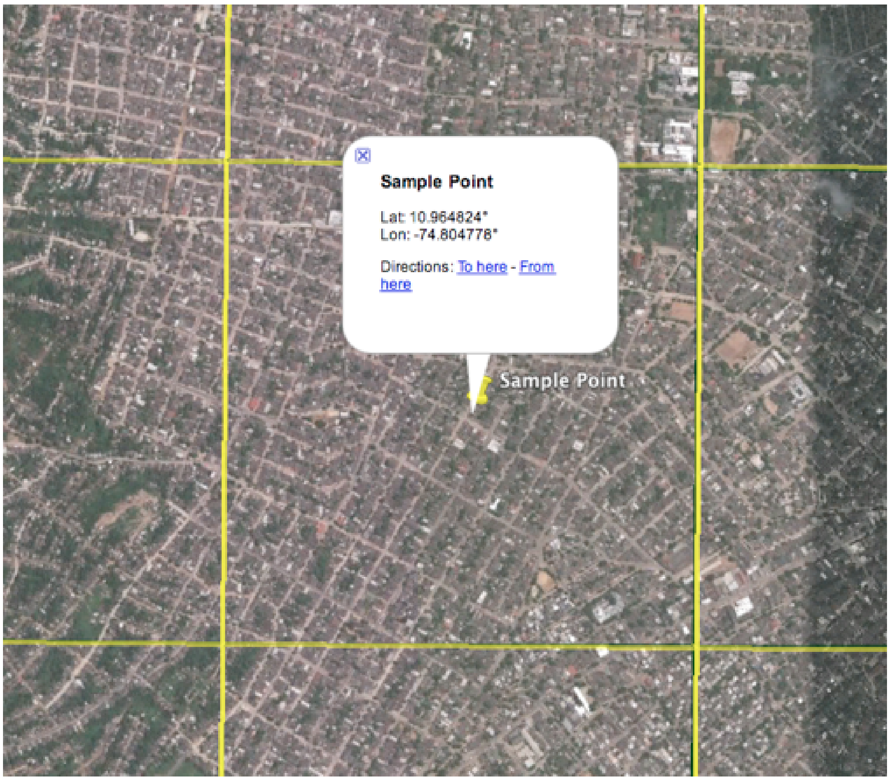
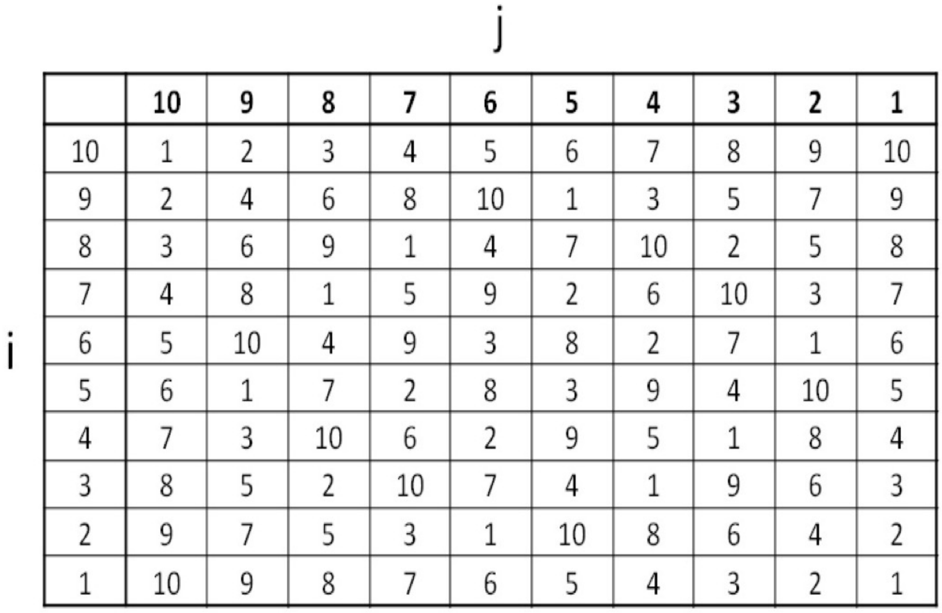


Figure 3.12: Example of a PoI within a grid where $k = 2$ on Google Earth and a POI within a cell where $ID = 7$

For the $k$th digit location reduction function, the $p$ variable should be 11 as it is the next prime number after 9, which is the maximum value that the location reduction function can generate. The addition of 1 to the reduced components of the coordinates ($i$ and $j$) is necessary in order to avoid the value 0, which would nullify the multiplication. Figure 3.12 shows a sample mapped geographic location into its cell ID; furthermore, the final grid for this mapping function is shown in Figure 3.13 for all possible values of $i$ and $j$.

Each ID exists $p-1$ times in a single grid (10 in this case). This implies that the resultant grid cell of a single location point queried by a user will be indistinguishable from other $p-2$ in the same grid, so there is not a unique geographical correspondence. The $p$ value must be a prime in order to guarantee that each id exists on each row; otherwise, all ID values multiples of $p$ will generate redundant patterns in a single row.

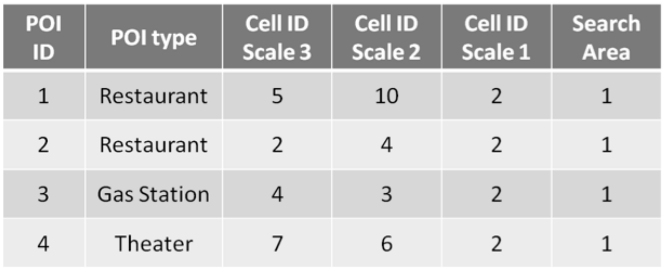
The overall MaPIR process involves both user and server; the user's mobile device obtains her location through a positioning system and translates it into the corresponding MaPIR ID, for this it is required that the user defines the scale that will be used; for an average city area, a value of $k = 2$ is recommended. As for the server side, the LBS provider must also translate the database of PoIs in a pre-processing stage using the mapping function for the offered scales; this in order to minimise query processing time.



| | 10 | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 |
|---|---|---|---|---|---|---|---|---|---|---|
| 10 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 9 | 2 | 4 | 6 | 8 | 10 | 1 | 3 | 5 | 7 | 9 |
| 8 | 3 | 6 | 9 | 1 | 4 | 7 | 10 | 2 | 5 | 8 |
| 7 | 4 | 8 | 1 | 5 | 9 | 2 | 6 | 10 | 3 | 7 |
| 6 | 5 | 10 | 4 | 9 | 3 | 8 | 2 | 7 | 1 | 6 |
| 5 | 6 | 1 | 7 | 2 | 8 | 3 | 9 | 4 | 10 | 5 |
| 4 | 7 | 3 | 10 | 6 | 2 | 9 | 5 | 1 | 8 | 4 |
| 3 | 8 | 5 | 2 | 10 | 7 | 4 | 1 | 9 | 6 | 3 |
| 2 | 9 | 7 | 5 | 3 | 1 | 10 | 8 | 6 | 4 | 2 |
| 1 | 10 | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 |

Figure 3.13: MaPIR granularity grid IDs

The LBS provider should identify the geographical areas where the service will be available in order to calculate the grids for those areas. Every time there is a new PoI added to the system, this process has to be performed just for the new location without affecting the rest of the points. The mapping must be restored completely only if the mapping technique itself is altered. PoIs are stored in the database along with their computed IDs and type of interest, Figure 3.14 depicts a sample view of the LBS database. A server side implementation for PostgreSQL using the PostGIS extension is available in (Zurbarán et al., 2018, Chapter. 12).



| POI ID | POI type | Cell ID Scale 3 | Cell ID Scale 2 | Cell ID Scale 1 | Search Area |
|---|---|---|---|---|---|
| 1 | Restaurant | 5 | 10 | 2 | 1 |
| 2 | Restaurant | 2 | 4 | 2 | 1 |
| 3 | Gas Station | 4 | 3 | 2 | 1 |
| 4 | Theater | 7 | 6 | 2 | 1 |

Figure 3.14: MaPIR granularity grid IDs

When a NN query is issued, the user's device sends it's calculated cell ID and a query predicate, e.g., restaurant, gas station, hotel, etc., in order to filter undesired PoIs to lower communication costs while allowing the LBS to get statistics on their user's interests; this predicate is not considered private. The server then retrieves all PoIs matching the requested ID and query predicate, resulting in an increased number of retrieved PoIs that act as dummy results which are later filtered by the user's device in order to present only relevant results, i.e. the ones that actually belong to the cell where the user is located. Figure 3.15 shows a visual output of a MaPIR request for a user within a cell ID of 6.

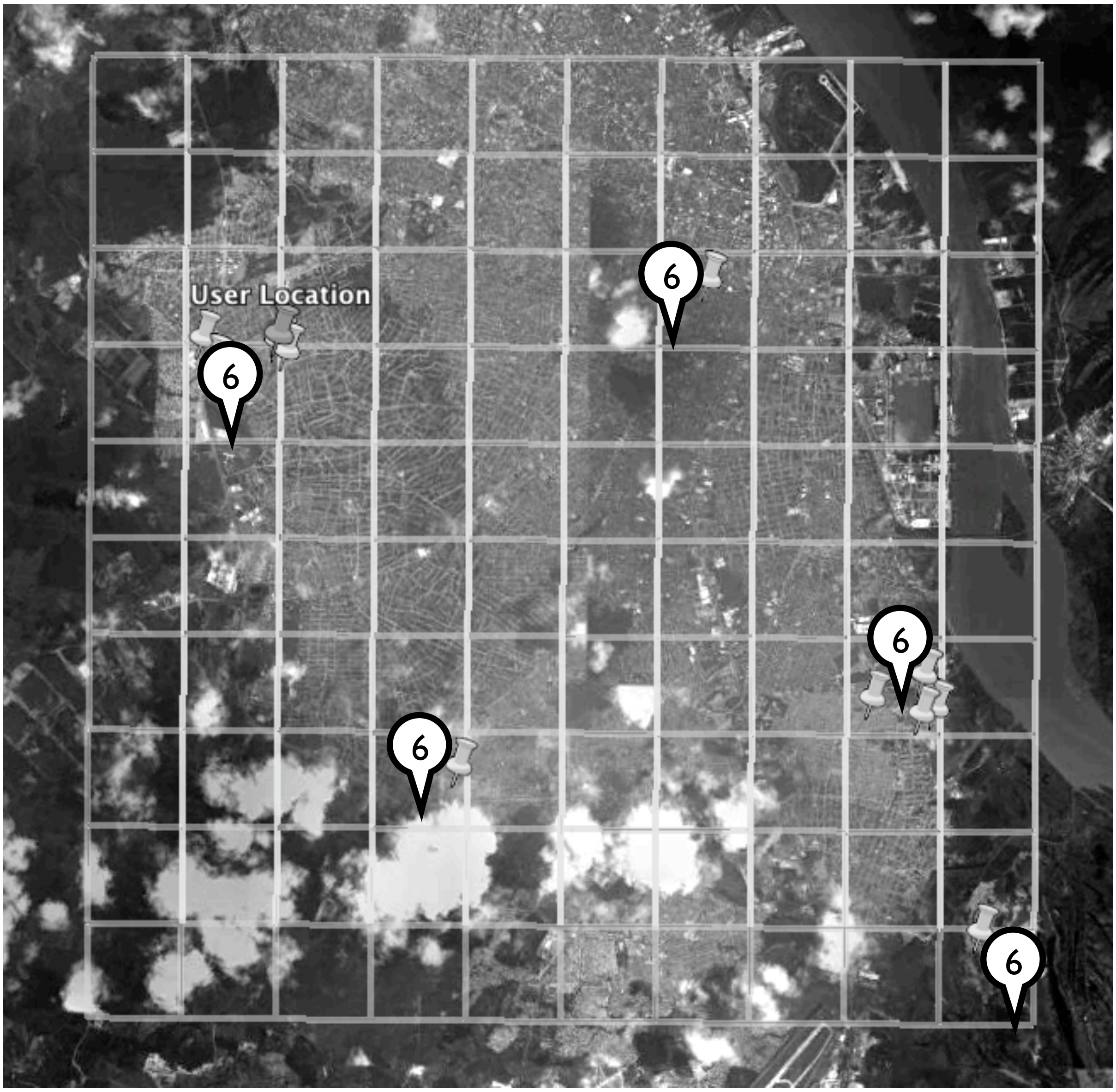


Figure 3.15: Example of a MaPIR request where $ID = 6$

The MaPIR mechanism enforces location privacy and allows querying PoIs with great accuracy, in an efficient manner, and without revealing the actual location of the users. Performance tests proved MaPIR query processing time to be more efficient than when using dummy queries and even when compared to a regular spatial query [2].

[2] See (Wightman et al., 2013, p. 969) for performance test details

## 3.4 Discussion

With noise-based techniques, once the user location is altered, the original data is lost. This means that the service provider will receive only a probabilistic hint of where the user was, which, depending on the induced noise, will have value for geospatial analysis.

Noise-base obfuscation techniques are susceptible to filtering attacks, when the Exponential Moving Average was used to filter noise, it proved that a uniform noise distribution, i.e., the one used in Rand and N-Rand, is more likely to be filtered than when the induced noise is somehow biased; therefore the need for creating non-uniform noise-based mechanisms like the ones presented in this chapter: $\theta$-Rand, Pinwheel and Near-Rand.

The MaPIR mechanism is based on cPIR and is applicable to reactive LBS; it requires a mapping function to be implemented in both, user's device and server.

The following is an extension of the table presented in Section 2.5.3.5 that compares the LPPMs presented in this chapter through characteristics based on concepts presented on Chapter 2.

| LPPM | Type | Techniques | Protected Region | Third Party | Alters LBS |
|---|---|---|---|---|---|
| Wightman et al. (2011) | All | Obfuscation | Neighbourhood / Variable | No | No |
| Wightman et al. (2014) | All | Obfuscation | Neighbourhood / Variable | No | No |
| Zurbaran et al. (2014) | All | Obfuscation | Neighbourhood / Variable | No | No |
| Wightman et al. (2013) | Reactive | PIR | City / Variable | No | Yes |

Table 3.1: Comparison of LPPMs

A Protected Region in Table 3.1 is said to be *variable* when the LPPM is able to adapt to different scales depending on the LBS requirements. In the case of proactive LBS, obfuscation mechanisms should use a small amount of noise[3], hence a neighbourhood sized area is suggested; but for reactive LBS, the location distortion could be greater. The amount of induced noise in location obfuscation will be further addressed on next chapters as well as the usability of the obfuscated data for geospatial analysis.

As for the estimated Protected Region of MaPIR, it is said to be variable due to the possibility of applying this mechanism on different scales; however, the use of a city scale is recommended

[3] As explored by Seidl et al. (2015), *"In Gaussian random perturbation tested in France, personal residences are identifiable with a standard deviation of 50 m, but are indiscernible beyond 200m according to Gambs et al. (2010). Given these considerations, the radii implemented here are 30, 100 and 250m."*

i.e. when $k = 2$, for communication costs might be prohibitive in a bigger Protected Region. In the scope of this work it is not contemplated the use of geographic information protected by MaPIR or any other PIR-based mechanism for geospatial analysis.

## 3.5 Chapter Summary

This chapter has introduced geoprivacy mechanisms developed in this work to be implemented while the LBS is offering its functionalities. Real-time LPPMs behave as if the LBS provider was untrusted and requires some knowledge of the LBS application in order to parametrise adequately the induce perturbation. Different mechanisms were introduced belonging to the Location Obfuscation techniques and a PIR mechanism. This chapter concludes that noise-based location obfuscation are less intrusive; however the MaPIR mechanism provides higher privacy protection.

# Chapter 4

# Location Privacy Protection Mechanisms for Third Party Information Release

## 4.1 Introduction

This chapter introduces LPPMs developed in this work to contribute to the thesis question and objectives; specifically, this chapter relates to Research Objective 2 in the Thesis Motivation: *Design LPPMs that are reliable, scalable, efficient, non-intrusive and that allow geospatial analysis*, emphasising in the utility of protected data for geospatial analysis.

It could be argued that LPPMs can be applied at two different moments: when the user is reporting the data to the LBS and after the data has been transmitted and stored, and privacy becomes only a concern when data leaves the service provider. In the first case, such mechanisms are referred in this work as *Real Time* LPPMs and are the ones explored so far. On the other hand, there are *LPPMs for Third Party Information Release* and under this model, the LBS is considered trusted.

It is not incorrect to assume a model where the LBS is trusted, since under current regulations a key element of the business model of social media platforms is to sell data from their users to third party companies. Therefore, the LBS have unrestricted access to the original information.

Even though it is expected that information released through a LBS API is already distorted or aggregated, the format of the information is not known by the public; so there is no certainty from the user's point of view of how much information is being revealed, despite the final use of the information e.g., marketing, research, government planning, etc. For these cases, there is a lack of specific legislation on location information, being this just regarded as general personal information.

The mechanisms introduced in this chapter aim to provide location privacy to the data, once the trusted service provider has it and before it is shared with third party actors. The objective for these mechanisms is to distort data in such a way that individual's location privacy is preserved, while maintaining the value of the data in terms of data representation, since this data can be used for research or commercial purposes.

## 4.2 VoKA: Voronoi K-Aggregation Mechanism

The Voronoi-based K-Aggregation technique (VoKA) was introduced by Zurbaran and Wightman (2017) and this section is partly based on this paper. VoKA is an LPPMs for Third Party Information Release that allows to group users reported locations based on density and not just by area; giving a much better idea of the real distribution of the data, while preserving a level of K-anonymity among the users.

Voka is an offline mechanism designed to aggregate datasets of geolocated data in a way that protects isolated sources from individualisation based on location. When aggregating datasets there are 4 main types of data geolocated sources, based on the number of entries and the number of nearby sources: popular areas, isolated points of interest, low density living areas and isolated living areas as depicted in Figure 4.1.

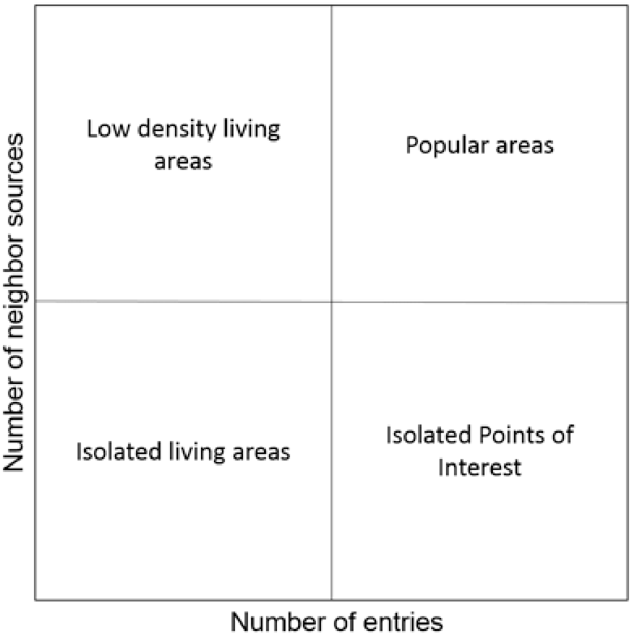


Figure 4.1: Identification of data sources based on number of entries and number of neighbours

- *Popular areas* are characterised by a large number of total entries and a large number of geolocated sources. This area is usually located in the city centres or in commercial districts, with many PoIs like restaurants, museum, etc.

- *Isolated points* of interest represent specific places that attract a large number of people and that generate a large number of entries, i.e. PoIs located in the outskirts of a city, an event like a concert, etc.
- *Low density* living areas represent residential districts of the city, in which people are less likely to generate geolocated entries, but there are a large number of sources.
- Finally, *isolated living areas* represent especial residential districts or low traffic points of interest, probably in rural areas, in which there is a small number of entries and not many different sources.

The last two being most critical of the geolocated sources because identification of the individuals, over time, may be fairly easy if the information is linked with other public information.

The VoKA algorithm will aggregate nearby sources with low number of entries, in order to generalise their location and merge the entries to a minimum amount that will reduce the probability of individualisation. This technique is designed to work *offline* i.e, after data collection, on a trusted service provider, and before the data is sent to a third party; either as open access, for commercial or for research purposes.

The first two variables that the algorithm requires to define before aggregating the data are: the number of entries that will identify an acceptable level of anonymity, K, and the maximum distance that will determine if two sources are neighbours, D. In general, there are no magic numbers for these variables, and they somewhat depend on the type of data, the city, and other parameters. However, in this work, we propose that the minimum number of entries, K, could be defined based on Equation 4.1:

$$K = max(P_{50}, K_{min}) \tag{4.1}$$

where $P_{50}$ represents the number of entries at the 50th percentile of the geolocated sources, and $K_{min}$ is a value defined by the administrator of the service provider. In this work we proposed using a $K_{min} = 20$, so the algorithm will try to ensure at least 20 entries per aggregated feature or that the feature is like the top 50% of the sources.

This algorithm is divided into two stages:

1. *Grid-Reduction*: The coordinates of each location are approximated to a certain number of decimal digits, e.g., 5, 4 or 3. This provides a first level of aggregation and privacy protection, by merging the location entries on a grid-based distribution. This approximation is similar to the one used by MaPIR [1]. However, this kind of aggregation can

[1] See Section 3.3 MaPIR: Mapping-Based Private Information Retrieval.

affect location information in different scales according to the geographic area where the location entries where generated. In other words, if the dataset is nearby the Equator, the scale presented for MaPIR in Figure 3.11 remains relevant; otherwise, the distance of the aggregated locations from original ones will depend on the Universal Transverse Mercator (UTM) zone where the data originated. i.e.; in locations around Milano, Italy, each decimal digit approximate a distance magnitude: the 5th decimal digit represents 70 cm; the 4th, 7 m; the 3rd, 70 m, etc.

2. *KD-Aggregation*: The second step starts by:
   (a) Sorting the geolocated sources by their number of entries, from smaller to largest, then by latitude and finally, by longitude.
   (b) Then, a greedy dominating set algorithm is applied over the list of points with less than K entries; in the beginning, all will start as not visited and not dominated: for each non visited location L, the algorithm will mark it as visited and look for other non visited locations within distance D (see Figure 4.3a).
   (c) Points within that distance D are evaluated with the Delaunay triangulation to identify direct neighbours, only those will be added to a list of closest neighbours and marked as visited and dominated (see Figure 4.3b).
   (d) This process is repeated until all nodes are visited (see Figure 4.3c). In this way, each visited node will represent a local Voronoi neighbourhood that does not extend farther that a distance D, increasing the anonymity of the entries and generalising the area.

   Given that only areas with isolated points are merged, the ones that already comply with the defined K-anonymity setting, remain the same as in original data. As seen in Figure 4.3b, the locations with small number of entries are now protected by absorbing the entries of their neighbours, representing now a wider area. The final Voronoi diagram generated with these points and their number of aggregated entries, will then represent the protected dataset.

Choosing a less precise approximation digit on the Grid-Reduction stage will create less number of sources, larger number of entries per aggregated location and a coarse representation of the data, as in Figure 4.2. On the other hand, selecting a more precise approximation digit, will create a large number of sources, and some of them will have very little number of entries or zero, especially on the low density and isolated living areas. Event though this aggregation could be used with even lower precision to include the isolated living areas, the amount of information that will be lost from popular areas would be devastating to the value of the information, which is the main issue this technique aims to avoid.

For the sake of evaluating the behaviour of this mechanism, a dataset of 18,966 tweets collected through the Twitter Streaming API [2] from Milano, Italy in 2016 is used as shown in Figure

[2] The code for data collection is available at: `https://github.com/mazucci/geocollect`. However, due

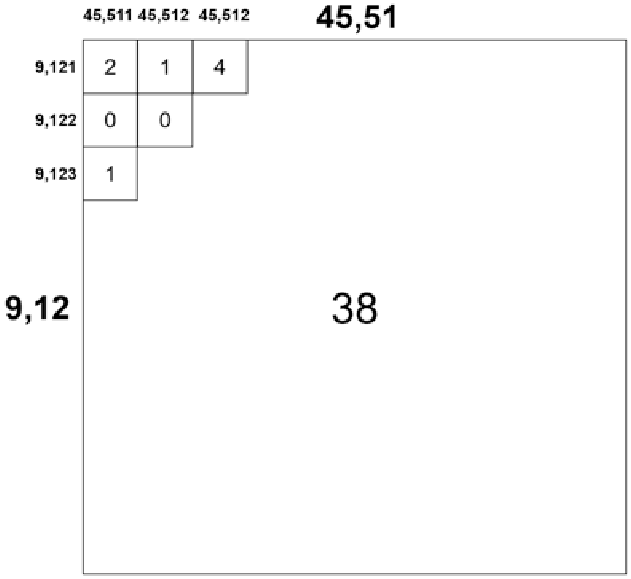


Figure 4.2: Example of grid reduction of data sources based on number of entries and neighbours

4.4. In order to visually gather information on the overall distribution of the tweet sample, a density Voronoi diagram is generated with the original data and shown in Figure 4.5; where unique coordinates were used to created weighted points with the count of repetitions of the same coordinates found in the sample.

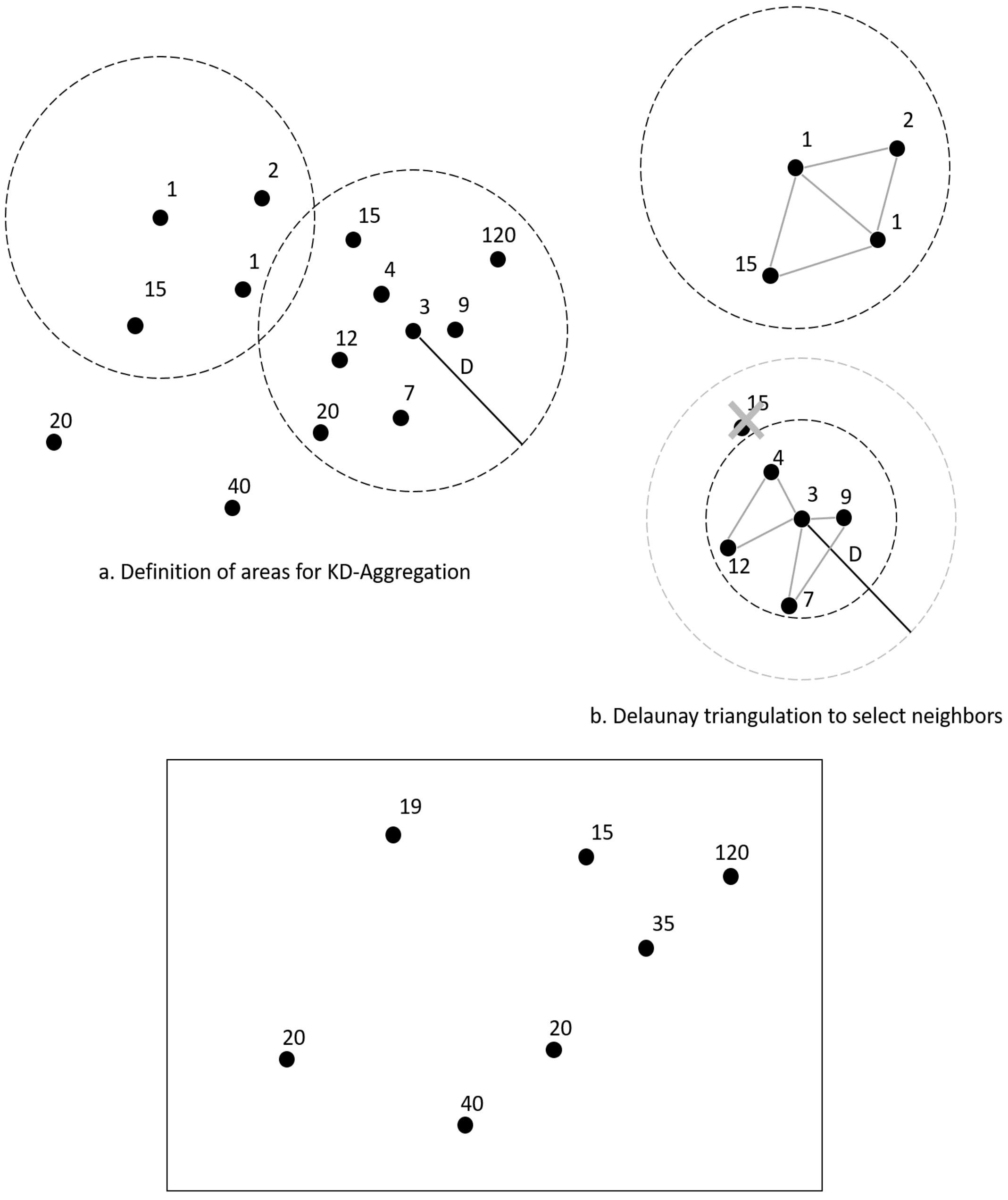
When the Grid-Reduction stage of the algorithm is applied, approximating from 5 to 2 decimal digits is explored. Figures 4.6 and 4.7 are the outputs generated from a density Voronoi diagram for these reduction settings.

Figure 4.6 a. and b., respectively, show the output when 5 and 4 digits are used. This, in comparison to Figure 4.7 a. and b., preserves more similarity to the Voronoi diagram from the original tweets in Figure 4.5. When more decimals digits are approximated, the distribution becomes more grid-like, as it is clearly observed when 2 digits are used in Figure 4.7 b.; this, off-course, is too much of a loss of data representation; therefore a value of 4 is recommended according to these trials.

For the second stage of Voka; KD-Aggregation, a 4 digit Grid-Reduction is selected and tested for a distance D of 500 m shown in Figure 4.8, 1 km and 2 km in Figures 4.9 a. and b. respectively.

From the KD-Aggregated density Voronoi diagrams, it can be seen how the distance D affects the output; when 500 m is used, more polygons are generated than with distances of 1 or 2 km.

to changes in APIs, this code is no longer maintained.



a. Definition of areas for KD-Aggregation

b. Delaunay triangulation to select neighbors

c. Final aggregation of points with aggregated entries

Figure 4.3: Example of the application of KD-Aggregation over data, assuming $K = 20$ and distance D.

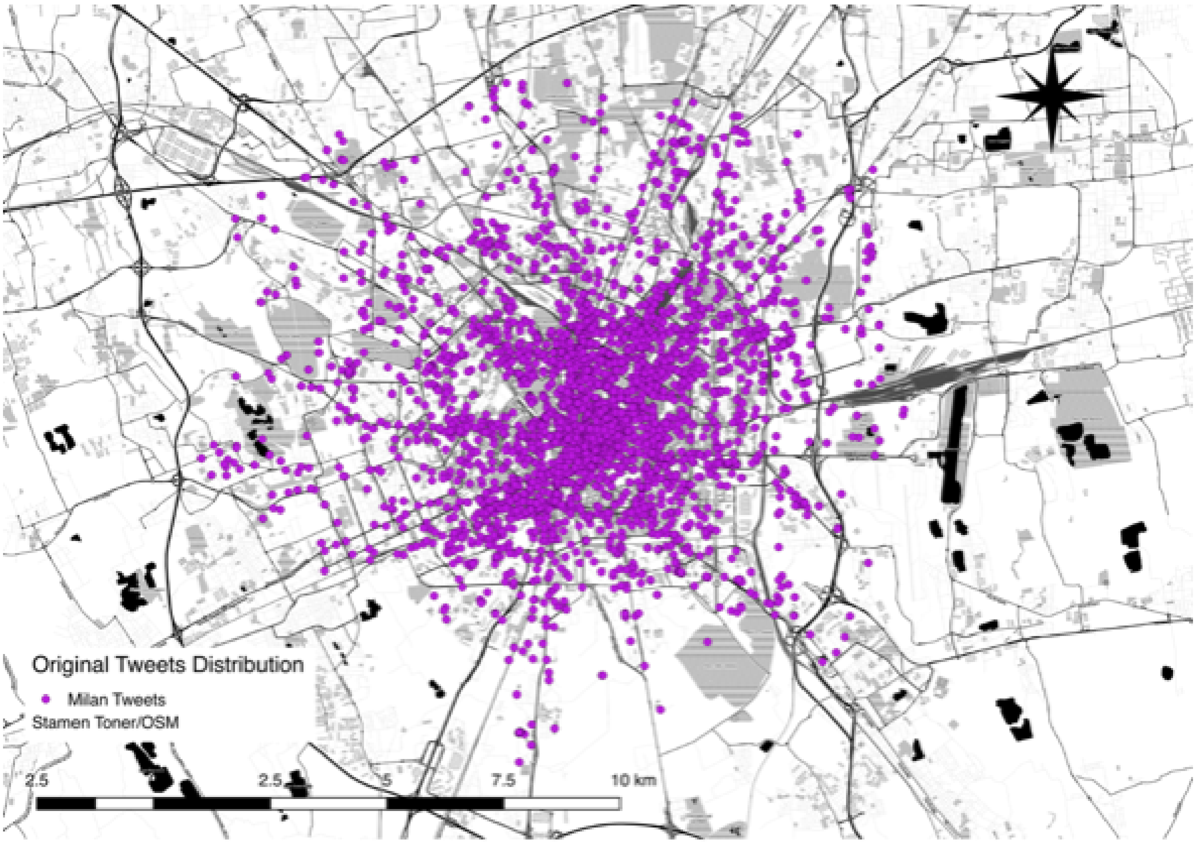


Figure 4.4: Original twitter sample data from Milano, Italy over a Stamen Toner/OSM basemap

Therefore, this distance parameter should be set according to the privacy concerns of the data owner. For this dataset, a distance of 500 m still provides a generalised representation of the original data where high density locations are found in the centre and only few in the outskirts of Milano.

Given that the VoKA algorithm obfuscates by aggregating, it produces data that appear more evenly distributed but with overlapping locations. For a visual understanding of the obfuscated points resulting from this, see Figure 4.10. When this map is compared to the one in Figure 4.4, it seems like the one produced with VoKA have fewer locations, but this is not the case, the locations are simply overlapped due to aggregation.

The Voronoi-based K-Anonymity algorithm, Voka, presents a viable alternative for data aggregation that takes into account locality and protects isolated sources of information, potentially vulnerable for individualisation attacks, while preserving the quality of information in locations that do not require privacy treatment. The main drawback of VoKA is that it can still generate locations that do not comply with the K-anonymity requirement due to locality; i.e. if an isolated point does not find enough nearby points to aggregate enough entries, it may still be vulnerable.

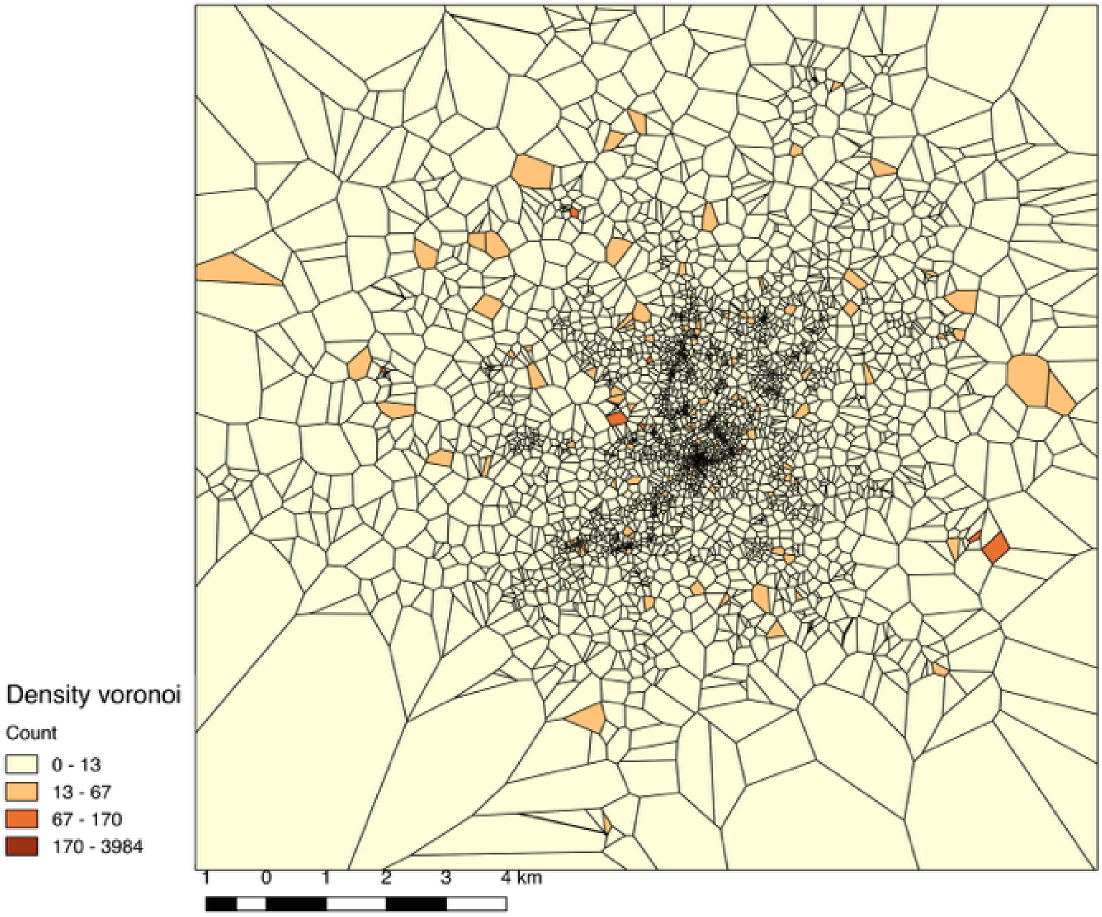


Figure 4.5: Density Voronoi diagram of original twitter sample data from Milano, Italy

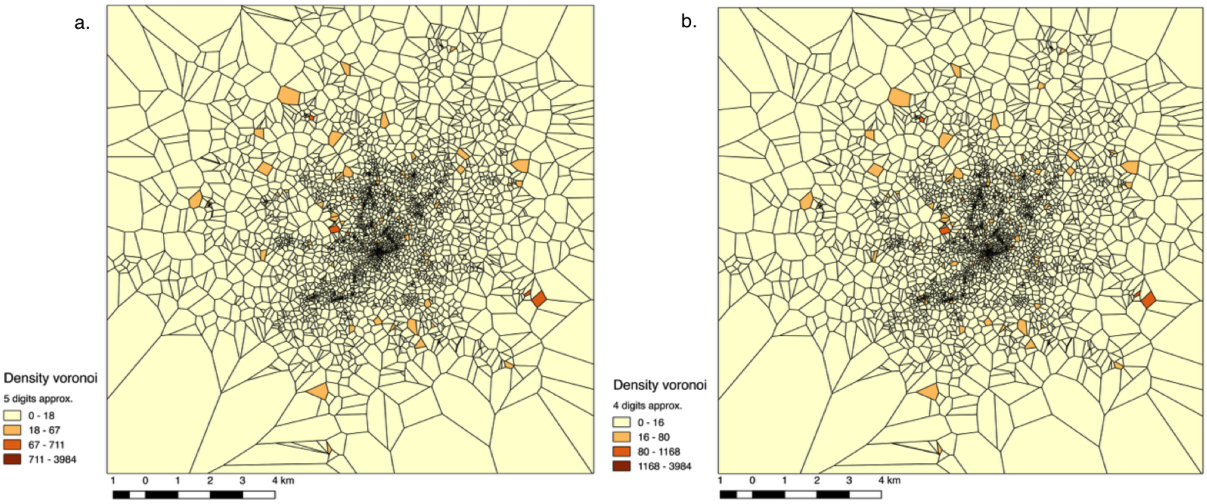


Figure 4.6: Density Voronoi diagram of the first stage of VoKA mechanism; Grid-Reduction, with 5 a. and 4 digits b.

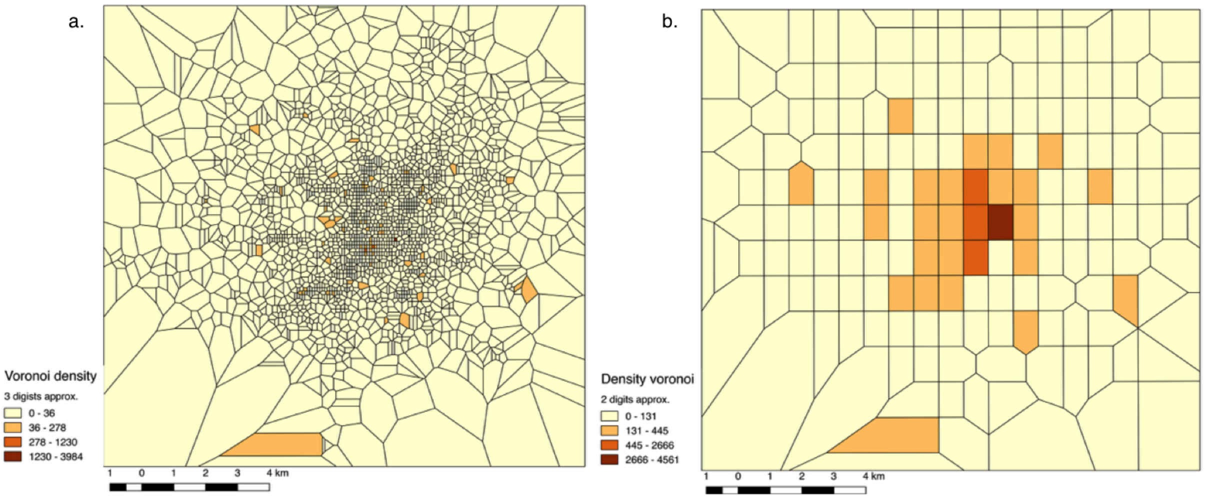


Figure 4.7: Density Voronoi diagram of the first stage of VoKA mechanism; Grid-Reduction, with 3 a. and 2 digits b.

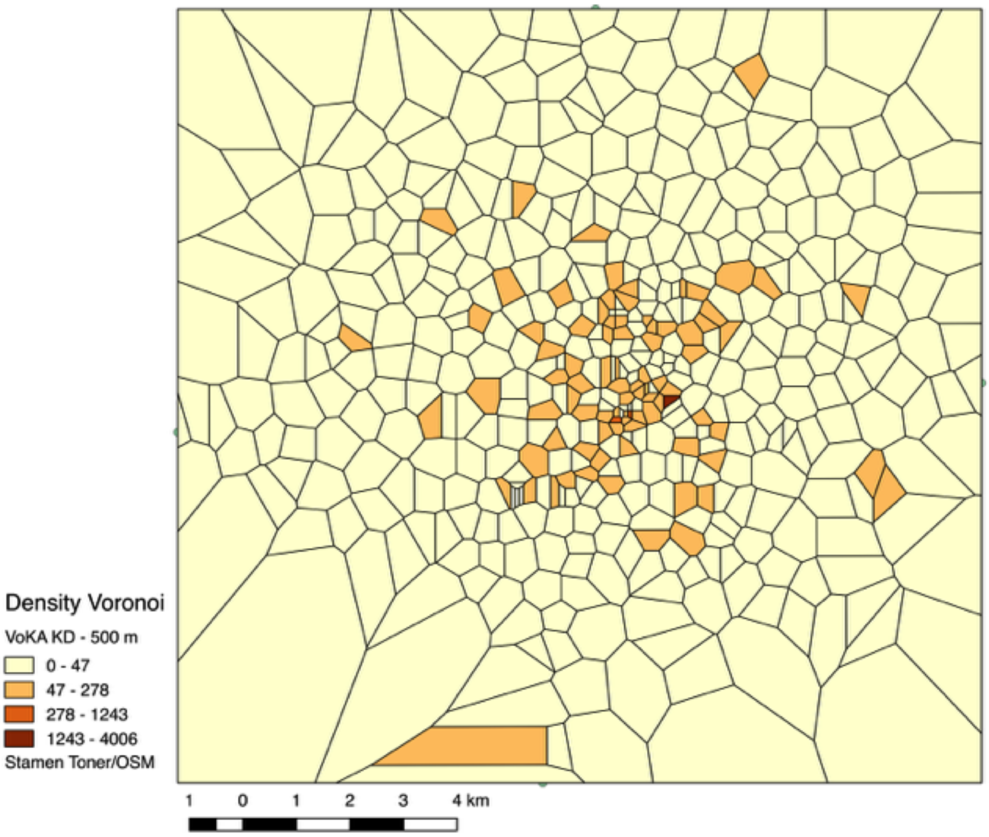


Figure 4.8: Density Voronoi diagram of KD-Aggregation using a 4 digits Grid-Reduction and the distance is set to 500 m

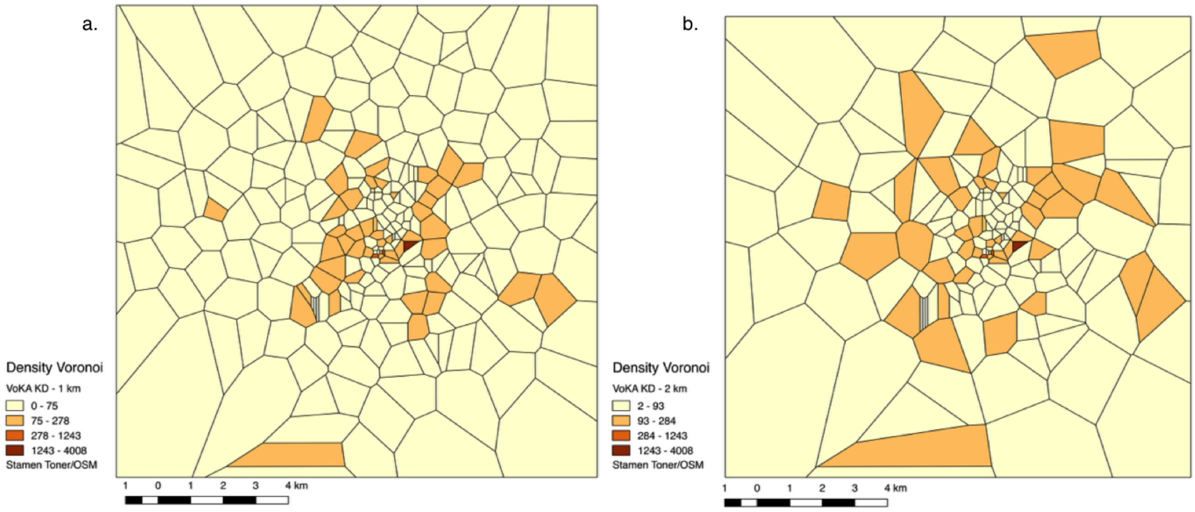


Figure 4.9: Density Voronoi diagram of KD-Aggregation using a 4 digits Grid-Reduction and the distance is set to 1 km a. and 2 km b.

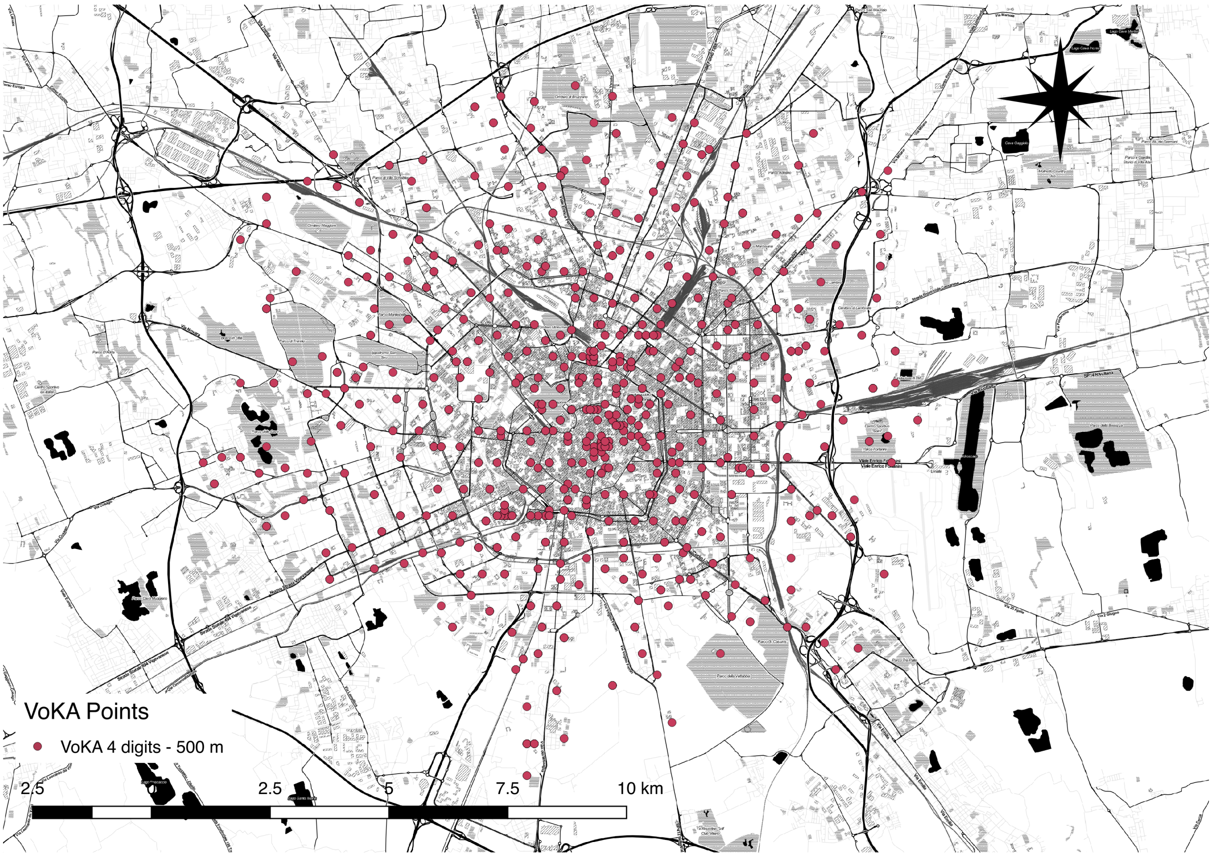


Figure 4.10: Output for hotspot analysis generated with obfuscated tweets in Milano, Italy with 4 digit Grid Reduction and 500 m used for KD-Aggregation

## 4.3 NRand-K

The N-Rand algorithm presented by Wightman et al. (2011) and reviewed in Chapter 2 Section Location Obfuscation is an alternative for obfuscating geographic information while using the LBS (real-time), with a low computational footprint, that does not require previous knowledge from the context of the user; like extension of the area of interest, how many other users are on the area, etc. However, deciding an optimal maximum noise radius becomes the main issue, as it happens with obfuscation techniques. Both before the implementation (which is a good noise level?) and after the data collection (how much did the data changed? Is the data still useful?). For a visual representation of the N-Rand algorithm see Figure 3.4a.

Koufogiannis and Pappas (2016) show how the exploitation of location privacy should depend on the population density, where sparsely populated areas need tighter privacy protection. This concept was introduced to the NRand-K algorithm, where a greater obfuscation takes place, only if there are less than K users in a delimited area; otherwise, the added noise is a minimum preset in order to still provide protection, but with limited distortion of the data. This is analogous to the concept of k-anonymity: where the location of a user is considered anonymous if k-1 other users are in the same area. Similarly, in NRand-K areas with higher density of reported locations are considered safer against inferences in an anonymous dataset than areas with fewer locations.

Since NRand-K is intended to be used by a trusted service provider when the collected location information is going to be released, this mechanisms does not require a third party nor a peer to peer approach, as is usually the case with techniques than enforce K-anonymity.

NRand-K uses aggregation to determine local density, this aggregation can be done using a grid of squared cells, a fishnet, administrative boundaries or by other means at the discretion of the service provider. The K parameter is set having into consideration the overall concentration of reported locations in the spatial extension of the study. This value could be set arbitrarily, but it is recommended to use the 75th percentile ($P_{75}$) of the distribution of the resulting point count within each polygon used for aggregating. Then, if this count is equal or greater than K, a minimum noise ($r_{min}$) is added using the NRand algorithm to each coordinate in the polygon; otherwise, if the number of points is less thank K, the full maximum noise ($r_{max}$) is applied. The algorithm is described as follows:

1. Set values for $r_{max}$ and $r_{min}$.
2. Aggregate geographical coordinates into polygons and calculate each polygon point density.
3. Let $K$ be $P_{75}$ of the calculated polygon densities.
4. For each polygon evaluate if its point count is less than $K$, if so, apply the N-Rand algorithm for all locations reported within the polygon with a radius of $max_r$; otherwise do so

with an obfuscation radius of $min_r$.

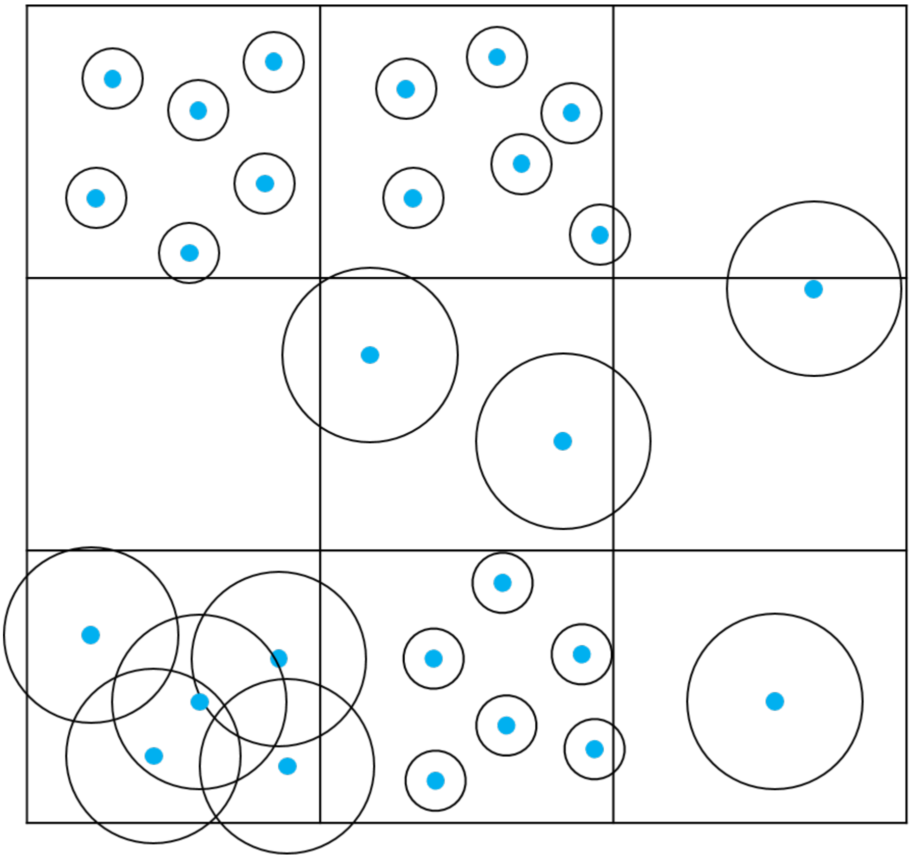


Figure 4.11: Sample of NRand-K where $K = 5$

Figure 4.11 shows a sample of the noise dominium for NRand-K, depicting smaller areas around the blue locations for the NRand algorithm where there is higher concentration of points, this is where there are at least 6 locations for this example (K=6). On the other hand, cells where there are isolated points or cells with low density, NRand-K uses a bigger dominium for inducing noise as shown in cells with less than 6 points, where the circles around the blue points are larger. With the addition of noise, some points may fall in a different polygon than the one they were initially aggregated; this consequence of altering the data is one that can cause an impact in spatial analysis. NRand-K aims to minimise this impact by reducing the dominium where this may happen. A variation to overcome this issue, is to generate random points until the obfuscated location falls within the polygon containing the original location.

Given that the main goal of the proposed NRand-K algorithm is to provide a decision factor for inducing noise based on the overall spatial density of the data, it can be used with other noise-based obfuscation mechanisms than NRand, such as $\theta-$Rand or Pinwheel.

## 4.4 Discussion

Mechanisms for protecting location information for release to third parties rely on the service provider; it is the LBS that ultimately decides wether or not to distort location information, and defines the parameters that these mechanisms require, in order to generate the allowed distortion. This privacy protection schema for publishing information is also explored by Wieland et al. (2008) while releasing health geolocated information.

In this chapter it was assessed the use of two LPPMs and according to performed experiments, some default settings are recommended for the parameters used by the introduced mechanisms. VoKA and NRand-K are both based on spatial aggregation, which is a common practice for geospatial analysis; therefore, considering this as a mean to obfuscate locations in a way that still represent the original data is not only reasonable, but consistent and only possible if the owner of the geo-referenced data performs these proposed mechanisms.

(Monreale et al., 2013, p. 11) stresses that *"arbitrary territory divisions, such as administrative districts or regular grids, do not reflect the spatial distribution of the data"* and suggests a Voronoi tessellation for aggregating. This concept is embedded in the VoKA mechanism, offering the possibility to aggregate based on the distribution of the reported locations and not depending on administrative boundaries or static grids.

The NRand-K algorithm presents a way to integrate the concept of K-anonymity and location obfuscation in real scenarios, offering the possibility to include the aggregation of the data in the decision of how much noise is required. Other metrics, such as the historical density of reported locations are proposed for estimating the K parameter. On the next chapter, experiments will be performed for NRand-K and VoKA exploring how these algorithms minimise the impact on geospatial analysis.

## 4.5 Chapter Summary

This section introduced LPPMs developed throughout this work that are applicable after the data collection stage; when the LBS has provided its services and wishes to reuse the collected data for sharing with third parties. These kinds of LPPMs benefit from the availability of the dataset to protect; having initial knowledge of the characteristics of the data enable a better selection of the input parameters for the algorithms. Both mechanisms presented in this chapter use K-anonymity as a criteria to provide geoprivacy.

# Chapter 5

# Assessing Impact of Location Privacy on Geospatial Analysis

## 5.1 Introduction

The need for addressing privacy issues while being able to obtain feedback in LBS has been a concern for some years, in the work of (Beresford and Stajano, 2004, p. 5), they state:

> *"Enabling location privacy is going to become increasingly important in a world where location-aware services are available over larger and larger geographical areas. Deploying systems that support location privacy for users and provide feedback about the level of anonymity users have may prove critical to the widespread adoption of location-aware services."*

Measuring the impact of LPPMs in geospatial analysis is *now* a necessary step, in order to understand how these analyses will be affected if there is a massive adoption of privacy protection techniques, either due to public awareness or due to legislation and enforced geoprivacy standards; probably both.

This chapter is dedicated to understand the impact that location obfuscation techniques have in geospatial analyses. For this, the performance of ESDA tools such as hotspot analysis and heatmaps is evaluated; the calculation of global spatial indices is also explored to quantify alterations of the data after applying privacy protection mechanisms. The select LPPMs to test in this chapter are Pinwheel, NRand-K and VoKA: Voronoi K-Aggregation Mechanism; the outputs of the geospatial analyses obtained with the original and obfuscated datasets will serve for comparison.

The chapter is organised as follows: a section dedicated to describe the methodology used for evaluating the LPPMs independent of the specificities of each mechanism, followed by a section dedicated to the performed experiments with each of the selected LPPM. At last, a comparison section is dedicated to evaluate the performance of each mechanism in terms of alterations on geospatial analysis. This chapter contributes to the Research Objective 3: *Asses the impact of LPPMs based on obfuscation and aggregation on geospatial analysis.* Which addresses the last component of the thesis question: *How to provide location privacy with mechanisms that are scalable, efficient, non-intrusive and with low impact on geospatial analysis?*

## 5.2 Methodology

The experiments were performed using the QGIS software in the version 2.18 *'Las Palmas'*[1] and other calculations were done using libreOffice[2]. The following subsections will explain in detail the data sources and the specificities on how the experiments were conducted; in general all software used is Free and Open Source (FOSS).

### 5.2.1 Data

The data used in this chapter is geo-referenced Twitter data collected through their Streaming API [3]. The collection was ongoing for the most part of 2016 and limited to tweets generated in Italy. This allowed to have a large dataset for different cities throughout a continuous period. For these experiments, the original dataset includes 18966 geo-referenced tweets from the city of Milano, Italy collected from February to May of 2017. From this thesis view, the collected data comes from a platform that offered an *'opt-in'* agreement at the time of collection; therefore, from what was stated in the literature review Section Crowdsourced Spatial Data Collection, this data is considered as VGI.

For performing hotspot analysis, data must be aggregated; for this aggregation, the administrative boundaries are used. In the case of the city of Milano, the neighbourhoods or *'Nuclei d'Identità Locale'* in Italian, are used [4].

[1] qgis.org/en/site/forusers/alldownloads.html

[2] libreoffice.org

[3] The code for data collection is available at: github.com/mazucci/geocollect. However, due to changes in APIs, this code is no longer maintained.

[4] Shapefiles available from the Milano Geoportale

### 5.2.2 Heatmaps

The computation of raster heatmaps on spatial data estimates the density based on pixel values of the data distribution. Each pixel is evaluated within a radius or bandwidth predefined by the analyser.

> The Heatmap plugin uses Kernel Density Estimation to create a density (heatmap) raster of an input point vector layer. The density is calculated based on the number of points in a location, with larger numbers of clustered points resulting in larger values[4].

For the creation of heatmaps, it was used the Heatmap plugin for QGIS[5]. The bandwidth for this analysis was set to 1km, using linear interpolation and a continuous mode to classify the heatmap rasters in all cases. The rasters are then normalised by subtracting the mean and dividing this by the standard deviation.

These classification maps, aid in exploring datasets in order to identify atypical concentrations relative to the events present in the analysis. For means of visualisation, the kernel density values are classified within ranges, these ranges are visually presented through a colour palette, where each colour represents a magnitude or temperature; the colour indicated with a higher value in the legend implies a higher concentration of features. This classification is good as an initial approach for understanding the distribution and key areas of the overall domain, but density maps alone cannot determine statistical significance.

### 5.2.3 Hotspot Analysis

Hotspot analysis calculates the Getis-Ord local statistic Gi$*$ (Ord and Getis, 1995), in order to define areas of atypical high point density occurrence, named hot spots, versus atypical areas of low occurrence, named cold spots. The z-scores and p-values resulting from the Gi$*$ are tested with the null hypothesis (i.e. complete spatial randomness) for each feature.

In order to perform this analysis on spatial coordinates with no predefined weigh variable associated, the aggregation of points in a certain area is the viable option. For this work, a vector layer of 85 neighbourhoods of Milan was used for aggregating; the resulting features for the analysis are the polygons associated with the corresponding count of tweets within each neighbourhood i.e., polygon density. These are the input for the Hotspot Analysis Plugin which was performed using a queen's case lookup to calculate hotspots statistical significance value.

The hotspot analysis in contrast with heatmaps, enables to determine statistical significance. For

[5] QGIS Heatmap Plugin

the performance of these experiments, the QGIS Hotspot Analysis plugin presented in (Oxoli et al., 2017) was used; refer to this work for a detailed explanation of the algorithm[6].

Reckoning the variations of hotspots of obfuscated and original data as a measure to compare LPPMs has been explored by Kounadi and Leitner (2015), the authors emphasise the importance of doing this given the relevance of clustering in geospatial analysis.

> Another effect of masks is the variations between hotspots derived from masked and original incident locations. This was an additional information dimension by Armstrong et al. (1999), where they examined the existence, the actual locations, and the relative locations of masked compared to original clusters. Similarly, Kwan et al. (2004) calculated the clustering distance using the cross K function analysis... Leitner and Curtis (2004) compared the similarity of observed hotspots, drawn by participants, which were derived from masked and original points...

The impact assessment in this work is based on quantifying the change between the outputs of the hotspot analysis for the different LPPMs compared to the output with the original data. For this, the number of new hotspots that appeared in the results after applying obfuscation (false positives), the hotspots that disappeared (false negatives) and statistical significance variations are considered.

### 5.2.4 Spatial Indices

To analyse other properties of spatial data, indices presented by (Kounadi and Leitner, 2015, p. 745) are evaluated for obfuscated and original datasets, specifically, the mean's divergence index (Mdi) and the orientation's divergence index (Odi) . The Mdi is calculated with Equation 5.1, where the mean values are calculated with the QGIS software and Odi is calculated as denoted in Equation 5.2, the values for the angle orientation of the standard deviational ellipse (SDE) are calculated using the QGIS plugin designed for this analysis [7].

$$Mdi = \frac{d(\text{original mean, obfuscated mean})}{d(\text{original mean, farthest point in study area})} * 100 \tag{5.1}$$

$$Odi = \frac{|\text{orientation angle of original SDE} - \text{orientation angle of obfuscated SDE}|}{180} * 100 \tag{5.2}$$

In order to verify how the original aggregation of the data is affected after obfuscation, the difference between the point count per each polygon used to aggregate the original data (original

[6] The code is available at github.com/danioxoli/HotSpotAnalysis_Plugin

[7] plugins.qgis.org/plugins/SDEllipse

point count) and the point count after obfuscation (obfuscated point count) is used to calculate the point count's divergence index (PCDi). This index is calculated using all polygons included in the analysis and is formally described in Equation 5.3; the MPCDi is designed to identify how variations in aggregation occur among different LPPMs.

$$PCDi = \frac{\sum_{i=1}^{n} |\text{original count}_i - \text{obfuscated count}_i|}{n}, \text{ where } n = \text{polygons in the analysis} \quad (5.3)$$

## 5.3 Experiments

The following subsections illustrate the outputs of the experiments performed with the original dataset and the ones with the resulting obfuscated datasets using the selected LPPMs; these output maps aim to provide a visual understanding of the alterations that LPPMs imply on geospatial analysis. Furthermore, the legends for both heatmaps and hotspot analysis, serve for quantifying alterations between datasets.

### 5.3.1 Original Dataset

This section shows the resulting outputs for heatmap and hotspot analysis using the original dataset of $18,966$ geo-referenced tweets in the city of Milano, Italy. In both maps, a concentration in the centre is spotted; in Figure 5.1 there are two high density clusters, but these appear as if they are merged into one big cluster, making it hard to differentiate between them.

As for the hotspot analysis, the five central neighbourhoods highlighted in Figure 5.2, correspond to the most touristic areas and oldest part of the city, some important landmarks in these neighbourhoods are the *Duomo di Milano*, the *Galleria Vittorio Emanuele* and the *Teatro alla Scala*. The concentration of tweets in this area could be explained by people tweeting for touristic purposes; however, the semantic analysis of the text in the tweets falls outside of the scope of this work.

The experimental results for this scenario using the original tweets will serve as a canon when comparing against analyses performed with obfuscated datasets; generated using the different LPPMs as described in the following subsections.

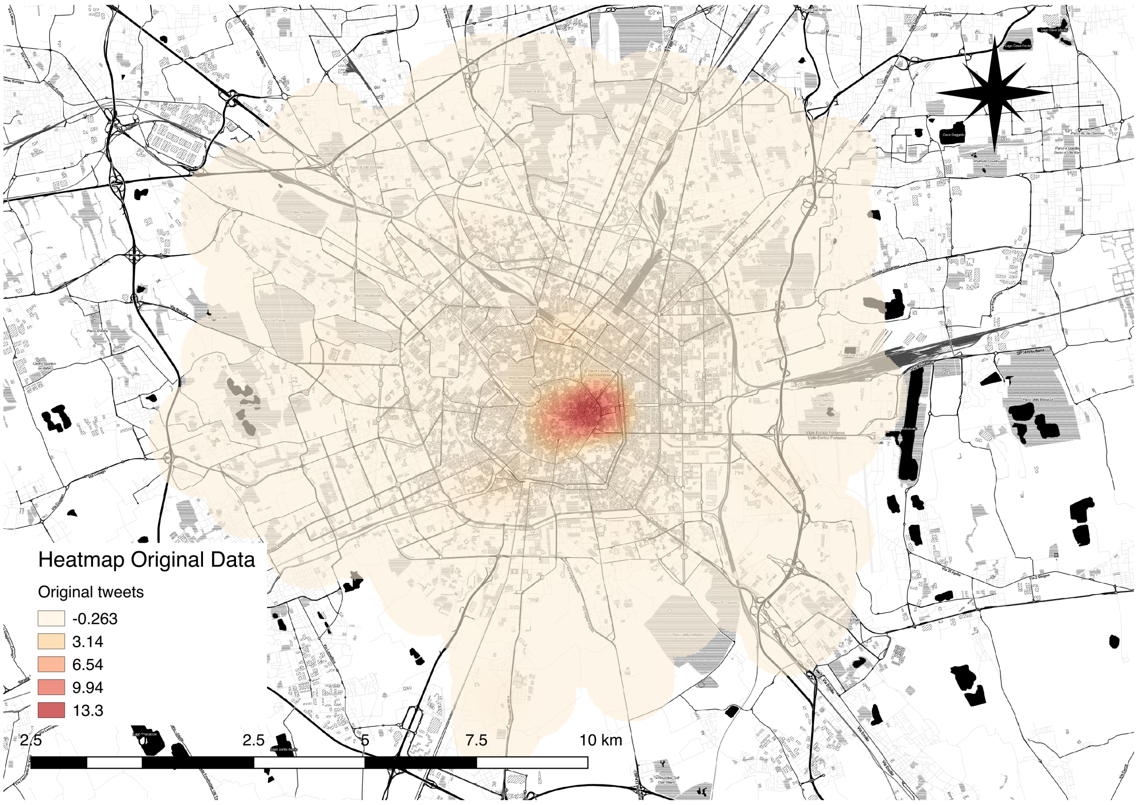


Figure 5.1: Heatmap generated with original tweets in Milano, Italy

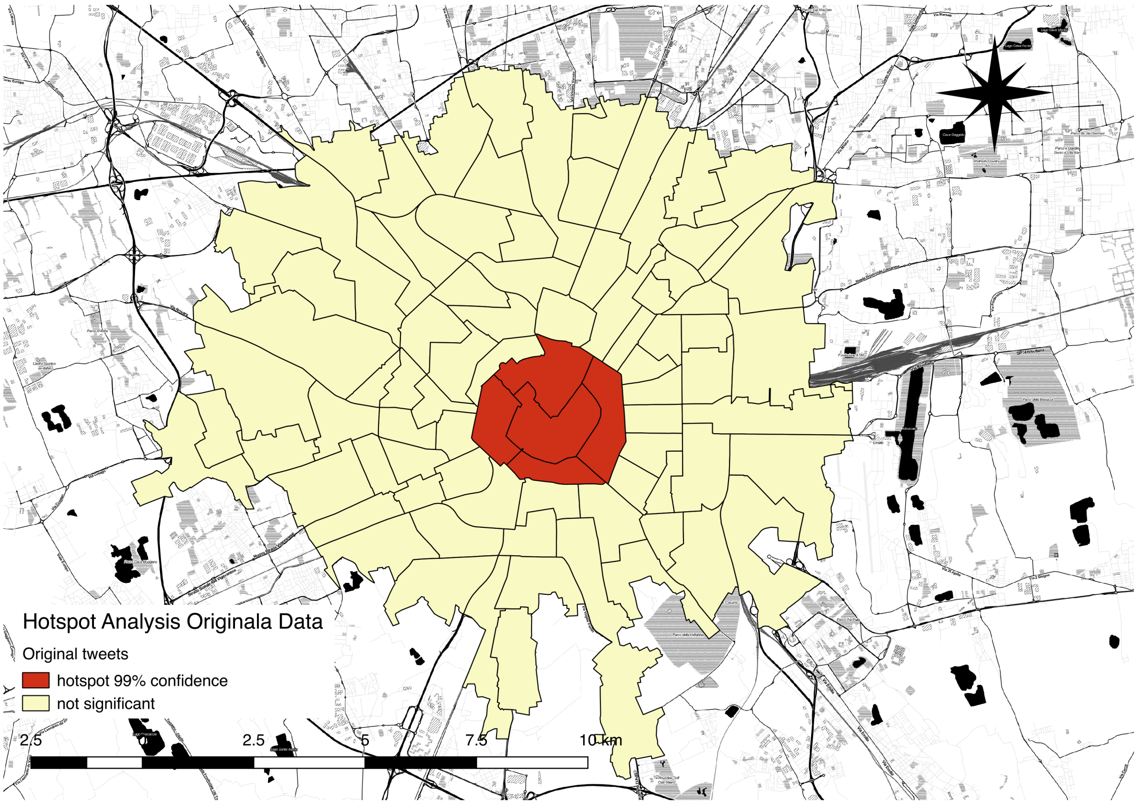


Figure 5.2: Output for hotspot analysis generated with original tweets in Milano, Italy

### 5.3.2 Pinwheel

The Pinwheel algorithm requires two parameters; a $\phi$ angle to determine the periodicity of the pinwheel vanes according to Equation 3.1 and an $r_{max}$, which is the maximum distance an obfuscated location can be from the original one. In these experiments, the Pinwheel algorithm was tested with different $r_{max}$ values: 500m and 1km, and the period parameter $\phi$ was set to 105°in all the experiments in order to have lower periodicity and an asymmetrical random point generation domain; making it more resilient against filtering attacks as mentioned in Subsection 3.2.3 of Chapter 3 where the mechanism was first introduced in this thesis.

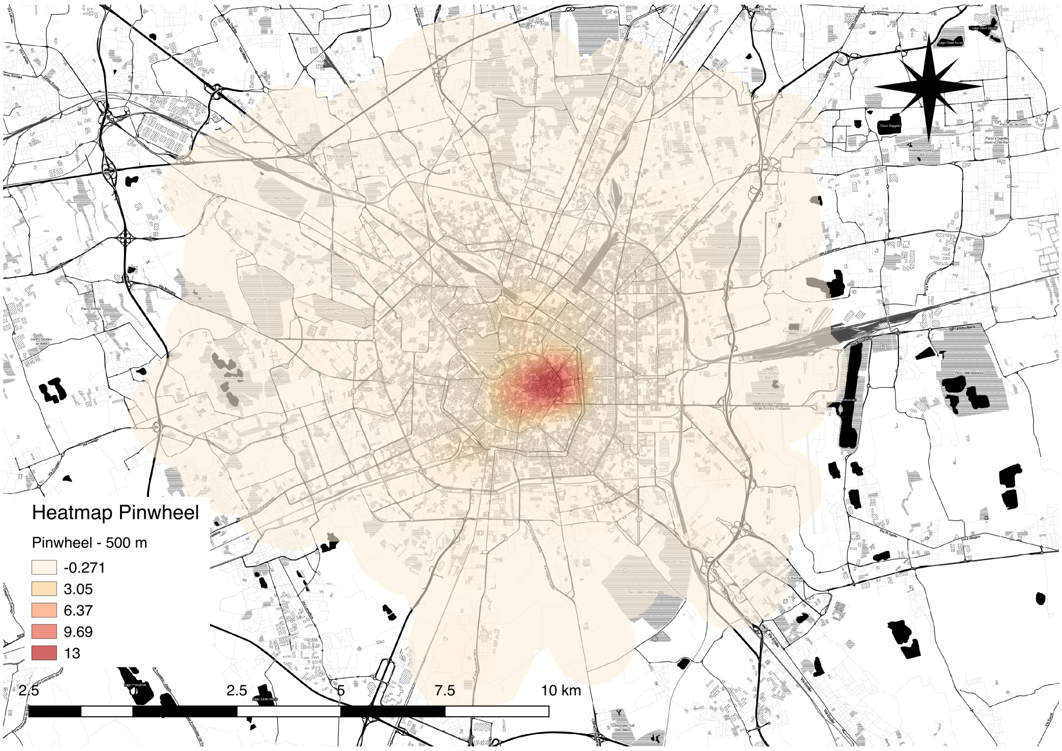


Figure 5.3: Heatmap generated with obfuscated tweets in Milano, Italy with $r_{max} = 500m$

Figure 5.3 shows a heatmap output using tweets obfuscated with Pinwheel with an $r_{max}$ of 500m. At first sight this map is very similar to the heatmap produced with the original tweets; however a diminished concentration can be appreciated when comparing the legends, where a difference of 0.3 can be appreciated in the highest temperature values. This becomes more evident with the output of the heatmap produced with the obfuscated dataset using an $r_{max}$ of 1km in Figure 5.4, in this case, the highest concentration in the legend decreases to 11.6 from that of 13.3 tweets on the heatmap produced with the original dataset in Figure 5.1, differing in a value of 1.7.

Further analising these maps by subtracting each of the RGB channels and generating a divergence heatmap (or heatmap of differences)[8] as the ones shown in Figure 5.5. These heatmaps

[8] Using the script by Mavridis (2012).

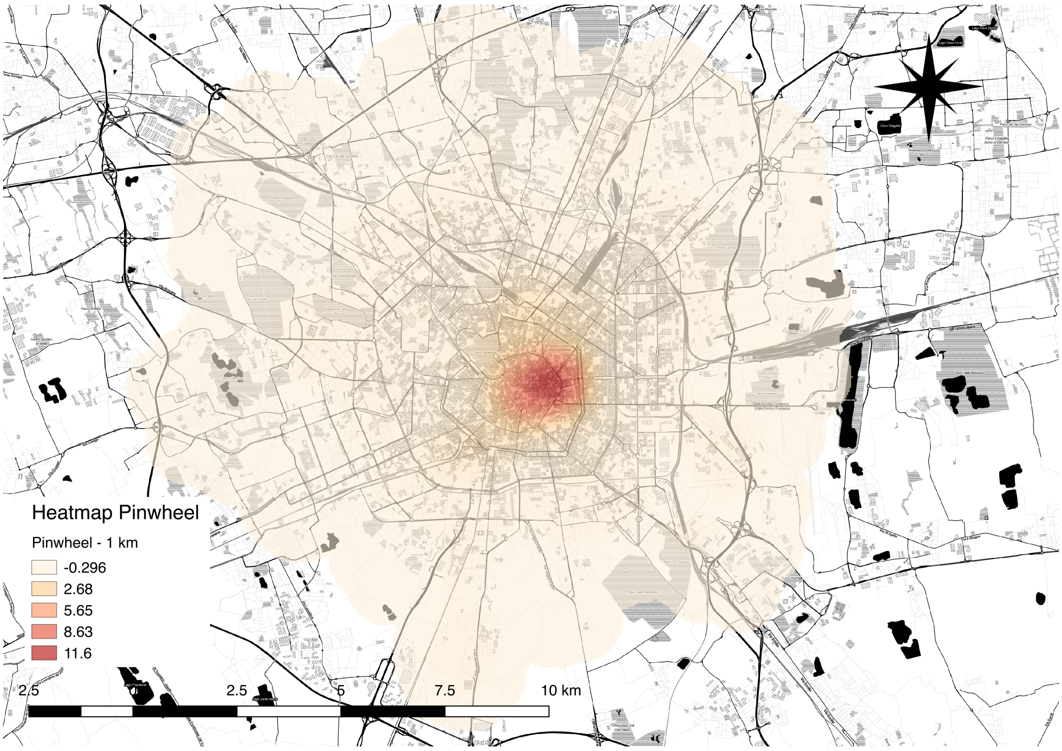


Figure 5.4: Heatmap generated with obfuscated tweets in Milano, Italy with $r_{max} = 1km$

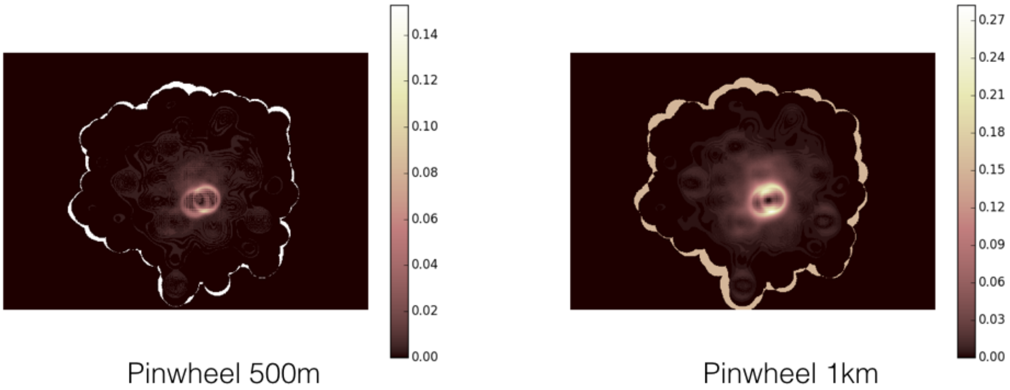


Figure 5.5: Heatmap divergence by pixels between the original data heatmap and the Pinwheel heatmaps generated with an $r_{max} = 500m$ and $r_{max} = 1km$

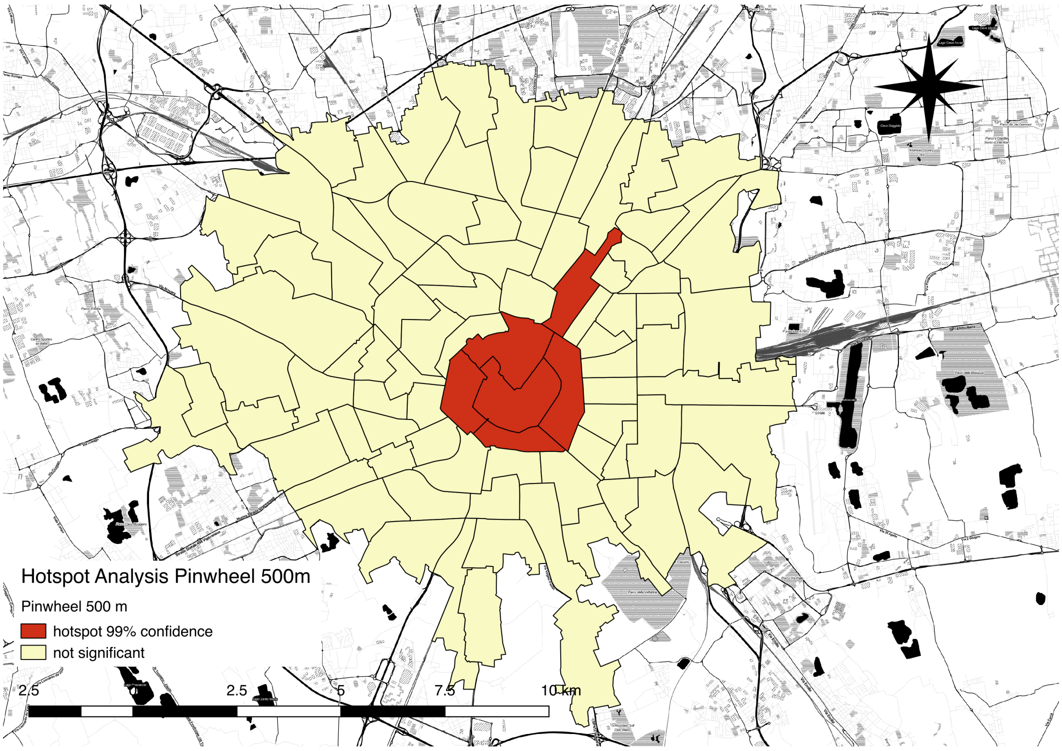


Figure 5.6: Output for hotspot analysis generated with obfuscated tweets in Milano, Italy with $r_{max} = 500m$

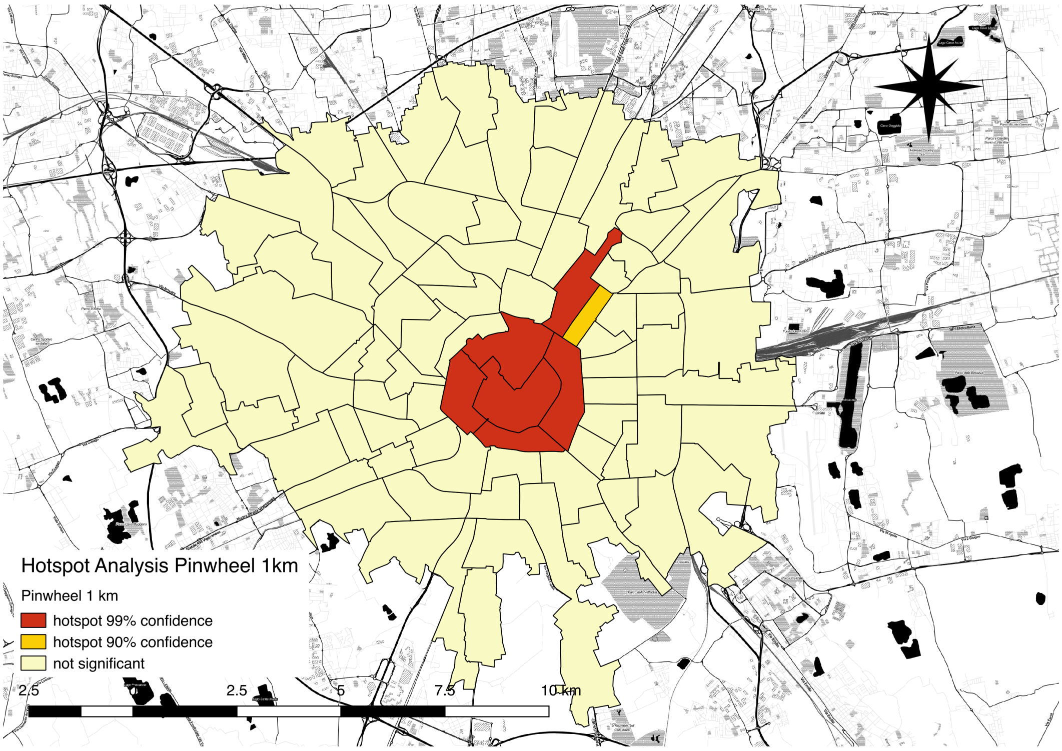


Figure 5.7: Output for hotspot analysis generated with obfuscated tweets in Milano, Italy with $r_{max} = 1km$

shows highlighted pixels where these images differ; the lighter the colour, the higher the error or difference between the original and the obfuscated heatmap. For the case of the obfuscated tweets with Pinwheel, the divergence heatmaps show more difference with an $r_{max} = 1$km than when the $r_{max}$ was 500m.

Even though the process for generating the heatmaps remains the same for all experiments; using a bandwidth of 1km, a linear interpolation and a continuous mode for classifying the heatmap rasters, the variations in the legend values of these maps indicate a more sparse point distribution. These variation in the legends serve to identify how the induced noise decreases the concentration in clusters originally identified.

The hotspot analysis outputs using these obfuscated datasets evidence differences from the one produced with the original dataset (Figure 5.2); in both variations of these experiments using $r_{max}$ of 500m (Figure 5.6) and 1km (Figure 5.7), now distinct hotspots appear on the output maps. There is a clear relation between the induced noise and sparsity of the obfuscated data with this mechanism, both heatmap and hotspot analysis reflect this; Figure 5.6 marks a new hot spot with a 99% statistical significance that was not present in the original analysis, while Figure 5.6 marks two new hot spots; the one mentioned previously when using $r_{max}$ of 500m and a new one with a 90% statistical significance.

### 5.3.3 NRand-K

For the NRand-K algorithm details refer to the Section 4.3 of Chapter 4 where this mechanism was introduced. In the experiments presented below, the K parameter was set to the 75th percentile ($P_{75}$) of the point count frequency of the 85 neighbourhoods in Milano, which resulted in a value of 149. This means that reported tweets within neighbourhoods with 149 or more tweets were altered with a minimum noise ($r_{min}$) that was set to 50m in all cases, while tweets reported within neighbourhoods with a point count of less than 149 were obfuscated using $r_{max}$. The $r_{max}$ parameter was set to 500m and 1km in order to be able to compare with both $r_{max}$ variations evaluated using the Pinwheel mechanism in the previous subsection.

In order to avoid generating obfuscated locations outside the area of the study or locations that fall in different neighbourhoods from the one where the original location was, the algorithm was restricted to provide only obfuscated locations within the neighbourhood polygon of the original location used to generate it. The results with both variations of NRand-K are presented as follows:

When compared to the heatmap from the original dataset in Figure 5.1, the output maps for NRand-K with $r_{max} = 500$m in Figures 5.8 and 5.15 show greater similarity than those generated with the Pinwheel mechanism; particularly if noting the values in the scales of the legends. For

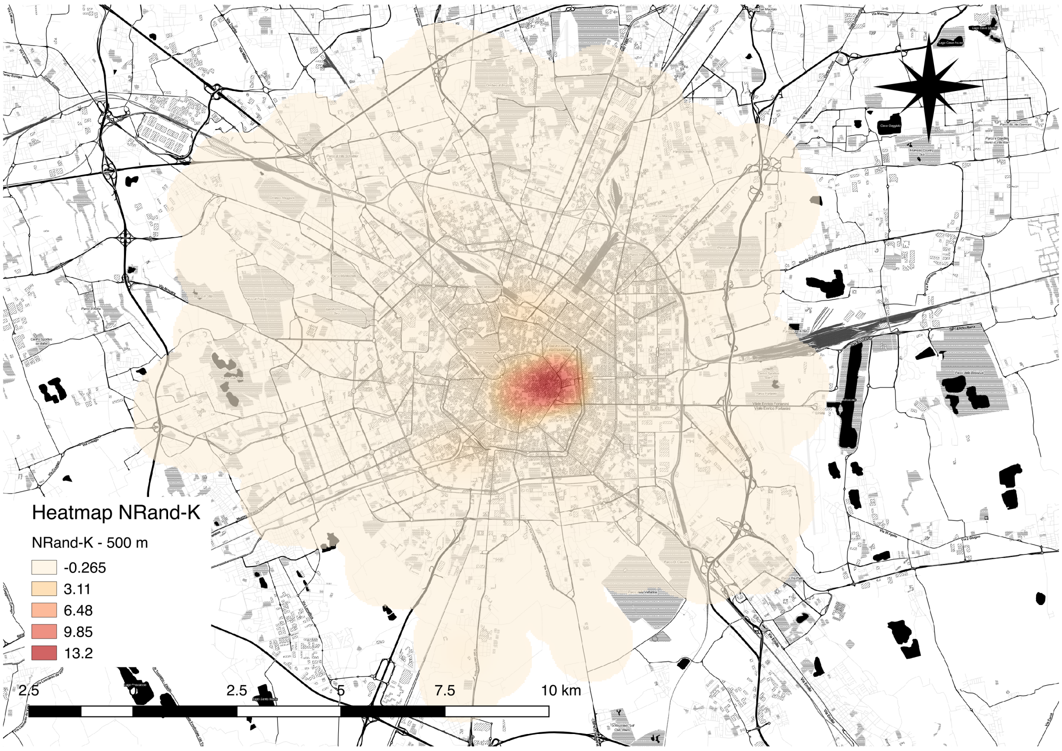


Figure 5.8: Heatmap generated with obfuscated tweets in Milano, Italy with $r_{max} = 500$m

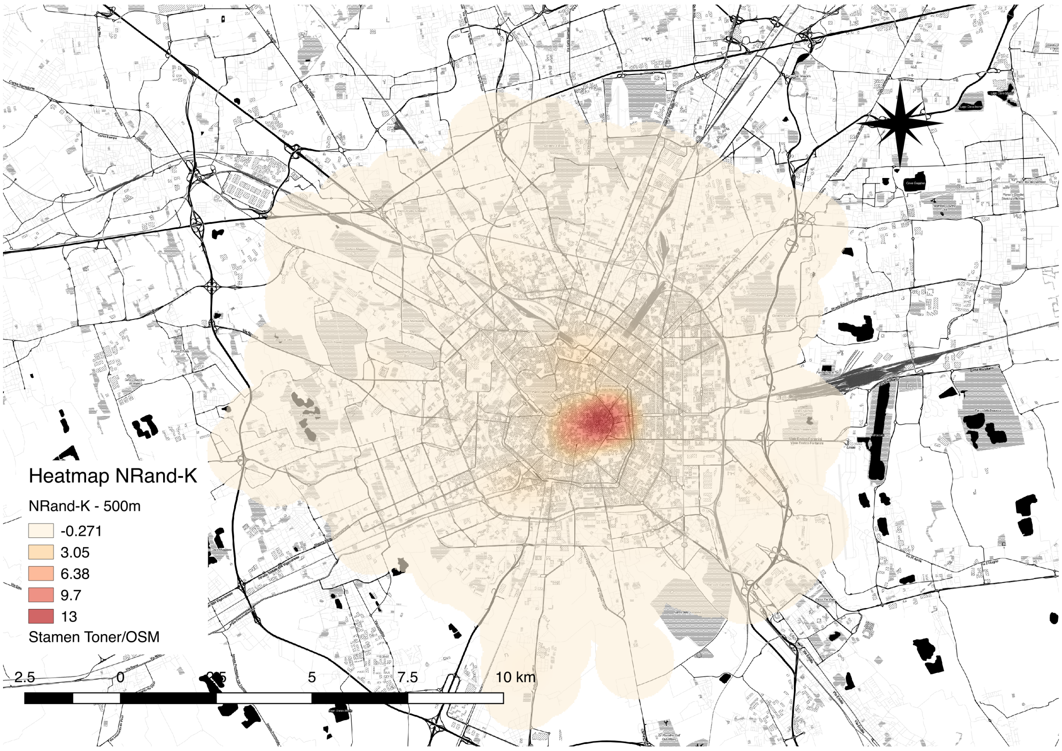


Figure 5.9: Heatmap generated with obfuscated tweets in Milano, Italy with $r_{max} = 500$m and polygon restriction

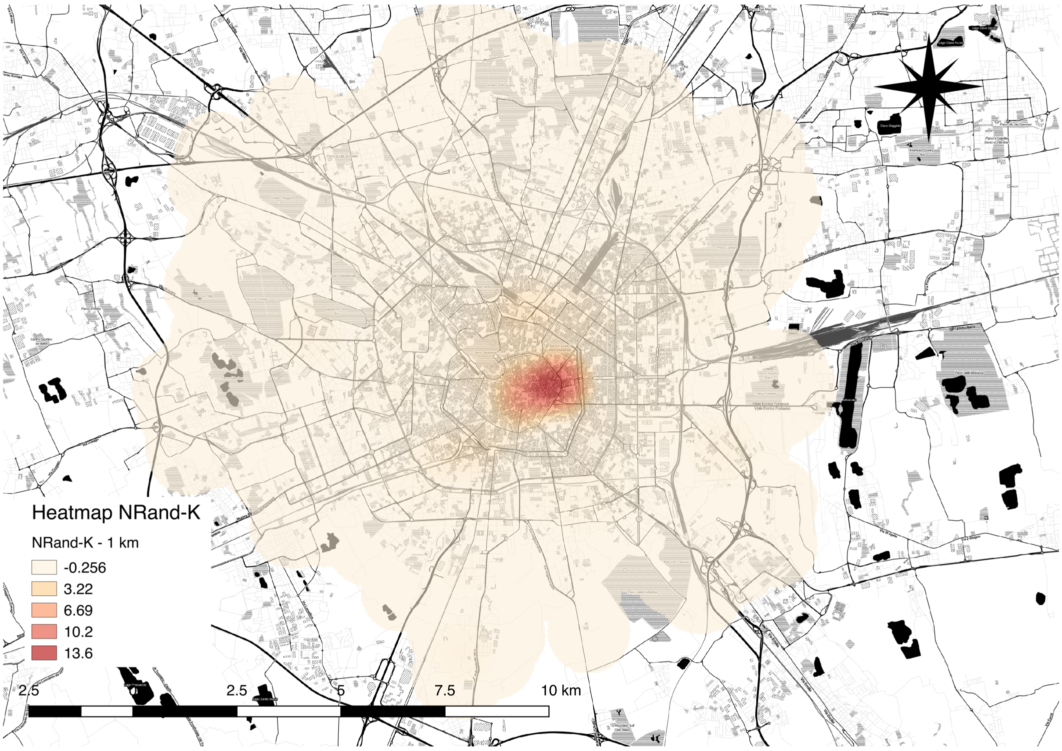


Figure 5.10: Heatmap generated with obfuscated tweets in Milano, Italy with $r_{max} = 1$km

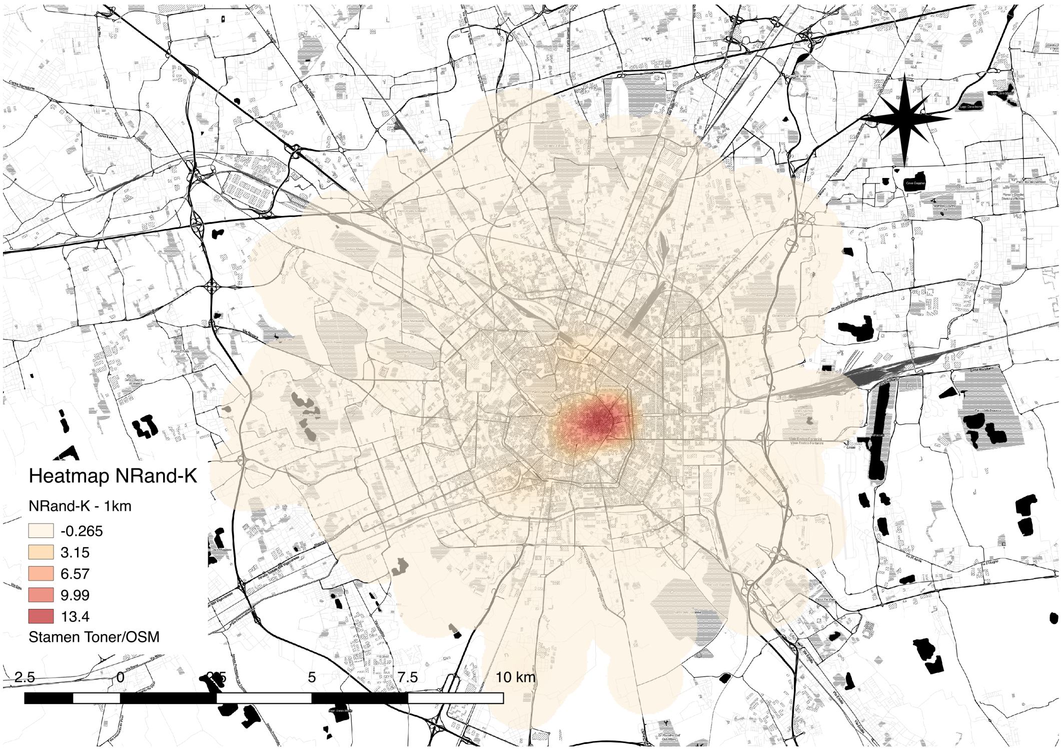


Figure 5.11: Heatmap generated with obfuscated tweets in Milano, Italy with $r_{max} = 1k$m and polygon restriction

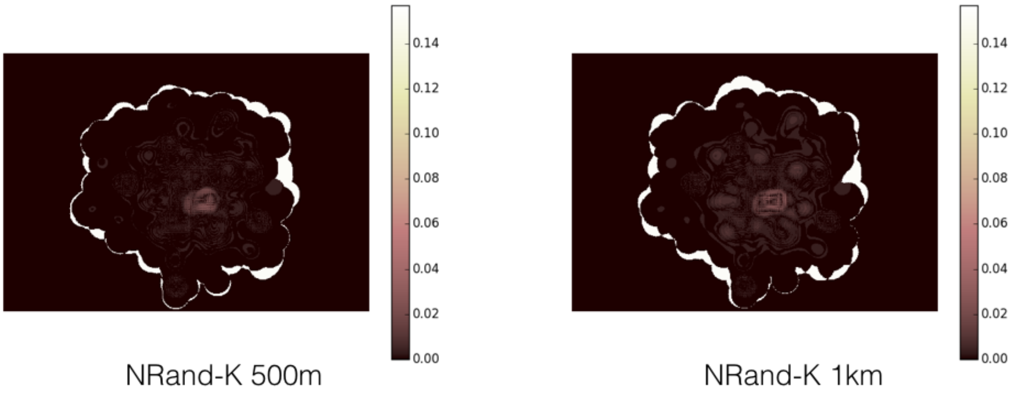


Figure 5.12: Heatmap divergence by pixels between the original data heatmap and the NRand-K heatmaps generated with an $r_{max} = 500m$ and $r_{max} = 1km$

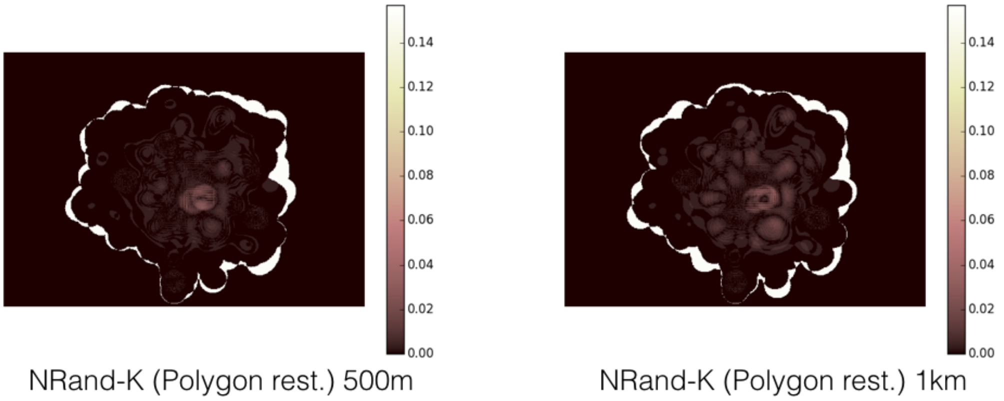


Figure 5.13: Heatmap divergence by pixels between the original data heatmap and the NRand-K (with polygon restriction) heatmaps generated with an $r_{max} = 500m$ and $r_{max} = 1km$

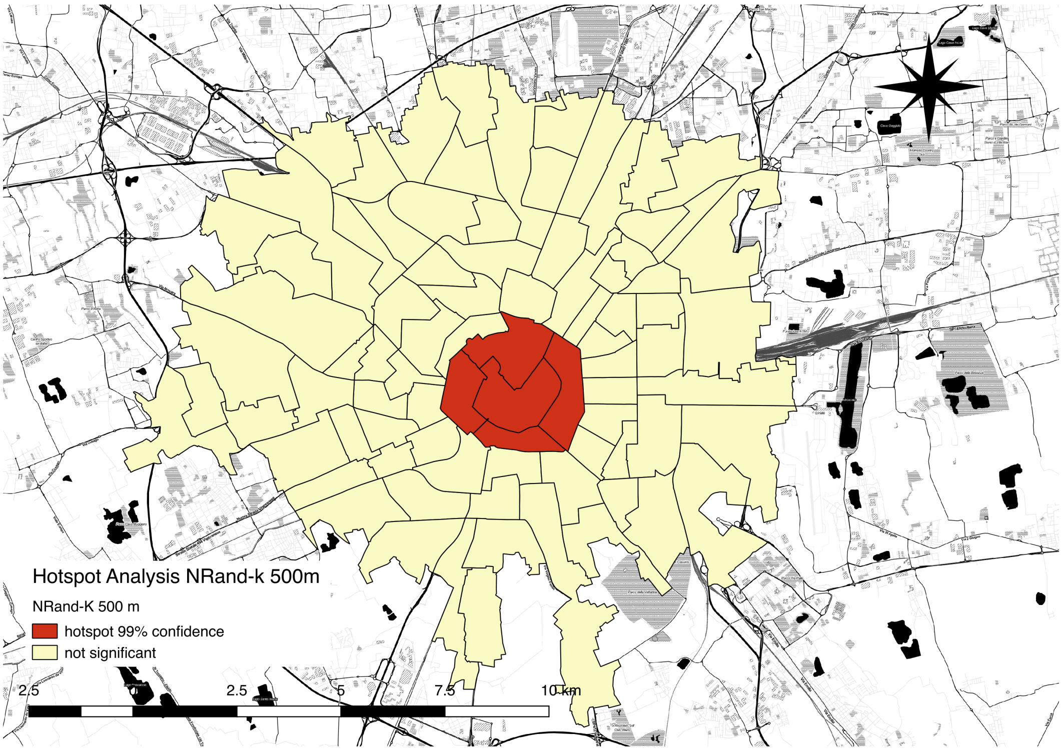


Figure 5.14: Output for hotspot analysis generated with obfuscated tweets in Milano, Italy with $r_{max} = 500$m

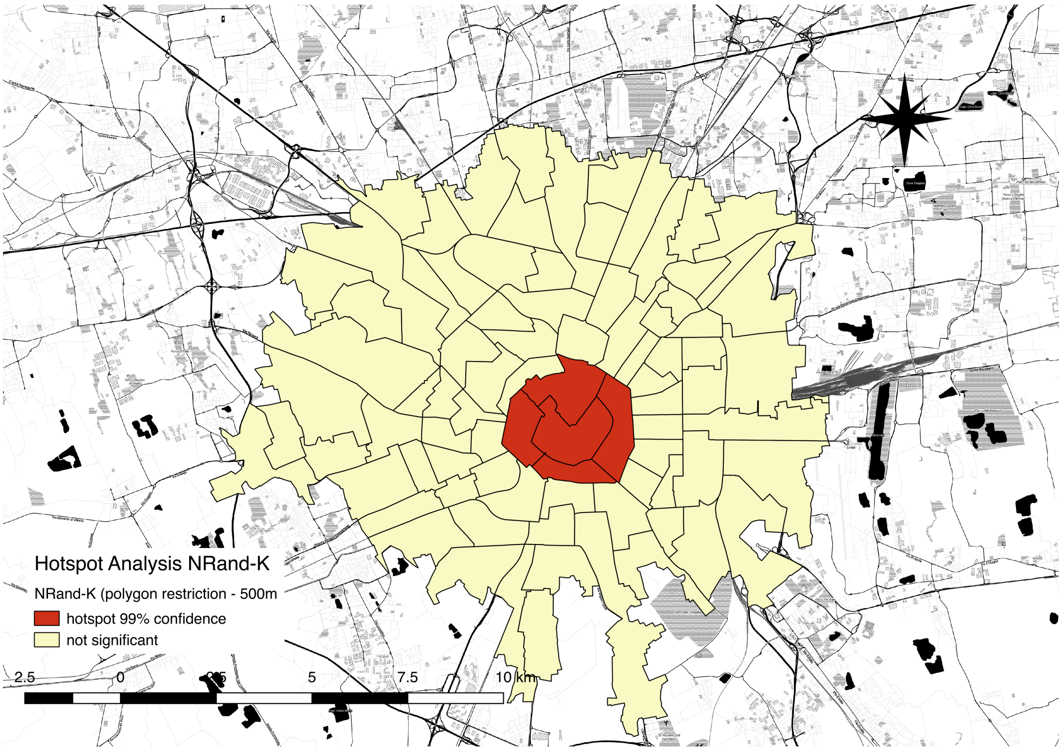


Figure 5.15: Output for hotspot analysis generated with obfuscated tweets in Milano, Italy with $r_{max} = 500$m and polygon restriction

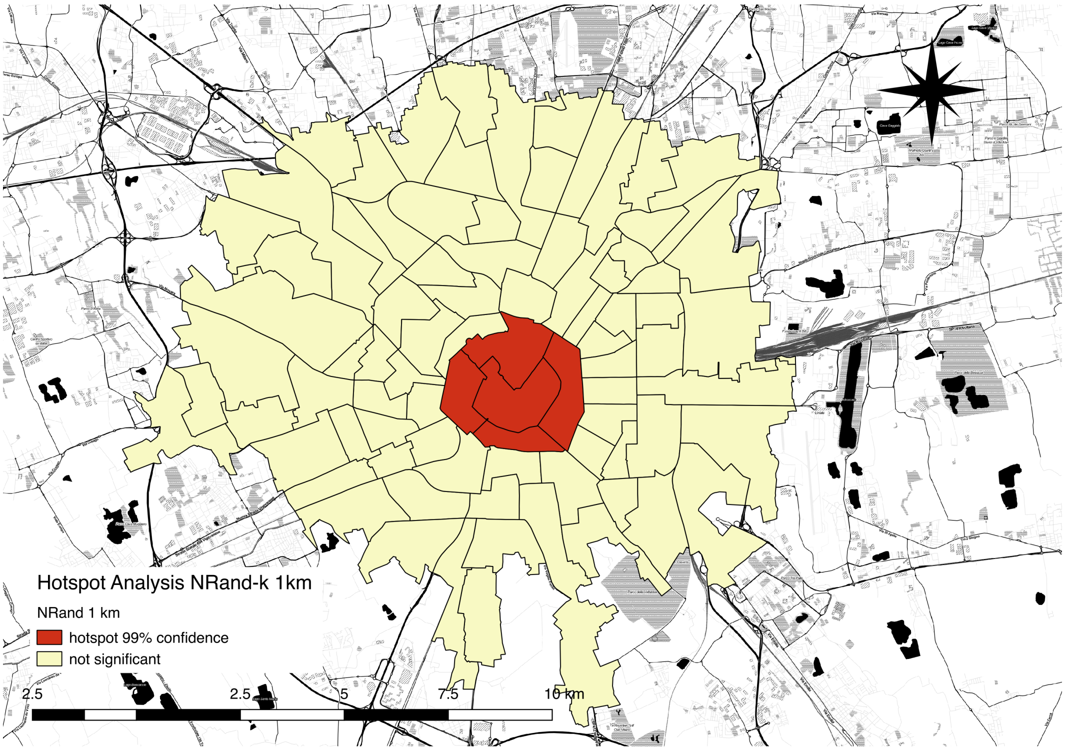


Figure 5.16: Output for hotspot analysis generated with obfuscated tweets in Milano, Italy with $r_{max} = 1$km

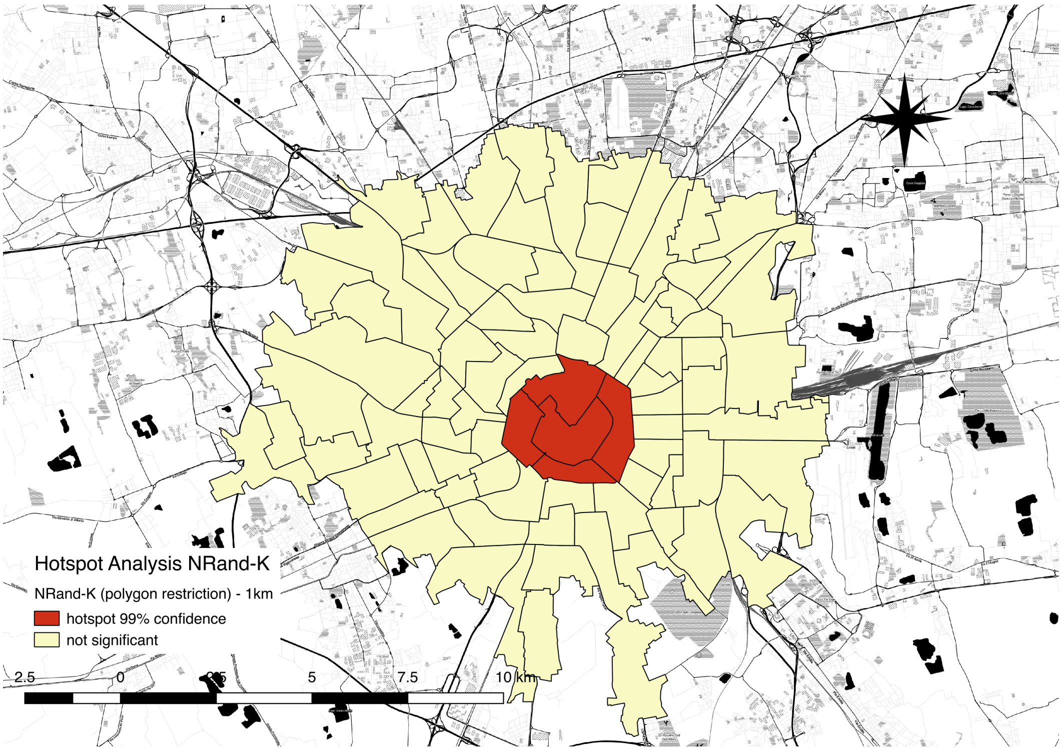


Figure 5.17: Output for hotspot analysis generated with obfuscated tweets in Milano, Italy with $r_{max} = 1$km and polygon restriction

the heatmap output with NRand-K where $r_{max} = 500$m, the highest concentration of tweets is of 13.2 compared to 13.3 in the heatmap with the original dataset; differing only by a value of 0.1 in terms of heatmap density. With an $r_{max} = 1$km the output in Figure 5.10 is visually more different from the original data heatmap than the one produced with an $r_{max} = 500$m. In the legend for this map, the density value for the highest concentration is 13.6, differing by 0.3 from 13.3 which was the highest concentration for the original data. However, the heatmap generated using NRand-K with polygon restriction and $r_{max} = 1$km, shows a maximum value of 13.4, which is closer to 13.3.

Analysing the heatmap divergence for both variation of NRand-K in Figures 5.12 and 5.13, again show fewer differences when the $r_{max} = 500$m is used. The average value for this measure can be found and compared in the last column of Table 5.1, evidencing what the heatmap figures suggest if comparing between both NRand-K variations: a greater difference when the polygon restriction is enforced in the algorithms for both $r_{max}$ scenarios.

The relation inferred from the NRand-K heatmap comparison using the values in the scales of the legends is not as clear as it was with the Pinwheel algorithm; the more induced noise, a greater decrease in the scale values. Since the noise induced by NRand-K not only depends on an $r_{max}$ parameter, but includes the criteria of a minimum K point count per neighbourhood that defines if an $r_{max}$ or an an $r_{min}$ is used; the distribution of the data cannot be said to become more sparse even if a greater $r_{max}$ is used, this mechanism rather aims to ensure that regardless the $r_{max}$ used, density clusters present in the original dataset are preserved and this is evidenced in these experiments.

The resulting outputs for hotspot analysis in both $r_{max}$ variations; Figure 5.14, Figure 5.16 and Figure 5.17 are identical to the one produced with the original tweet coordinates in Figure 5.2. These results support what was inferred from the heatmap calculation; the NRand-K preserves better the cluster density regardless of the $r_{max}$ when compared to the Pinwheel mechanism. It is worth mentioning, that since the input for the hotspot analysis is the polygon shapefile with their respective counts, the NRand-K mechanism with polygon restriction will preserve the hotspot analysis result in all cases.

### 5.3.4 VoKA

According to the VoKA algorithm there are two stages as described in Section 4.2, namely, Grid Reduction and KD Aggregation. For the Grid Reduction stage, a 4 digit reduction was chosen; this translates into approximately 10m of accuracy loss. For the KD-Aggregation stage, a distance value D of 500m was used.

Figure 5.18 shows the output for the heatmap calculation with the tweet coordinates after applying the VoKA obfuscation algorithm. When comparing the values on the heatmap scale to

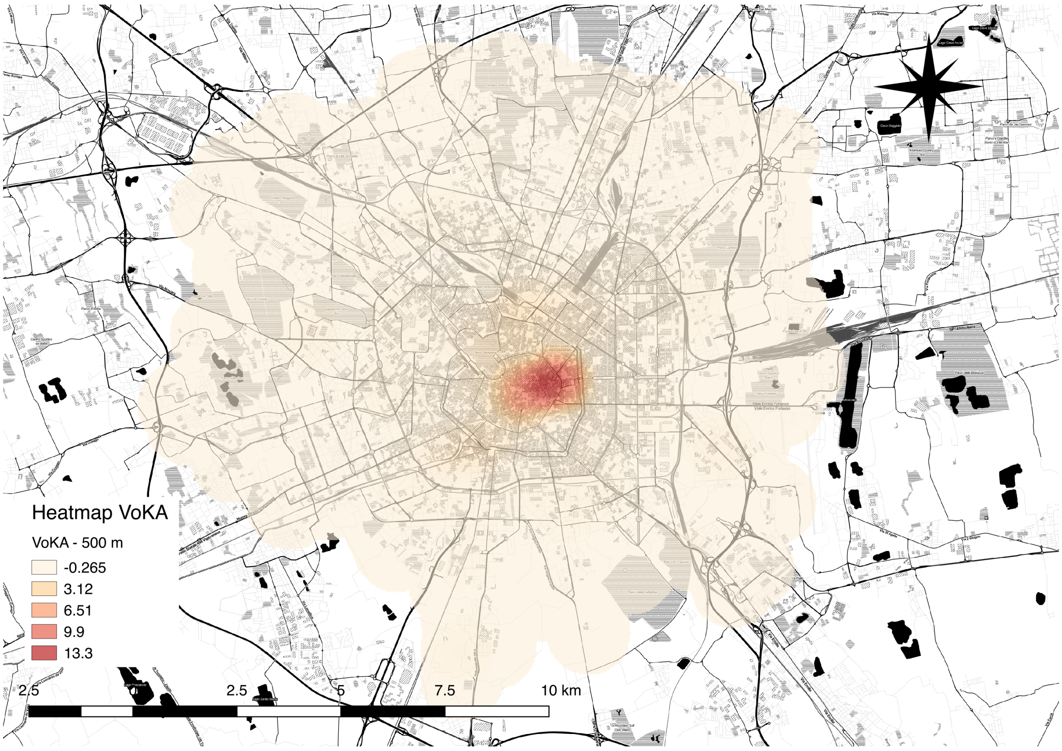


Figure 5.18: Heatmap generated with obfuscated tweets in Milano, Italy with 4 digit Grid Reduction and 500m used for KD-Aggregation

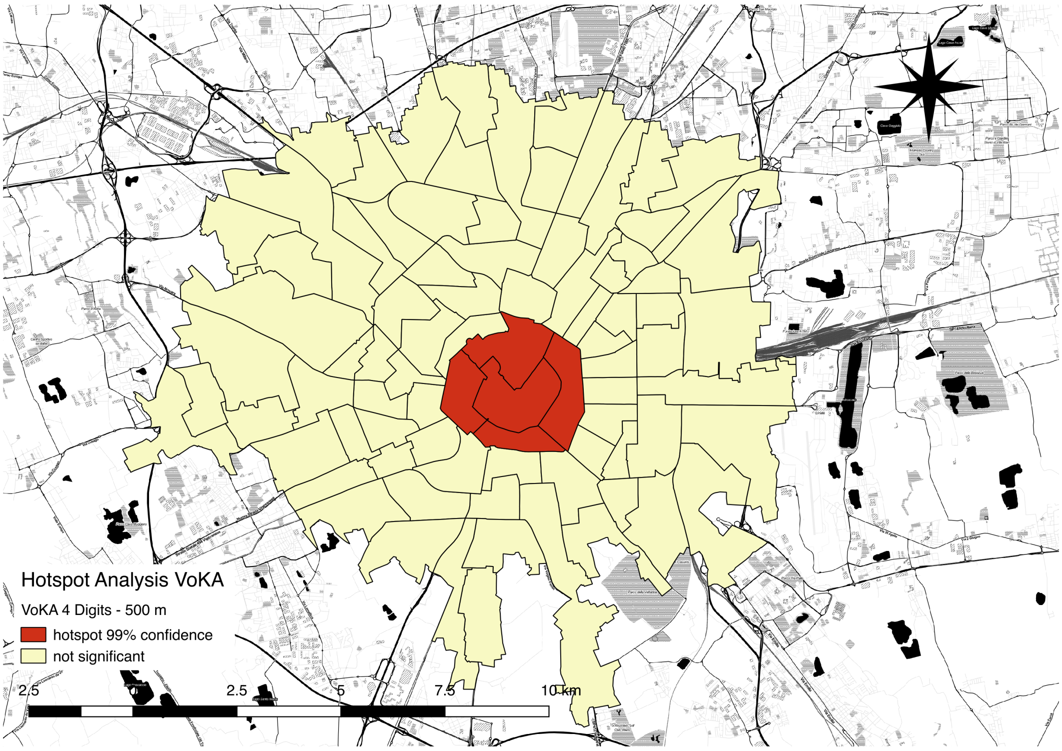


Figure 5.19: Output for hotspot analysis generated with obfuscated tweets in Milano, Italy with 4 digit Grid Reduction and 500m used for KD-Aggregation

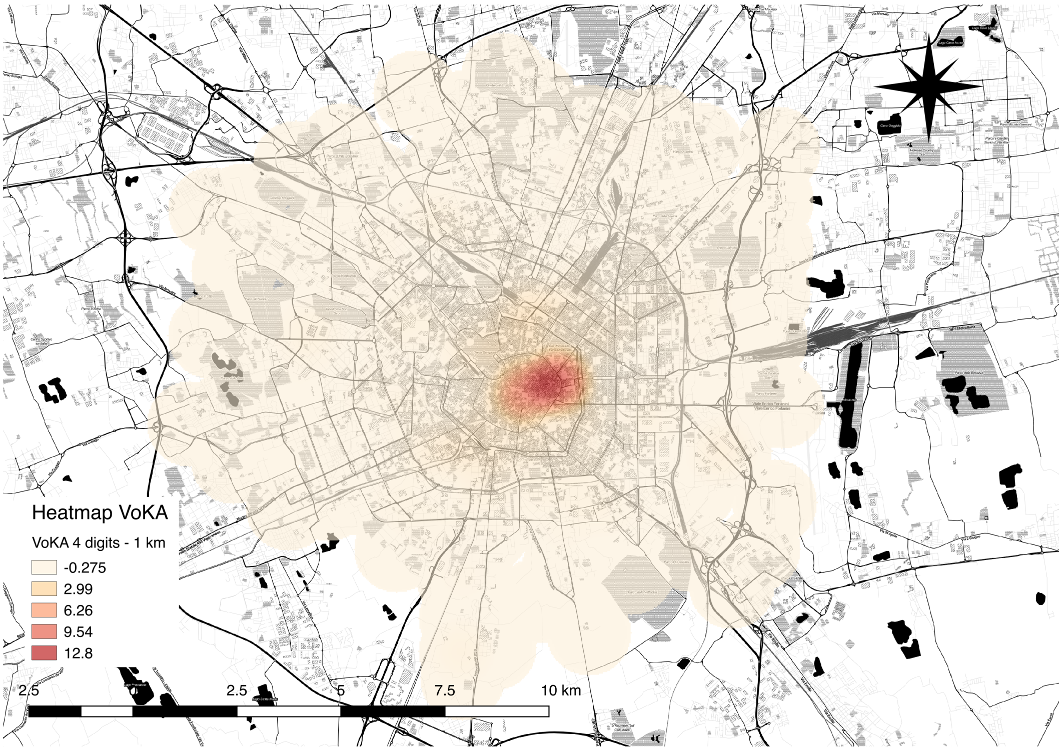


Figure 5.20: Heatmap generated with obfuscated tweets in Milano, Italy with 4 digit Grid Reduction and 1km used for KD-Aggregation

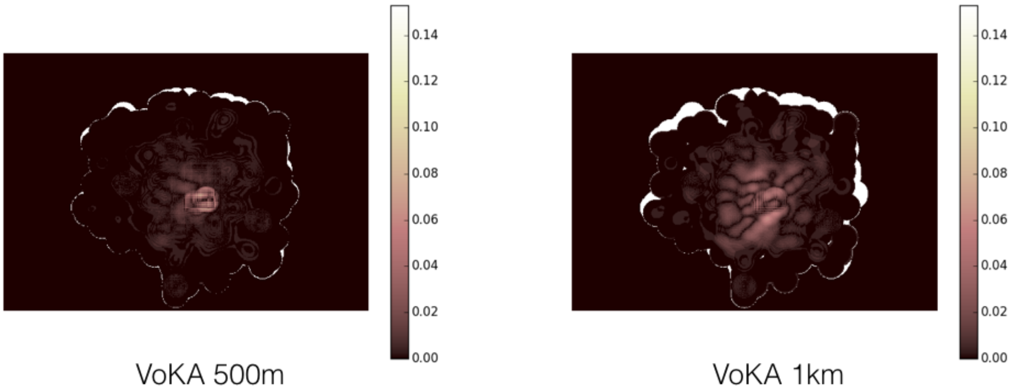


Figure 5.21: Heatmap divergence by pixels between the original data heatmap and the VoKA heatmaps generated with an $r_{max} = 500m$ and $r_{max} = 1km$

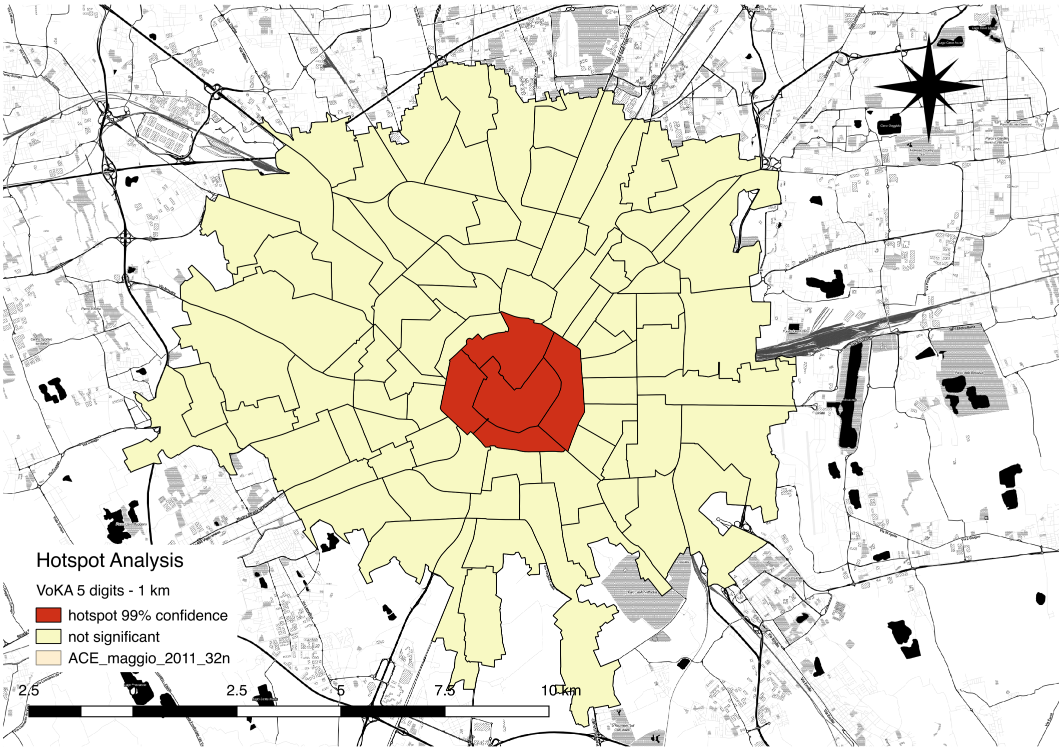


Figure 5.22: Output for hotspot analysis generated with obfuscated tweets in Milano, Italy with 4 digit Grid Reduction and 1km used for KD-Aggregation

those of the original data heatmap, these are very similar; differing only by 0.02 in the lowest temperature and equal in the highest. This reflects that an *aggregation based on the original data distribution reduces the impact on geospatial analysis more than when aggregating by administrative boundaries*; as it was the case for the experiments with NRand-K. Figure 5.20 shows the output for a heatmap done with obfuscated data using VoKA with 1km of distance for KD-Aggregation. Here, the maximum temperature is of 12.8, decreasing by 0.5 from the heatmap with the original data, and suggesting a greater impact than that of NRand-K with an $r_{max}$ of 1km.

The divergence heatmap of VoKA with an $r_{max} = 500$m shows the least alterations or highlights when compared to the one of $r_{max} = 1$km as well as to all other evaluated scenarios. This is better seen when comparing the mean heatmap error in Table 5.1, where the one for VoKA with an $r_{max} = 500$m was 6.67, the smallest error of all tests.

As for the hotspot analysis in Figures 5.19 and 5.22, which resulted in identical outputs as the one produced with the original dataset, proving the preservation of the hotspots identified with the original tweets; as it was the case as well with NRand-K mechanism analysed in the previous subsection.

## 5.4 Comparison

In order to calculate the Spatial Indices, the original dataset was used to obtain values necessary for these calculations. For the MDi described in Equation 5.1, the $d(\text{original mean}, \text{farthest point})$ resulted in a value of **10,610m**; for the ODi described in Equation 5.2, the SDE orientation angle was **63.51**°. It is worth noting that all indices presented here are better if the resulting value is close to zero. The following table shows the resulting values for each experimental scenario described in the LPPM column, the euclidean distance for the means, the results for MDi, the SDE orientation angle, the resulting ODi and the PCDi. The values in bold are the ones that performed best from all tested scenarios.

| LPPM | d(original mean, obfuscated mean) | MDi | SDE orientation | ODi | PCDi | Heatmap Mean Error |
|---|---|---|---|---|---|---|
| Pinwheel 500m | **5.84 m** | **0.06** | 60.05° | 1.93 | 67.74 | 12.04 |
| Pinwheel 1km | 13.65m | 0.13 | 48.54° | 8.32 | 102.86 | 31.3 |
| NRand-K 500m | 20.05m | 0.19 | 53.76° | 5.42 | 25.52 | 16.03 |
| NRand-K 1km | 18.07m | 0.17 | 48.10° | 8.56 | 31.81 | 20.73 |
| NRand-K 500m (polygon restriction) | 22m | 0.21 | 56.83° | 3.71 | **0** | 20.18 |
| NRand-K 1km (polygon restriction) | 18.32m | 0.17 | 54.77° | 4.86 | **0** | 22.35 |
| VoKA 500m | 88.06m | 0.83 | **64.84**° | **0.74** | 18.68 | **6.67** |
| VoKA 1km | 246.4m | 2.32 | 66.4° | 3.53 | 59.56 | 17.33 |

Table 5.1: Comparison of spatial indices

From the results obtained by calculating the described indices, the VoKA mechanism is the one that obtained the best results in most cases; it has the lowest orientation and point count divergence; however it produced the highest value for the MDi. On the other hand, the mechanism that produced a mean coordinate most similar to the original dataset was Pinwheel when an $r_{max} = 500$ m was used, and its ODi value follows that of VoKA being the second lowest; indicating less impact in the orientation of the data. The NRand-K mechanism performed averagely for the ODi when an $r_{max} = 500\text{m}$ was used, but shows more similarities when the polygon restriction is enforced, in this case the PCDi is 0, as expected.

## 5.5 Discussion

Spool (2011) states that the default configuration for an application will not be altered by most users, making the definition of defaults critical for the future of a service. This consideration should be a priority for deploying applications that include privacy protection policies, which in LBS translates into finding the appropriate settings for LPPMs to work under essential and practical scenarios. Assessing the impact of LPPMs on the quality of information can provide an initial estimate on the best parameter values to reduce the degradation while guaranteeing a minimum level of protection. These values could serve as default values that later on, the LBS manager can adjust to a specific application, even for real-time LPPMs.

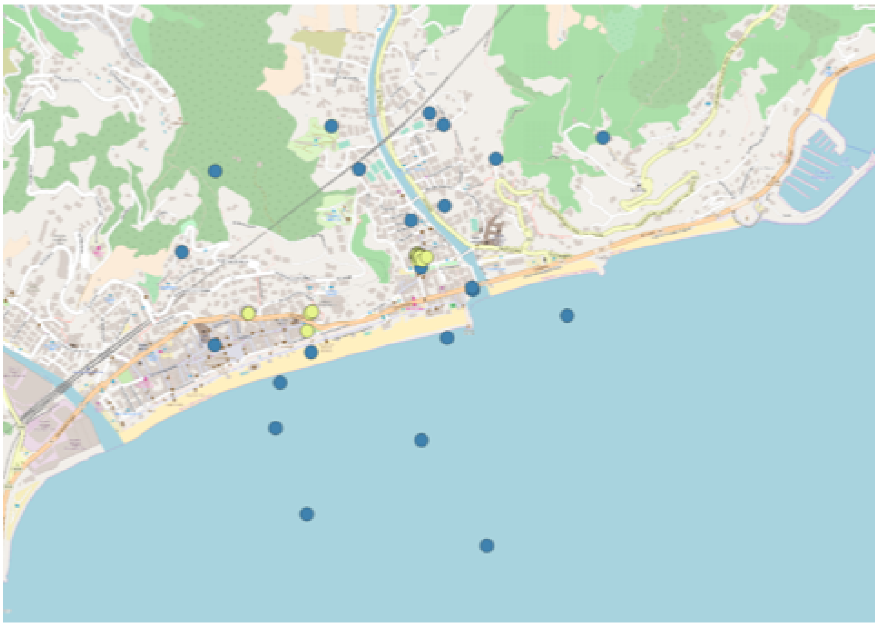


Figure 5.23: Obfuscated locations (blue dots) falling outside study boundaries

The results of these experiments to assess the geospatial analysis impact, highlight the importance of adjusting parameters for both obfuscation algorithms and analysis tools. The scale of the study is critical to determine a feasible noise to induce; while 500m can work for performing geospatial analysis in a city, a noise of 5km can also perform well if the analysis area is in the scale of a country. The results obtained here prove that if correctly parametrised, a privacy preserving geospatial analysis can be done with similar outputs as if the data was unprotected.

It is worth noting that none of the presented algorithms introduce any knowledge of the geographical area of the study hence, being prompt to produce obfuscated locations falling in improbable areas as evidenced in Figure 5.23.

## 5.6 Chapter Summary

This chapter presents a methodology to compare different obfuscation LPPMs using different spatial analysis and indices, the results conclude that geospatial analysis is still possible and valid even with obfuscated data.

# Chapter 6

# Discussion

## 6.1 Chapter Introduction

Since the emergence of location-aware devices, the offer of LBS should have been accompanied by the mechanisms necessary to ensure privacy. The fact that it is only in recent years that we have become *aware* [1] of this is due to unseen risks, from which we are now suffering.

LPPMs should evolve as LBS do. In order to create relevant protection mechanisms, the growing spectrum of LBS functionalities is a challenge for their implementation in commercial applications; companies may not be willing to experiment with privacy algorithms, at the risk of losing accuracy or compromising data analysis, which is part of their business model. However, as the user community become aware of the value of their privacy, this may become an incentive for further research driven to commercial adoption of location privacy solutions; specially, if accompanied by legal enforcement from governmental institutions or international organisations.

## 6.2 Geoprivacy Incidents

Xu and Gupta (2009) mention how user's privacy concerns translate into the fear of losing control of personal information and can even lead to cause stress and anxiety. This is ameliorated by the adoption of protective measures, the perception of privacy and users' trust on the providing companies of a service can increase, leading to more compliance to adopt a LBS.

The recent events of the Strava fitness application that unintentionally gave away the positions of military bases reported in (Hern, 2018) and the Facebook incident documented in (Schechter,

[1] *"having or showing realization, perception, or knowledge" from Merriam Webster (a).*

2018), that allowed Cambridge Analytica[2] to manipulate the United States presidential elections, 2016, or as described in their slogan: *"Data drives all we do. Cambridge Analytica uses data to change audience behaviour."* are just some of the possibilities that the existence of geo-referenced big data enable. Geographic information is not a dismissible component of big data nowadays, it is key to conducting analysis; since data analysis missing context is unlikely to provide informed results.

What is particular in the case of Facebook or any other platform that allow the introduction of promoted unregulated information is their potential to not only gather data, but to process it and feed it back to the user according to the user's specific profile; while serving their clients advertising interests. In this scenario, the user is off course accountable for voluntarily engaging in the use of these services and for neglecting the exercise of informed judgement on topics of her interests, but is there any real alternative other than to renounce the use of these platforms?

The ethical dilemma that this poses is not new. If unrestricted, companies will take as much advantage of their resources to be profitable; machiavellian[3]. It is not their lawful duty to see for the preservation of rights of individuals. It is the job of governmental institutions to regulate supported by non-governmental organisations (NGOs), e.g., the United Nations.

## 6.3 Privacy Advocates

In the pursuit for finding privacy advocates in this era, two organisations stand out: the Electronic Frontier Foundation (EFF, 2016), based in the U.S and Privacy International (Privacy International, a), based in the U.K. These organisations are established on the common principle of defending individual's rights in a digital world; Privacy International specifically focusing on the right to privacy. Among the many activities and projects they support, engaging with NGO is one of them. According to Privacy International (2017):

> Engaging with the UN human rights mechanisms is not an end in itself: it serves to develop common understanding of the scope of the right to privacy, spells out the obligations of states to uphold the right, and the responsibilities of companies to respect the right. Very often, recommendations by the UN serve national advocacy strategy: the recommendations of these mechanisms can be fundamental to supporting national campaign for changes in laws and practices. Additionally, some UN bodies can offer a mechanism of redress for victims of violations and act as an

[2] cambridgeanalytica.org

[3] from Machiavellianism: *"the political theory of Machiavelli; especially: the view that politics is amoral and that any means however unscrupulous can justifiably be used in achieving political power"* defined in (Merriam Webster, b).

> international 'watchdog', by, for example, raising concerns of individuals at risk or legislative proposals that could, if implemented, violate the right to privacy.

Privacy advocates can raise awareness, but it is a long way for regulations to be in place given that ultimately, each country is in charged of their jurisdiction. However:

> Data isn't stopping at the borders and our rights and protections shouldn't either. Companies are using countries with low protections as testing ground for worst practices.
> Privacy International (b)

This thesis argues that it is a matter of international laws to regulate privacy on digital services, since service providers may take advantage of poor privacy regulations in specific countries to exploit data that originates anywhere. If the situation is already challenging in areas with clear privacy regulations, what about those without minimum privacy guidelines or non-existent?

## 6.4 Towards Ethical Geodata Management

The implementation of a regulatory measure for location privacy may happen in the near future, at least by political unions like the European Union where reports like the one presented in (Winterbottom et al., 2009) attempt to provide clear standards for geoprivacy and that in general, the EU is more protective of the user's privacy rights when compared to other states. Under this scenario, it is critical to understand the impact that regulatory measures for geoprivacy may have on deospatial analysis. This thesis assesses geospatial analysis with different location privacy protection mechanisms to provide a baseline of what it might be like in a future with an ethical geodata management.

The complexity of the situation given the diverse actors and interests involved, begs at least the following question: *When does CGI data becomes usable in the way VGI data would?*

Which is quickly followed by: *What are the limits of VGI data usability, if any?*

This thesis takes the view that we are all stakeholders in this matter; service providers do have a great moral responsibility. A responsibility that should perhaps be decentralised; if an open data approach is carried out, considering releasing data with implemented privacy protection mechanisms accessible to not only a few, but to everyone; the relation between LBS providers and users could become symbiotic. In this way, users would not only be generators of the *product*, but benefit from it; avoiding the need of recurring to specialised companies with privileged access to this data, which in most cases is not accessible to laymen.

While this research is of a technical nature with ethical implications, such implications are out of the scope of this work and it stands an open discussion. However if taking an ethical standpoint, this work is more inclined towards Kant's Categorical Imperative[4], than a machiavellian approach; in this regard, service providers should not only wish for their own well-being and *happiness*, but consider their yet unregulated actions as becoming a universal law and their consequences.

A popular argument that pro-surveillance organisations present is the fact that if someone has nothing to hide, then privacy should not be a concern. On the contrary, universal rights exist for everyone regardless of their conditions or in other words:

> "The main problem with the *nothing to hide* argument is assuming that privacy is important only if you have something to hide," said Ignacio Cofone, a New York University research fellow and privacy expert. "Privacy is portrayed, then, as something for deviants. If we start arguing how everyone has something to hide, then we've already lost." Harding (2018)

This chapter draws together the case studies and LPPMs examined in the previous chapters to answer the research question of this thesis: **How to provide location privacy with mechanisms that are scalable, efficient, non-intrusive and with low impact on geospatial analysis?**.

Chapter 3 and Chapter 4 were dedicated to describe the implementation of proposed LPPMs that comply with being *scalable, efficient, non-intrusive* and how these can be adopted by service providers, while Chapter 5 describe how it will reflect on service providers geodata management model in respect to having a *low impact on geospatial analysis*.

## 6.5 Chapter Summary

This chapter finalises the discussion of the concepts and the technical work that this thesis has undertaken. Taking a view on further ethical implications that are up to discussion. The proceeding chapter concludes this thesis.

[4] *"act only in accordance with that maxim through which you can at the same time will that it become a universal law"* from (Johnson and Cureton, 2018)

# Chapter 7

# Conclusion

## 7.1 Chapter Introduction

This thesis examined geoprivacy; the existing techniques and proposed new mechanisms considering a balance between privacy protection and realistic interest of LBS providers. In doing so, two stages for geodata protection were identified: While the LBS is being used and after the LBS collected the data from its users. This chapter concludes the thesis, examining the research questions, aims and objectives and how they were achieved. Further it discusses the limitations of this research, sets a future research agenda and closes this thesis.

## 7.2 Meeting the Research Questions, Aims and Objectives

Chapter 1 introduced the research question, aims and objectives of investigating location privacy, its applications and implications. This necessitated an approach to first examine the existing mechanisms in the literature, then to demarcate a canon for what geoprivacy protection mechanisms should do and then to design such mechanisms. Going further, in order to prove that the developed mechanisms comply with the research question, a replicable process involving geospatial analysis was presented as a model to quantify the impact of such protection mechanisms. The following subsections examine how the research questions, aims and objectives were met in the context of this research.

### 7.2.1 Identifying Common Characteristics of LPPMs

LPPMs aim to protect users geographic information from LBS or from third parties. In order to do so, the different existing techniques in the literature were categorised under three types according to the type of LBS they can serve, i.e., reactive, proactive or both. This classification allowed to understand the characteristics of the numerous mechanisms that may use one or more of these techniques to achieve location privacy; with variations in their approaches and their specific applications. The characteristics identified served to compare between the surveyed mechanisms and establish a canon to develop geoprivacy algorithms that were feasible to implement in existing LBS.

The identified characteristics that an ideal LPPM should comply with were identified as foolows: serve for all types of LBS, provide as much privacy as possible (the entire spatial domain), while not requiring the use of a third party, and that does not alter the existing LBS architecture. These ideal characteristics are hard to meet, but some trade-offs are acceptable: a mechanism that does not alter the LBS but that provides adjustable privacy protection in a decentralised way and without including a third party is the viable option. From here on, the thesis focused mainly on location obfuscation algorithms that proved to meet these requirements in both Table 2.1 of Chapter 2 and Table 3.1 of Chapter 3.

### 7.2.2 Developing Scalable, efficient & non-intrusive LPPMs

After identifying that the ideal but feasible characteristics to provide to an LPPM are possible with the use of the location obfuscation technique, privacy protection mechanisms based on this technique were developed for their implementation during real-time use of an LBS. The presented mechanisms for real-time use ($\theta$-Rand, Pinwheel & Near-Rand) are *scalable* in the way that the maximum noise threshold is possible to fix according to the necessity of the service; *efficient*, because the computational cost for $\theta$-Rand and Pinwheel is constant $O(1)$, and for Near-Rand is linear $O(n)$, not taking more than milliseconds to process in a mobile device; and non-intrusive, since none of these mechanisms require alterations in the LBS to enable their functionality.

### 7.2.3 Minimising Geospatial Analysis Impact

By analysing the behaviour of the outputs in geospatial analysis with location obfuscation techniques before and after collection, it was identified that aggregation based on the original data distribution reduces the impact on geospatial analysis. Mechanisms like NRand-K and VoKA are a contribution in this aspect, and are not necessarily only applicable after the collection has occurred; if enough historical geographic information is considered, the aggregation can be performed according to those historical records while the LBS is being used.

## 7.3 Suggested Further Research

This research has illuminated the need to consider not only LBS functionalities, but geodata usability for analysis when developing geoprivacy techniques. There are other aspects to geospatial analysis that this thesis does not cover, such as semantic analysis for contextual information and other implications that alterations in geographic information may have in the LBS agenda. It is encouraged to follow up the developments of geoprivacy violation issues, regulations, and the pertinency of storing reported locations for unlimited time.

## 7.4 Final Conclusions

The conclusion of this thesis is that performing reverse engineering in algorithms for geospatial analysis, provides a common ground for understanding how these analyses can be affected by alterations in geographic information. This understanding leads to the creation of mechanisms that leverage users and geodata stakeholders interests to engage in privacy protective measures without risking the possibility of performing geospatial analysis.

Arguably, this is the first study of LPPMs of this depth combining the interests of both users for privacy protection and LBS providers for geodata re-usability. Demonstrably, this thesis is very timely due to numerous recent outrages that digital privacy violations imply. Even though the *geo* component is just one dimension of digital privacy, it is of great importance in any data analysis; for without context, the analysis would lack relevance.

While the problem of risking location privacy was foreseen by the academic community since the emergence of LBS, it is now that by ignoring this, the consequences are evidenced in a catastrophic manner. It is not hopeless; however to think that location privacy mechanisms will be in place for the continuity of massive adoption of LBS.

# Bibliography

Aggarwal, C. C. (2005). On k-anonymity and the curse of dimensionality. In *Proceedings of the 31st international conference on Very large data bases*, pp. 901–909.

Ardagna, C., M. Cremonini, S. De Capitani Di Vimercati, and P. Samarati (2011, jan). An obfuscation-based approach for protecting location privacy. *IEEE Transactions on Dependable and Secure Computing 8*(1), 13–27.

Ardagna, C. A., M. Cremonini, E. Damiani, S. De Capitani di Vimercati, and P. Samarati (2007). Location Privacy Protection Through Obfuscation-Based Techniques. pp. 47–60.

Armstrong, M. P., G. Rushton, and D. L. Zimmerman (1999). Geographically Masking Health Data to Preserve Confidentiality. *STATISTICS IN MEDICINE Statist. Med 18*, 497–525.

Bellavista, P., A. Kupper, and S. Helal (2008, apr). Location-Based Services: Back to the Future. *IEEE Pervasive Computing 7*(2), 85–89.

Beller, M. and J. Leerssen (2007). *Imagology: The cultural construction and literary representation of national characters. A critical survey*. Number 13. AmsterdamRodopi.

Beresford, A. R. and F. Stajano (2004). Mix zones: User privacy in location-aware services. pp. 127–131.

Cheng, R., Y. Zhang, E. Bertino, and S. Prabhakar (2006). Preserving user location privacy in mobile data management infrastructures. *Privacy Enhancing . . .* , 393–412.

Chor, B. and N. Gilboa (1997). Computationally private information retrieval. *Proceedings of the twenty-ninth annual ACM symposium on Theory of computing*, 304–313.

Chor, B., O. Goldreich, E. Kushilevitz, and M. Sudan (1998, nov). Private information retrieval. *Journal of the ACM 45*(6), 965–982.

Chow, C.-Y., M. F. Mokbel, and W. G. Aref (2009, dec). Casper*: Query processing for location services without compromising privacy. *ACM Transactions on Database Systems 34*(4), 1–48.

Congreso de Colombia (2012). Ley 1581 de 2012.

Department Of Defense, U. (2008). Global Positioning System Standard Positioning Service. *Www.Gps.Gov* (September), 1 – 160.

Duckham, M. and L. Kulik (2005, jan). A Formal Model of Obfuscation and Negotiation for Location Privacy. Lecture Notes in Computer Science, pp. 152–170. Springer Berlin Heidelberg.

Duckham, M. and L. Kulik (2006). Location privacy and location-aware computing. *Dynamic & mobile GIS: investigating change in space and time 3*, 35–51.

ECHR (1998). Human Rights Act 1998.

EFF (2016). Electronic Frontier Foundation | Defending your rights in the digital world.

EU (2002). Directive (EU) 2002/58/EC: Directive on privacy and electronic communications. *Official Journal of the European Communities L*(201), 37–47.

EU (2016). Regulation (EU) 2016/679: General Data Protection Regulation. *Official journal of the European Union*.

European Commission (2016, nov). Guidelines for public administrations on location privacy: European Union Location Framework - EU Science Hub - European Commission.

European GNSS Agency (2010, may). Opportunities abound in growing location-based services market.

Franken, A. (2015). S.2270 - 114th Congress (2015-2016): Location Privacy Protection Act of 2015.

Gambs, S., M.-O. Killijian, and M. N. del Prado Cortez (2010). Show me how you move and I will tell you who you are. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS - SPRINGL '10*, pp. 34.

Ghinita, G., P. Kalnis, and A. Khoshgozaran (2008). Private Queries In Location Based Services Anonymizers Are Not Necessary Categories And Subject Descriptors. *Proc. of the 2008 ACM SIGMOD . . . 8*(ii), 1–12.

Gong, Z., G. Z. Sun, and X. Xie (2010). Protecting privacy in location-based services using K-anonymity without cloaked region. In *Proceedings - IEEE International Conference on Mobile Data Management*, pp. 366–371.

Goodchild, M. F. (2007, apr). Citizens as sensors: The world of volunteered geography.

Google (2014). Understanding Consumers' Local Search Behavior. (May), 2014.

Gruteser, M. and D. Grunwald (2003). Anonymous Usage of Location-Based Services Through Spatial and Temporal Cloaking. In *Proceedings of the 1st international conference on Mobile* *systems, applications and services - MobiSys '03*, MobiSys '03, New York, NY, USA, pp. 31–42. ACM.

Harding, X. (2018). âĂIJWho cares, I have nothing to hideâĂİ âĂŤ Why the popular response to online privacy is so flawed.

Harvey, F. (2013). To volunteer or to contribute locational information? Towards truth in labeling for crowdsourced geographic information. In *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice*, Volume 9789400745, pp. 31–42. Dordrecht: Springer Netherlands.

Hern, A. (2018). Fitness tracking app Strava gives away location of secret US army bases.

Hjaltason, G. R. and H. Samet (1999). Distance browsing in spatial databases. *ACM Transactions on Database Systems 24*(2), 265–318.

Huang, X., X. Huang, M. L. Yiu, M. L. Yiu, H. Lu, H. Lu, C. S. Jensen, and C. S. Jensen (2008). SpaceTwist: Managing the Trade-Offs Among Location Privacy, Query Performance, and Query Accuracy in Mobile Services. *00*, 366–375.

Iliffe, M. P. (2017). *The Praxis of Community Mapping in Developing Countries*. Ph. D. thesis.

Johnson, R. and A. Cureton (2018). Kant's Moral Philosophy.

Kalnis, P., G. Ghinita, K. Mouratidis, and D. Papadias (2007, dec). Preventing location-based identity inference in anonymous spatial queries. In *IEEE Transactions on Knowledge and Data Engineering*, Volume 19, pp. 1719–1733.

Khoshgozaran, A. and C. Shahabi (2007). Blind Evaluation of Nearest Neighbor Queries Using Space Transformation to Preserve Location Privacy. *Sstd 4605*, 239–257.

Khoshgozaran, A., C. Shahabi, and H. Shirani-Mehr (2011). Location privacy: Going beyond K-anonymity, cloaking and anonymizers. *Knowledge and Information Systems 26*(3), 435–465.

Khoshgozaran, A., H. Shirani-Mehr, and C. Shahabi (2008, apr). SPIRAL: A Scalable Private Information Retrieval Approach to Location Privacy. *Mobile Data Management Workshops, 2008. MDMW 2008. Ninth International Conference on*, 55–62.

Kido, H., Y. Yanagisawa, and T. Satoh (2005). An anonymous communication technique using dummies for location-based services. In *Proceedings - International Conference on Pervasive Services, ICPS '05*, Volume 2005, pp. 88–97.

Koufogiannis, F. and G. J. Pappas (2016, dec). Location-dependent privacy. In *2016 IEEE 55th Conference on Decision and Control, CDC 2016*, pp. 7586–7591.

Koukoletsos, T. (2012). *A Framework for Quality Evaluation of VGI linear datasets Thesis* *submitted for the Degree of Doctor of Philosophy ( PhD ) University College London ( UCL ) Author ' s Declaration*. Ph. D. thesis.

Kounadi, O. and M. Leitner (2015, oct). Spatial Information Divergence: Using Global and Local Indices to Compare Geographical Masks Applied to Crime Data. *Transactions in GIS 19*(5), 737–757.

Krumm, J. (2007, jan). Inference Attacks on Location Tracks. In A. LaMarca, M. Langheinrich, and K. N. Truong (Eds.), *Pervasive Computing*, Lecture Notes in Computer Science, pp. 127–143. Springer Berlin Heidelberg.

Kushilevitz, E., R. Ostrovsky, and T. Bellcore (1997). Replication is not needed: Single database, computationally-private information retrieval. *{IEEE} Symposium on Foundations of Computer Science*, 364–373.

KuÌĹpper, A. (2005). *Location-based services : fundamentals and operation*. John Wiley.

Kwan, M.-P., I. Casas, and B. Schmitz (2004). Protection of Geoprivacy and Accuracy of Spatial Information: How Effective Are Geographical Masks? *Cartographica: The International Journal for Geographic Information and Geovisualization 39*(2), 15–28.

Labrador, M. A. and P. M. Wightman (2009). *Topology Control in Wireless Sensor Networks âĂŞ with a companion simulation tool for teaching and research*. Springer Publishing Company, Incorporated.

Labrador, M. A., A. J. Perez, and P. M. Wightman (2010). Location-Based Information Systems: Developing Real-Time Tracking Applications.

Lu, H. and C. S. Jensen (2008). PAD: Privacy-Area Aware, Dummy-Based Location Privacy in Mobile Services. *MobiDE*, 16–23.

Mascetti, S., D. Freni, C. Bettini, X. S. Wang, and S. Jajodia (2011). Privacy in geo-social networks–proximity notification with untrusted service providers and curious buddies.

Mavridis, P. (2012). Visualizing the difference of two images as a Heatmap.

Merriam Webster. Aware | Definition of aware by Merriam-Webster.

Merriam Webster. Machiavellianism | Definition of Machiavellianism by Merriam-Webster.

Micciancio, D. (2010, mar). A first glimpse of cryptography's Holy Grail. *Communications of the ACM 53*(3), 96.

Monreale, A., W. H. Wang, F. Pratesi, S. Rinzivillo, D. Pedreschi, G. Andrienko, and N. Andrienko (2013). Privacy-preserving distributed movement data aggregation. *Lecture Notes in Geoinformation and Cartography 2013-Janua*, 225–245.

Olumofin, F., P. K. Tysowski, I. Goldberg, and U. Hengartner (2010). Achieving efficient query privacy for location based services. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Volume 6205 LNCS, pp. 93–110.

Ord, J. K. and A. Getis (1995, oct). Local Spatial Autocorrelation Statistics: Distributional Issues and an Application. *Geographical Analysis 27*(4), 286–306.

Oxoli, D., G. Prestifilippo, D. Bertocchi, and M. Zurbarán (2017). Enabling spatial autocorrelation mapping in QGIS: The hotspot analysis Plugin. *Geoingegneria Ambientale e Mineraria 151*(2), 45–50.

Papadopoulos, S., S. Bakiras, and D. Papadias (2010). Nearest neighbor search with strong location privacy. *Proceedings of the VLDB ... 3*(1), 619–629.

Perilla, J. R., O. Beckstein, E. J. Denning, and T. B. Woolf (2006). Location privacy and location-aware computing. *English 819*(3), 34–51.

Privacy International. PI Privacy International.

Privacy International. Why the Cambridge Analytica-Facebook scandal is a wake-up call for all governments: Seven steps for a global response | Privacy International.

Privacy International (2017). How to Talk About the Right to Privacy at the UN.

Quercia, D., I. Leontiadis, L. McNamara, C. Mascolo, and J. Crowcroft (2011). SpotME if you can: Randomized responses for location obfuscation on mobile phones. In *Proceedings - International Conference on Distributed Computing Systems*, pp. 368–372.

Rannenberg, K., D. Royer, and A. Deuker (2009a). *The Future of Identity in the Information Society (FIDIS) – Challenges and Opportunities*. Berlin; London: Springer.

Rannenberg, K., D. Royer, and A. Deuker (2009b). *The Future of Identity in the Information Society (FIDIS) – Challenges and Opportunities*. Berlin; London: Springer.

Riboni, D., L. Pareschi, and C. Bettini (2011, jun). Integrating identity, location, and absence privacy in context-aware retrieval of points of interest. In *Proceedings - IEEE International Conference on Mobile Data Management*, Volume 1, pp. 135–140. IEEE.

Schechter, A. (2018). Roger McNamee: âĂIJI Think You Can Make a Legitimate Case that Facebook Has Become ParasiticâĂİ -.

Seidl, D. E., P. Jankowski, and M.-H. Tsou (2015, apr). Privacy and spatial pattern preservation in masked GPS trajectory data. *International Journal of Geographical Information Science 8816*(December), 1–16.

Skyhook. WiFi Location Services for Marketing & Advertising.

Spool, J. (2011). Do users change their settings?

The Tor Project (2016). Tor Project: Anonymity Online.

United Nations (1948). Universal Declaration of Human Rights (United Nations).

van Diggelen, F. (2009, mar). *A-GPS: Assisted GPS, GNSS, and SBAS* (1 edition ed.). Boston: Artech House.

Wang, S. and X. S. Wang (2009). AnonTwist: Nearest neighbor querying with both location privacy and K-anonymity for mobile users. In *Proceedings - IEEE International Conference on Mobile Data Management*, pp. 443–448. IEEE.

Warren, S. D. and L. D. Brandeis (1890). The Right to Privacy. *Harvard Law Review 4*(5), 193.

Wieland, S. C., C. A. Cassa, K. D. Mandl, and B. Berger (2008). Revealing the spatial distribution of a disease while preserving privacy. *Proceedings of the National Academy of Sciences of the United States of America 105*(46), 17608–13.

Wightman, P., W. Coronell, D. Jabba, M. Jimeno, and M. Labrador (2011). Evaluation of location obfuscation techniques for privacy in location based information systems. In *2011 IEEE Latin-American Conference on Communications, LATINCOM 2011 - Conference Proceedings*, pp. 1–6.

Wightman, P. and M. Zurbarán (2018). An Initial Evaluation of the Impact of Location Obfuscation Mechanisms on Geospatial Analysis. In S. V. Ukkusuri and C. Yang (Eds.), *Transportation Analytics in the Era of Big Data*, pp. 28. Springer International Publishing.

Wightman, P., M. Zurbaran, and A. Santander (2014). High variability geographical obfuscation for location privacy. In *Proceedings - International Carnahan Conference on Security Technology*.

Wightman, P. M., M. A. Jimeno, D. Jabba, and M. Labrador (2012). Matlock: A location obfuscation technique for accuracy-restricted applications. In *IEEE Wireless Communications and Networking Conference, WCNC*, pp. 1829–1834.

Wightman, P. M., M. Zurbarán, M. Rodríguez, and M. A. Labrador (2013). MaPIR: Mapping-based private information retrieval for location privacy in LBISs. In *Proceedings - Conference on Local Computer Networks, LCN*, pp. 964–971.

Williams, P. and R. Sion (2008). Usable PIR. *Network Security 10*, 329–37.

Winterbottom, J., M. Thomson, and H. Tschofenig (2009). GEOPRIV Presence Information Data Format Location Object (PIDF-LO) Usage Clarification, Considerations, and Recommendations. Technical report.

Xu, H. and S. Gupta (2009, jun). The effects of privacy concerns and personal innovativeness on potential and experienced customers' adoption of location-based services. *Electronic Markets 19*(2-3), 137–149.

Zandbergen, P. A. (2009). Accuracy of iPhone locations: A comparison of assisted GPS, WiFi and cellular positioning. In *Transactions in GIS*, Volume 13, pp. 5–25.

Zhong, G., I. Goldberg, and U. Hengartner (2007). Louis, Lester and Pierre: Three Protocols for Location Privacy. *Pets'07*, 62–76.

Zurbaran, M., K. Avila, P. Wightman, and M. Fernandez (2014, nov). Near-Rand: Noise-based location obfuscation based on random neighboring points. pp. 1–6.

Zurbaran, M. and P. Wightman (2017, oct). VoKA: Voronoi K-aggregation mechanism for privacy in location-based information systems. In *2017 International Carnahan Conference on Security Technology (ICCST)*, pp. 1–6. IEEE.

Zurbarán, M., P. Wightman, C. Paolo, S. V. Mather, T. J. Kraft, and B. Park (2018). *PostGIS Cookbook* (Second ed.). PACKT PUBLISHING LIMITED.

Zurbarán, M. U. d. N., L. U. d. N. Gonzalez, P. U. d. N. Wightman, and M. U. o. S. F. Labrador (2014, jul). A Survey on Privacy in Location-Based Services. *Ingeniería y Desarrollo 32*(2), 314–343.