Adaptive Voronoi Masking

A method to protect confidential discrete spatial data

Master Thesis

August 10, 2020

Fiona S. Polzin
Student №: 6393225
F.S.Polzin@students.uu.nl
Supervisor: Ourania Kounadi
Responsible Professor: Raul Zurita-Milla

Acknowledgements

This is my thesis *"Adaptive Voronoi Masking - A method to protect confidential discrete spatial data"* which I conducted as part of my Master's degree *"Geographical Information Management and Application"* at Utrecht University within the last eight months. This work would not have been as successful without the help from a number of individuals:

First, I would like to thank my supervisor "Rania" Kounadi for giving me the best support and guidance throughout this research. Thank you for all your feedback and for helping me out when I struggled. I very much enjoyed the interesting discussions that clearly helped in completing this thesis. As well, I would like to thank my mom and my brothers for always encouraging and believing in me. I also want to thank my friends for always making me laugh and forget the studies for a brief moment. Finally, I would like to thank my boyfriend for making me smile, understanding this lengthy period, and keeping me company (even with the best food).

Abstract

Geomasks have been developed to assure the protection of individuals in a discrete spatial point data set by transferring the data points to a new location. Several obfuscating techniques exist but the risk of false re-identification is a commonly discussed problem. This thesis develops an alternative approach, referred to as Adaptive Voronoi Masking (AVM), which is based on the concepts of Adaptive Aerial Elimination (AAE) and Voronoi Masking (VM). It considers the underlying population density by establishing areas of K-anonymity in which Voronoi polygons are created. Complementary to many other geomasks, the proposed method considers the underlying topography and displaces data points to street intersections decreasing the risk of false-identification immensely since residences are not endowed with a data point.

The spatial characteristics of the new method are examined by the mean and median centers, the Nearest Neighbor Hierarchical Cluster Analysis (NNHCA), Ripley's K-function, a visualization of the masked data points, and are subsequently compared with the output of AAE, VM, and Donut Masking (DM). VM attains the best efficiency for the mean centers whereas DM does for the median centers. Regarding the NNHCA, DM demonstrates the strongest performance since it's cluster ellipsoids are the most similar to those of the ODP in terms of orientation, size, number of clusters, and mean points. In regard to the Ripley's K-function, AVM and DM succeed the other geomasks since they achieve the most similar values as the ODP retaining the point pattern of the original data sets. However, AVM clearly outperforms all methods regarding the risk of false re-identification as observed from the visualization of the MDP because no data point is moved to a residence which is a major step forward in academic research in geomasking techniques. Furthermore, AVM has the ability to maintain the spatial K-anonymity which is also done by AAE and partly by DM. Hence, based on these three factors, AVM is the best obfuscating method.

Additionally, this research analyzes whether any rules and regulations exist to protect individual data. Hence, the European GDPR was investigated to examine whether personal - particularly locational and health data - are protected. Thereby, it is concluded that the GDPR is vaguely worded and not efficient enough to protect the individual in terms of data processing. Besides, most geomasks do not comply with the rules and regulations of the GDPR since their masked data points can be traced back to their original location disclosing an individual by an address. Unlike this, AVM complies with the rules and regulations since no data point is moved to a residence that might be associated with an address and, therefore, with an individual.

Keywords

Geomasking — Adaptive Voronoi Masking — Voronoi Masking — Adaptive Aerial Elimination — GDPR

List of Abbrevations

AAE	Adaptive Areal Elimination
APA	Adaptive Point Aggregation
ARP	Adaptive Random Perturbation
AVM	Adaptive Voronoi Masking
NNHCA	Nearest Neighbor Hierarchical Cluster Analysis
DM	Donut Masking
DV	Disclosure Value
ESDA	Exploratory Spatial Data Analysis
expK-value	Expected K-value
EU	European Union
GDPR	General Data Protection Regulation
GIS	Geographic Information System
GPS	Global Positioning System
HiConfEnv	Upper confidence envelope
ITDB	Incident and Trafficking Database
Kact	actual K-anonymity
Kest	estimated K-anonymity
k-NN	K-nearest neighbor
LBS	Location-based services
LBSN	Location-based social networks
LwConfEnv	Lower confidence envelope
MDP	Masked Data Points
obsK-value	Observed K-value
ODP	Original Data Points
OSM	Open Street Map
PSA	Point Similarity Analysis
RFID	Radio Frequency Identification
RoRi	Risk of re-identification
RP	Random Perturbation
SKA	Spatial k-Anonymity
VM	Voronoi Masking

Contents

List of Figures

List of Tables

1 Introduction

1.1 Research Context

In the last years, the amount of high-resolution spatial data has been enormously growing (Fronterrè, 2018). This is due to the enclosing of positioning capabilities in mobile devices (i.e., global positioning system (GPS)) (Ghinita et al., 2010) and, hence, the identification of individuals' locations (Blumberg and Eckersley, 2009). Furthermore, the development of geospatial technologies (Bridwell, 2007), and the great interest for data (El Emam, 2013) have resulted into exciting applications and, thus, have been taking essential positions in our daily lives (Blumberg and Eckersley, 2009; Ghinita et al., 2010): these advancements promise us unprecedented services ranging from route navigation, finding friends (Bridwell, 2007), providing localized news (Furini and Tamanini, 2015) or even alert about health-related circumstances such as air pollution (Kounadi and Resch, 2018). This is possible due to the conjunction of advanced Geographic Information Systems (GIS), statistical methods (Zandbergen, 2007), and algorithms (Bridwell, 2007), which have established remarkable capabilities to analyze regional geographic variations in the fields of epidemiology (Zandbergen, 2007), health (Zimmerman and Pavlik, 2008), disaster and crisis management (Roche et al., 2013), crime (Kounadi and Leitner, 2014), and navigation (Bridwell, 2007).

Especially in the field of public health, the advances of GIS and the interest in spatial analysis have led to an increase of thematic maps in research (Zandbergen, 2014; Brownstein et al., 2006) and online platforms visualizing point data (Krieger, 2003; Zimmerman et al., 2007; A. J. Curtis et al., 2006). However, several studies in health geography (Brownstein et al., 2006), reproductive and sexual health (Haley et al., 2016) did not anonymize or aggregate data; instead, the original data were used (Brownstein et al., 2006; Haley et al., 2016; Kounadi and Leitner, 2014). Certainly, there is a persuasive justification to precisely observe data due to the straightforward identification of "hot spots", spatial patterns of a sickness (Zandbergen, 2014; Hampton et al., 2010; Zhang et al., 2017) and to evaluate the relationships between diseases and environmental exposures (Armstrong et al., 1999). Besides, accurate information can be communicated with the public, a disease surveillance system can be implemented (Armstrong et al., 1999; Olson et al., 2006), prevention and intervention can be supported (Zandbergen, 2014; Hampton et al., 2010), social inequalities within communities can be concluded (Hardwick and Patychuk, 1999), and health care systems can be improved (Cavoukian in El Emam, 2013).

Another problem with publishing or sharing data is the lacking awareness of the public and practitioners regarding the revelation risk of the data (Kounadi and Resch, 2018). Users are not always informed about what kind of personal data is compiled, utilized or by whom and miss the opportunity to disclose their personal and spatial information (Alrayes and Abdelmoty, 2014; Ricker et al., 2015; Thatcher, 2013).

However, publishing an individual's location either in paper or digital form - knowingly or unknowingly - increases the risk of re-identification by *reverse geocoding* and can seriously violate the individual's privacy. For instance, when publishing a data point, coordinates can be simply calculated and connected with an address as illustrated in figure 1.

This address can be easily associated with an individual by searching for a property database or directories (Zandbergen, 2014). Besides, by using Google Street View, a house and its front yard can be represented as well. How simple the re-identification of individuals is, was already demonstrated by Brownstein et al. (2006): By applying the reverse-identification method, 79% of the spatially coded addresses were accurately identified while all 550 of the plotted address points were disclosed within 14 m of the right address. Furthermore, Kounadi and Leitner (2014) exposed that within an eight-year duration, almost 70.000 home addresses had been disclosed in academic research. Adding to this, the authors emphasized that this is still an ongoing problem. However, Ajayakumar et al. (2019) criticized that geomasks are still unavailable for many institutions due to the lack of expertise in geospatial proficiency although the awareness of the power of mapping has grown particularly in health organizations and clinics which have become spatially literate lately. The authors stress that geomasks need to become more of a real-world requirement. Another example of the simplicity of re-identification can be represented by the leading software developer ESRI, which has created the tool "Reverse Geocode". This tool creates addresses from point data (ESRI, 2020e). The free and open-source software QGIS also developed a tool alike called "GeoCoding". It can be installed for free and allows the user to obtain addresses from point data as well (Pasotti, 2020; Tor, 2020).

Figure 1: The disclosure by identifying confidential data using coordinates.

The consequences are vast; an individual being identified as an HIV-patient - correctly or wrongly - can affect him or her by discrimination or social stigmatization (Seidl et al., 2018). Identifications can also cause harassment (Duckham and Kulik, 2006), unwanted advertisement or humiliation (Schwab et al., 2011; Schilit et al., 2003).

Kounadi and Leitner (2016) criticize that general rules on privacy do not include details of the spatial re-identification risk notwithstanding the fact that research and reports about guidance on the challenge of publishing or sharing spatial data exist (Graham, 2012; Wartell and McEwen, 2001; Kounadi and Resch, 2018). Consequently, confidential spatial data sets do not only have to be preserved but also need to comply with present-day restrictions and regulations on the right to privacy (Kounadi and Leitner, 2016). Accordingly, the importance of protecting confidential discrete spatial data particularly in regard of health data to assure an individual's confidentiality is evident: Health data is the most intimate and sensitive information of an individual's life - it can contain information about the mental state, physical fitness as well as the health history of someone's ancestors (Cavoukian in El Emam, 2013). As a result, the requirement for the individual's protection has grown resulting in the development of various geomasking methods to decrease the risk of re-identification.

1.2 Problem Statement

The growing need for protecting discrete geodata results from technological proliferation and the penchant of releasing data online along with academic publications where confidential discrete locational data is issued as thematic point maps (Kounadi and Leitner, 2014). Hence, researchers have been developing different methods to preserve the individual's privacy, such as geomasking techniques (Zandbergen, 2014). Geomasking techniques are meant to secure confidentiality when publishing geodata while preserving geographic detail to enable a precise spatial analysis of the data set (Seidl et al., 2018; Zimmerman and Pavlik, 2008; Zhang et al., 2017; Bridwell, 2007;

Hampton et al., 2010 Kounadi and Leitner, 2016). Simultaneously, it is tried to diminish the peril of re-identification (Seidl et al., 2018; Zandbergen, 2014). For further comprehension, the address attribute within spatial data is considered: The address attribute is supposed to be processed as confidential information because it can reveal an individual and, thereby, ought to be deleted before publishing a data set. However, the deletion of the addresses from a published data set results in a less accurate data utility from an academic point of view. Thus, geomasks aim to obfuscate the real locations of an individual. Through this, the address attribute does not have to be eliminated but can be transferred to a different position to maintain the issued data set for valuable research (Gupta and Rao, 2020).

Some geomasks displace the data points a specific distance aside from its original location (Seidl et al., 2018; Zandbergen, 2014; Bridwell, 2007; Zimmerman and Pavlik, 2008) (e.g., *random perturbation* (RP) (Armstrong et al., 1999), *flipping methodology* (Leitner and Curtis, 2004), and *Voronoi masking* (VM) (Seidl et al., 2015)), while others aggregate data points (e.g., *spatial* and *point aggregation* (Armstrong et al., 1999)). A more detailed clarification of the former mechanisms is given in the chapters 2.3, 2.3.1, and 2.3.2.

Recently, some geomasks also consider the underlying population density adapting the displacement error (Hampton et al., 2010; Wieland et al., 2008; Gruteser and Grunwald, 2003; Cassa et al., 2006).

By considering the underlying population density, the user of the mask is able to determine a level of *K-anonymity* (Cassa et al., 2006; Wieland et al., 2008; Hampton et al., 2010; Gruteser and Grunwald, 2003 in Kounadi and Leitner, 2016). By providing K-anonymity, each record or person within a masked data set will not be identified from at least k-1 record or person whose details lie within the data set as well (Sweeney, 2002). Regarding spatial data sets, K-anonymity assures that every location such as household, address or an individual's location cannot be differentiated from minimum k-1 locations (Kounadi and Leitner, 2016; Ghinita et al., 2010; Gruteser and Grunwald, 2003). This means, that *spatial k-anonymity* (SKA) is used to describe the probability of identifying a location that can be linked to an individual by reverse geocoding. It can be implemented when analyzing the degree of privacy as well as when measuring the width of displacement (Seidl et al., 2018). It is plausible that especially in the field of health and epidemiology, geomasks have demonstrated to be useful contributors since data support the investigation of crucial health and disease patterns, but also include sensitive information (Zhang et al., 2017; Bridwell, 2007).
Many masking methods exist and can be easily applied, but limitations and negative aftereffects exist (Council et al., 2007): As already mentioned in chapter 1.1, published locations can be re-engineered to spot the original locations (Brownstein et al., 2006; Zandbergen, 2014). Adding to that, identified locations can be linked with further information (Krumm, 2007) creating personal data (Kulk and Van Loenen, 2012). A possible solution could be the aggregation of point data. However, when aggregating point data, the ability to distinguish spatial relations or clusters and deriving persuasive information is decreased (Zandbergen, 2014; Armstrong et al., 1999; Kwan et al., 2004). Obviously, the data becomes less useful for research purposes (Zandbergen, 2014; Gupta and Rao, 2020). Contrary to masking methods applying aggregation, masking methods that modify the locations of data points can be applied (Armstrong et al., 1999). Nevertheless, the transferred points can be moved to a position which has real observations (Zimmerman et al., 2007) or where they cannot exist (Council et al., 2007), resulting in false identification (Seidl et al., 2015; Council et al., 2007):

False identification represents the incorrect linking of a household or person to a data point. Contrary to that, correct identification is the correct linkage of a household or person to a data point (Seidl et al., 2015). The consequences resulting from identification were already emphasized in the previous chapter 1.1 and can result in many negative effects impeding an individual's social prominence (Gupta and Rao, 2020). Besides, it can unintentionally involve individuals, who were not part of the research (Council et al., 2007). It is evident, that these limitations influence both the disclosure risk (Council et al., 2007) and a successful investigation of spatial patterns (Kwan et al., 2004).
There is no generally recommended nor approved geomask (Zandbergen, 2014; Gupta and Rao, 2020) and each method has disadvantages and advantages. Moreover, no geomasking method currently applicable is able to serve a thorough solution to protect locational privacy (Duckham and Kulik, 2006). For this reason, Zandbergen (2014) suggests counterbalancing data utility and confidentiality protection. Furthermore, no research known to the author has combined two methods to create an anonymization algorithm which is taking topological polygon relationships into account

(Seidl et al., 2018). Also, not a lot of geomasks consider the underlying topography except for the *Street aggregation at intersection or at midpoint* (Leitner and Curtis, 2004) or the *Location Swapping*-method (Zhang et al., 2017) (see chapter 2.3). Finally, yet importantly, it is uncertain whether laws and policies protect the individual's privacy due to the increasing tendency of publishing data online (El Emam, 2013; Kulk and Van Loenen, 2012). It is of great significance to mention, that many users are not fully aware that location information is being amassed, how the data is exploited, and by whom. Therefore, many users are not capable of disclosing their spatial information (Alrayes and Abdelmoty, 2014). In May 2018, the General Data Protection Regulation (GDPR) came into effect, empowering users by implementing constraints and transparency for the processing and storage of personal data (Kounadi et al., 2018). Hence, it is of valuable research, whether current or a new geomasking algorithm will comply with those.

1.3 Objective and Research Questions

1.3.1 Scope

The main objective of this research is to combine two geomasking techniques VM and AAE, which will be called *Adaptive Voronoi Masking* (AVM). Through that, it is aimed to protect the individual's privacy while at the same time decreasing the false identification risk.

1.3.2 Scientific Objectives and Research Questions

The study objective will evaluate commonly applied geomasking techniques and compare these with the proposed method AVM. As a result, this research can examine whether AVM is able to reduce the false re-identification risk and provide confidentiality while maintaining data utility. Additionally, we will analyze whether AVM complies with the rules and regulations by the GDPR. Therefore, the following subquestions will be answered to investigate these objectives:

1. Study Objective 1:

   (1) *How do commonly used geomasking methods affect the false re-identification risk?* As elaborated in subchapter 1.2, false re-identification is a common problem among geomasking techniques. To attain a full understanding and to develop an algorithm which intends to abate this issue, commonly used geomasking methods and their risk of false re-identification will be examined.

   (2) *To what degree can the proposed geomasking technique reduce the false re-identification risk?* Based on subqestion one, a visualization of the masked point data will be employed to evaluate the risk of false re-identification of AVM based on SKA and to compare the result with other commonly used geomasking techniques.

   (3) *To what degree can the proposed geomasking technique provide confidentiality for the individuals while maintaining data utility?* Since a growing pressure to issue raw health data for commercial purposes, research, and policy can be observed and detailed data sets are of critical importance for successful population-based research (El Emam, 2013), it must be scrutinized whether AVM will retain data utility while at the same time support confidentiality. Only through that, correct spatial analysis can be conducted. Besides, it is a well-known dispute that geomasking techniques attempt to achieve a balance between preserving data utility and the risk of re-identification (Zandbergen, 2014).

2. Study Objective 2:

   (4) *How does the proposed method comply with the rules and regulations by the General Data Protection Regulation (GDPR)?* Due to the increasing tendency of publishing data online, it is ambiguous whether laws and policies protect the individual's privacy (El Emam, 2013; Kulk and Van Loenen, 2012).

   (5) *To what level is the GDPR protecting the individual's privacy with respect to information derived from spatial data?* As described in subchapter 1.2, many users are not informed about their data being assembled, by whom and how the data is being applied. Besides, it is a widely spread problem of sharing confidential data. Therefore, it is of great significance to evaluate whether the GDPR is preserving the individual's privacy in the case of individuals not protecting their data by themselves.

1.3.3 Research Delimitations

This research intends to merge the VM and AAE-method creating a new algorithm to protect the individual's privacy while decreasing the false identification risk. Nevertheless, the time constraints of this research limit the comparison and combination of other methods. Therefore, the focus will be on commonly applied geomasking techniques and *not* on other anonymization methods. Furthermore, it will not be investigated whether a lower possibility of false identification risk can be achieved by either merging VM or AAE with a different geomask (i.e., donut masking, grid masking, etc.) or by combining two completely different methods. Besides, for evaluating the data utility and the risk of re-identification various statistical methods exist, however, only a few statistical approaches will be implemented. Thus, this research will not evaluate which statistical method is the best technique to evaluate spatial information. Also, this research does not aim to endorse the *best* or *the* universally accepted geomasking technique.

Moreover, since this research focuses on rules and regulations, only the European Union (EU) and its GDPR will be taken into consideration. Countries and their legal framework outside the EU will not be considered to not exceed the frame of this thesis.

1.4 Reading Guide

The structure of this thesis is illustrated in figure 2 and is organized as follows: The first chapter aims to give a preface of geomasks, their relevance, and the problem of re-identifying individuals by using location as an identifier. The research objectives and questions are introduced.

Chapter 2 describes the theoretical framework. Hereby, geoprivacy is elaborated due to its vital role in the protection of confidential discrete spatial data followed by an outline of the GDPR and its aim and limitations regarding locational data and data processing. Thereafter, various geomasking techniques are explained, particularly the two to be merged methods VM and AAE. Here, it is the goal to represent the various obfuscating methods existing. The last subchapter introduces the exploratory spatial data analysis methods to be used to compare the original data points (ODP) with the outcome of the geomasks.

Chapter 3 introduces the study area followed by the software and data to be applied. Next, the new obfuscating mechanism AVM is elaborated followed by the implementation of the statistical methods. Chapter 4 discusses the results of the statistical methods implemented on the masked data set as well as on the ODP and compares them. Chapter 5 answers the research questions whereas chapter 6 summarizes the findings of this thesis including recommendations for future research. Further, the limitations of this study as well as future recommendations are highlighted in chapter 6.1.

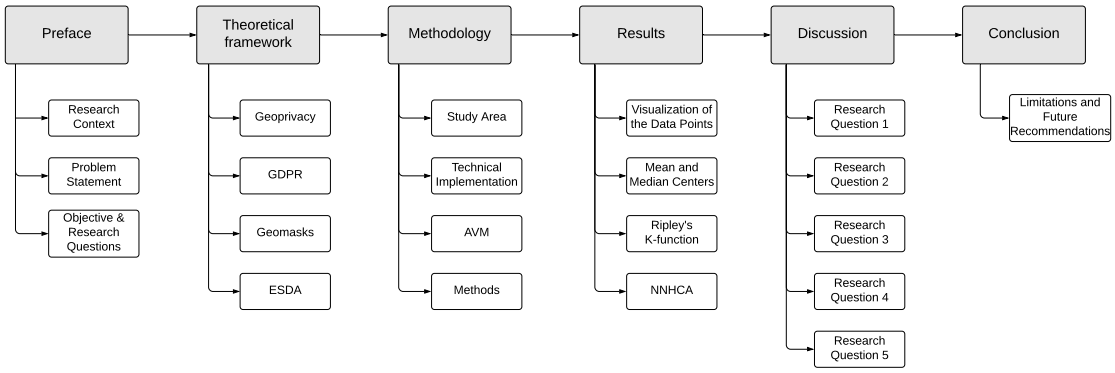


Figure 2: The work flow of this research.

2 Theoretical Background

2.1 Geoprivacy

Due to the extensive assemblage of locational data through geolocation technologies, our positions can be identified with GPS, radio frequency identification (RFID), IP address location, cellphone network triangulation (Furini and Tamanini, 2015), and high-resolution aerial images (Council et al., 2007). The dissemination of locational data ensues either through cellphones or surveys, reports, and publications which can contain confidential spatial data as well as through location-based social networks (LBSN) (i.e., Twitter, Foursquare, and Facebook) (Kounadi, 2015) and location-based services (LBS) (Bridwell, 2007; Kounadi, 2015), which are services using the users' geographic position to provide them with spatially adaptive content. Through the collection of data, planners and scientists are provided with data about spatio-temporal patterns of mobility, activity, and social interaction (Bridwell, 2007). As a result of the various types collecting spatial data, a subjective categorization of spatial data at risk (Kounadi and Resch, 2018) is listed in table 1.

Nº	Spatial data at risk	Characteristics
1	Mobile phone data	Includes the user's time stamp as well as location
2	Sensitive discrete spatial data about individuals	Normally crime and health geocoded data sets including residential positions of victims or patients
3	Confidential discrete spatial data about individuals	
4	LBS data	Navigation services which gather user's temporal and spatial information
5	LBSN data	Referring to social media applications, the user may reveal her or his position with attribute information and time stamp
6	Confidential discrete spatial data	The least debated spatial data set in literature ⇒ Incident and Trafficking Database (ITDB) evolved by the International Atomic Energy Agency comprising illegitimate relocation of radioactive and nuclear materials (International Atomic Energy Agency, 2015)
7	Data from mobile technical sensors transported by users (= sensor operators)	Spatiotemporal data which is collected from participatory mobile sensing applications ⇒ refers to mobile phone users who gather data about their surroundings, i.e., noise, air quality, and traffic
8	Data from mobile devices transported by users (= subjective sensors)	Spatiotemporal data which is collected from participatory mobile sensing applications ⇒ refers to mobile phone users who gather data about their own subjective impression on the sensed attribute (either themselves (emotions) or environment (i.e. road or public safety)
9	Data from mobile technical sensors transported by users (= objective operators)	Spatiotemporal data which is collected from participatory mobile sensing applications

Table 1: Spatial data at risk and their characteristics (Kounadi and Resch, 2018).

It is noteworthy that the various spatial data sets in the previous table distinguish between their protection characteristics: For example, data set three illustrates confidential locations of *individuals* while the sixth data set is about confidential discrete spatial data. LBSN data can direct to the unveiling of personal convictions and interests. The mobile phone and LBS data sets are alike regarding their attributes but differ in their temporal frequency. Moreover, the last two data sets are far more complicated to protect because they demonstrate high sensitivity and diversity of individual information resulting in a considerable privacy loss in relation to the other data sets. Therefore, the procedure to secure these data sets regarding anonymity level, anonymity

measure, protection method, or data access can vary (Kounadi and Resch, 2018).

Examining the sources of dispersing locational data, it can be differentiated between two types of spatial disclosure (Kounadi, 2015):

1. The disclosure of a person's spatial patterns (unveiling of private spatial information);
2. The disclosure of a person's confidential information in spatial data sets or on maps (unveiling of confidential spatial information) (Kounadi, 2015).

The second kind resembles the type of disclosure which this research focuses on. The disclosure of both types has happened (Kounadi, 2015) due to technological proliferation and the gathering of locational data. Through that, our personal privacy has been undertaken changes since workplaces or even our home addresses can be identified and connected with other spatial information (Onsrud et al., 1994; Armstrong, 2002; Monmonier, 2004). Consequently, an increasing public awareness about the jeopardy of unveiling discrete spatial information (Beresford and Stajano, 2003; Bettini et al., 2005) and about its availability and usage have arisen (Kounadi et al., 2018) on a global scale (Keßler and McKenzie, 2018):

After the devastating hurricane Katrina, a local newspaper in the capital of Louisiana, Baton Rouge, published a map illustrating the locations of deaths caused by the natural disaster (A. J. Curtis et al., 2006). In Europe, in 2013 the privacy campaign group *Big Brother Watch* levied legal concerns over tracking devices in recycling bins in London, which tracked mobile devices of passers-by, after technological details implemented in the bins were published in the online magazine *Quartz* (Miller, 2013). In the state of Florida, at the beginning of March 2020, a man was accused of robbing an elderly woman since his mobile data was investigated by the police. At the time of the burglary, the accused man was using an exercise-app to track his work-out, which was linked to his Google account. The Florida police department applied their "geofence warrant", which resembles a systematic search merged with crime scenes to collect Google location data derived from the user's Bluetooth, Wi-Fi, GPS, and cellular connections. Afterward, the department had returned to Google to request more information on the suspect (Schuppe, 2020).

Due to the current COVID-19 pandemic, the state of Israel has approved its internal security service *Shin Bet* to start the surveillance of infected citizens and the spreading of the virus by analyzing mobile phone data, which, firstly, occurred without court order authorization (Bob and Hoffman, 2020, March 17; Föderl-Schmidt, A., 2020, March 21). Within a day, 400 citizens were sent to quarantine, resulting in demonstrations about the threat to democracy (Föderl-Schmidt, A., 2020, March 21). Simultaneously, the German telecommunications company *Telekom* provided the spatial data of 46 million mobile phone users to the *Robert-Koch-Institute* to decrease the spread of the virus (Balser and Hurtz, 2020, March 18).

These cases not only support the widely accepted axiom that technology is progressing faster than the law or policy regarding locational privacy, spawning conflicts all over the world (Keßler and McKenzie, 2018). The high media coverage of allegations of locational privacy violations is also emblematic for a growing public recognition of location privacy concerns (Duckham and Kulik, 2006) and has resulted in unease and open discussion about policies and implications needed for the public use of discrete data (Kounadi et al., 2015). Bridwell (2007) stated over a decade ago, that the existing regulations can be rather defined as *"disclosure limitation strategies"*. This statement was supported by Ricker et al. (2015), who describe the app development and the re-purpose of collected data as the *"wild west"*, where laws are written *after* security breaches had been violated. Boulos et al. (2009) described data security as the *"missing ring"* and describe a phenomenon which is compromising on neglecting data loss, data theft or data disclosure by non-authorized groups (Kounadi et al., 2018).

Due to the former examples of identifying individuals by their locations by applying LBS-services and researchers deprecating the deficiency of fixed rules and regulations, it is disputable whether rules and regulations exist to secure spatial data. For this, locational privacy or geoprivacy has to be defined. Kwan et al. (2004) described geoprivacy as:

> *"individual's right to prevent disclosure of the location of one's home, workplace, daily activities, or trips. The purpose of protection geo-privacy is to prevent individuals from being identified through locational information"* (p.3).

This definition can also be supported by AbdelMalik et al. (2008), Duckham and Kulik (2007), and Bridwell (2007).

However, the definition by Kwan et al. (2004) refers to geoprivacy for the individual's spatiotemporal activity and confidential discrete locational data (Kounadi et al., 2018). A definition for the privacy in participatory sensing, i.e., as seen on Facebook or Foursquare, is required, too. Here, Christin et al. (2011) defined description for individuals using participatory sensing applications:

> *"Privacy in participatory sensing is the guarantee that participants maintain control over the release of their sensitive information. This includes the protection of information that can be inferred from both the sensor readings themselves as well as from the interaction of the users with the participatory sensing system".*

It has to be mentioned, that this definition refers to health monitoring, e-diaries and other applications (Kounadi et al., 2018).

As can be seen, by the two definitions and the examples of disclosing discrete information, individuals surely lost control over their spatial data since their daily activities, trips, and residences were identified and, resulting, their identities. Ironically, former research has shown that the majority of people possessing smartphones are aware of these privacy concerns, "but did not change behaviours based on their concerns" (Kar et al., 2013 in Ricker et al., 2015, p. 4). Kramer et al. (2014) observed that social media companies (e.g., Facebook) have started to obscure the boundaries of what is socially responsible and ethical. Similar observations regarding social media were made by Furini and Tamanini (2015), who also discovered that the majority of people active on social media, did not know about the "nature of their profile" nor privacy settings. It seems that the proliferation of smartphones dominates the disadvantages (Kar et al., 2013).

Nonetheless, privacy settings do exist. Yet studies have revealed that the majority of users most likely do not spend more than one to five minutes on reading privacy policies (Van den Berg and Van der Hof, 2012) because it is a very time-consuming procedure (McDonald and Cranor, 2008). Some might even consent to the privacy policies without fully reading or grasping them (Toubiana and Nissenbaum, 2011). Albeit some users try to read them, many are not able to thoroughly understand them due to their technological and legalistic tests. In the case of succinct and straightforward privacy policies, it is doubtful how much information they really provide (Toubiana and Nissenbaum, 2011).

It is evident that geoprivacy is underlying great complexity: governmental institutions as well as companies are interested in locational data while users seek the greatest proliferation for their daily activities and interests on the one hand but feel threatened in their personal privacy on the other hand. Figure 3 gives a detailed overview of the complexity of geoprivacy: The *user* is in constant interaction with its *tools* that provides him or her with *applications*. These applications run on their *utility*, which the user is asking for, while the utility relies on *precision*, which the user has for his or her *location*. The location information is provided via, i.e., mobile phone data, LBS, LBSN, or mobile technical sensors underlying different protection methods as specified at the beginning of this chapter. The cycle between utility, tools, precision, and location information is influenced by external factors: First of all, the user's *education and information* will decide, whether a user is considering a tool as secure to use.

Secondly, the tools are dependent on the *technological developments*, which are reliant on the *service provider*. These are influenced by the *economic interest*, which is affected by *legal framework* that, again, is influenced by *ethical aspects* (Keßler and McKenzie, 2018). It must be mentioned that due to distinctions in legal frameworks and cultures, the legal implementation to protect data distinguishes (Custers et al., 2018) among countries (Lynskey, 2015 Custers et al., 2018; Cole and Fabbrini, 2016; Botha et al., 2017). For instance, while the right to data protection is fundamentally grounded within the EU (Lynskey, 2015), the US manages data security as well as privacy on a sectoral basis (Kaminski, 2015). Therefore, it can be assumed that not only legal frameworks but also the perception of breaching geoprivacy distinguish from country to country highlighting the complexity of geoprivacy and leading to the uncertainty of how the individual's confidential spatial data is being preserved by a legal framework.

Adding to that, the service provider might share the gathered data with *third party businesses and services*, for either an advantage the third party business is paying for, or to save the data on a platform. In either way, there is a possibility of this happening with or without the user's permission (Keßler and McKenzie, 2018). The third-party businesses and services have also an economic interest: The value of re-using only public sector information in Europe was expected to reach up to €140 billion (Dekkers et al., 2006), emphasizing the great commercial potential (Kulk and Van Loenen, 2012).

Nonetheless, third-party businesses and services operating in the EU are affected by the *GDPR* which aims to protect users regarding the collection and further processing (Custers et al., 2018; Kounadi and Resch, 2018) as well as re-selling (Keßler and McKenzie, 2018) of their personal data. How the GDPR aims to secure personal data especially regarding spatial data, and whether commonly applied geomasks comply with these rules and regulations is exemplified in the following subchapter.

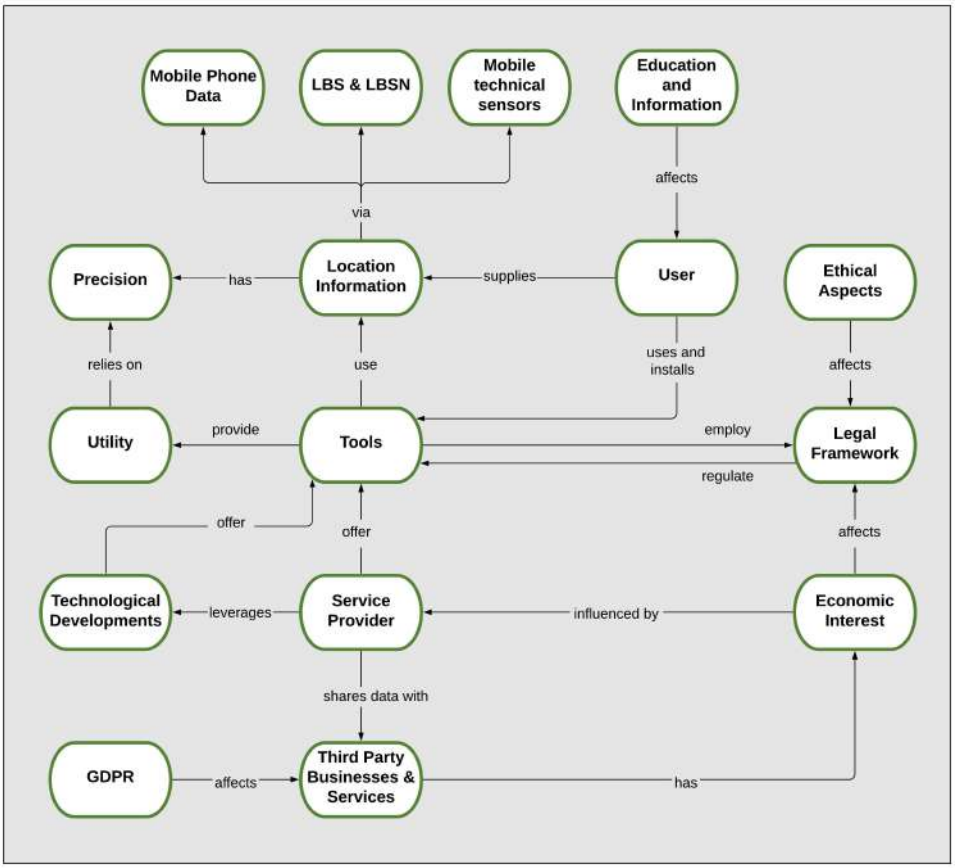


Figure 3: The complexity of geoprivacy in the EU (adapted from Keßler and McKenzie (2018)).

2.2 General Data Protection Regulation (GDPR)

The main objective of the GDPR is to unify data privacy laws across the EU and has been in effect since May 25th, 2018 (European Parliament, 2016). The GDPR particularly addresses the improvement of the legal framework of users regarding the constraints and transparency of processing and collecting personal data (Custers et al., 2018; Kounadi et al., 2018). Hereby, organizations, governments (Custers et al., 2018), and LBSN engaging within the EU or the European economic area have to develop an approach of gathering data that excludes data misuse (Kounadi et al., 2018). The GDPR obligates to inform individuals about the processing of their data, to reduce the time of data storage, and to limit data usage (European Parliament, 2016). It is noteworthy that the implementations of the rules and obligations set by the GDPR distinct between EU member states due to cultural diversity and different legal systems (Bamberger and Mulligan, 2015). The investigation of the GDPR was chosen for this research because, firstly, this research's study area takes place in Germany and, secondly, because no research known to the author has examined whether geomasking techniques comply with the GDPR. This scrutiny is also incumbent to answer the research questions five and partly four (see chapter 1.3.2).

Therefore - also to not exceed the frame of this thesis -, the GDPR will be analyzed regarding valuable factors for academic research in geomasking techniques and their relation to the re-identification of individuals, such as *personal data*, and *processing data*. Besides, the GDPR will be examined regarding *health data* since this research is predominantly focusing on this kind of data.

2.2.1 Personal data

Within the Charter of Fundamental Rights of the European Union, the articles 8(1) and 16(1) state that every person *"has the right to the protection of personal data"* (European Parliament, 2016). However, *"personal data"* is a broad term as can be observed by the definition by the GDPR:

> *"personal data means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person"* (art. 4, paragraph 1).

Evidently, *"locational data"* belongs to personal data since it can be used as an identifier - directly or indirectly. This can be explained by the example of observing an individual's spatial data for a specific period of time. Through that, it is most likely to discover a user's identity as illustrated in chapter 2.1.

The definition of personal data in art. 4, paragraph 1, also described *physical, physiological, genetic, [and] mental* factors as personal data. These terms draw a strong link to health data, which this research is predominantly focusing on. Health data or - as referred to in the GDPR - *data concerning health* is defined as:

> *"personal data related to the physical or mental health of a natural person, including the provision of health care services, which reveal information about his or her health status;"* (art. 4, paragraph 15).

Personal health data, too, is underlying special safeguards. Therefore, processes that aim to encourage improved healthcare and innovation (i.e., mobile health), require strong data protection regulations to protect confidential data (European Data Protection Supervisor, 2020). Hence, the following subchapter will define the processing of data.

2.2.2 Processing Personal Data

Due to the aforementioned examples of identifying individuals by their locational data, it is questionable how personal data can be processed. The GDPR defines *processing* as:

> *"any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission,*

dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction" (art. 4, paragraph 2).

In a succinct summary, the GDPR describes principles of processing personal data concerning the following aspects:

- **Lawful processing**: Personal data has to be processed in a transparent and fair manner in case of no user consent (art. 5, paragraph 1(a));
- **Purpose limitation**: Personal data is collected for legitimate, specified, and explicit purposes. It is not supposed to be further processed in an incompatible way, except for statistical purposes, historical or scientific research purposes or for the public interest (art. 89, paragraph 1) (art. 5, paragraph 1(b))
- **Data minimisation**: Personal data is supposed to be limited regarding to the necessities of the processing purpose (art. 5, paragraph 1(c));
- **Accuracy**: Any inaccurate personal data are to be erased or rectified (art. 5, paragraph 1(d));
- **Storage limitation**: Data can be stored for the shortest time possible. The institution needs to specify the time limits to access or erase the stored data (art. 5, paragraph 1(e));
- **Integrity and confidentiality**: Personal data has to be processed in a specific manner that assures the protection against unlawful or unauthorized processing as well as against damage, destruction, and accidental data loss (art. 5, paragraph 1(f));
- **Right of access**: The data subject has the right to access his or her personal data (art. 13, paragraph 2b));
- **Records of Processing Activities**: Data processing steps have to be documented including data processing steps and their purposes, the time limits for erasure, general definition of security measures, group of data subjects, data categories, and data recipients (art. 32).
- **Right to erasure**: The data subject has the opportunity to erase all of his or her personal data (art. 13, paragraph 2(b));
- **Pseudonymization**: By applying pseudonymization to personal data, parts of information is replaced with random information, reducing the risk of re-identifying data subjects (art. 4, paragraph 5).
- **Data protection officer**: Every institution requires a data protection officer (art. 37);
- **Data breaches**: In case of any data breach, the data controller is compelled to inform the supervisory authority (art. 33, paragraph 1);
- **Fines and Penalties**: Audits, warnings, and fines can be assessed Violators of the GDPR may be fined up to €10 Mio. or "up to 2% of the total worldwide annual turnover of the preceding financial year" (art. 83, paragraph 4);

Regarding the aspect of lawful processing and that data has to be processed transparently, Custers et al. (2018) discovered that transparency practices on personal data processing is significantly low within the EU member states.

It is of great significance that the GDPR does not cover the following cases:

- Military, justice, police, lawful interception, and national security (paragraph 104; L 119/19);
- Deceased people (L 119/5);
- Scientific and statistical analysis;
- Data subjects to national legislation;
- In case of a dedicated law on employer-employee relationships; (Kounadi et al., 2018)

- "Processing of personal data by a natural person in the course of a purely personal or household activity" (Kounadi et al., 2018, p. 11).

As can be seen, scientific and statistical analyses are excluded from the GDPR, however, paragraph 33 states that "data subjects should be allowed to give their consent to certain areas of scientific research" (L119/6). There is no further interpretation of the type of scientific research, creating interpretability based on personal interest. Still, the statistical analyses are of great interest for companies, because they can analyze individual data and create inferences on the individuals developing benefits for their own company. As highlighted in chapter 2.1, the company might sell the data to third party businesses or saves the data on a platform. In that case, there is a chance that this is occurring without the user's permission (Keßler and McKenzie, 2018). Therefore, by simply stating "statistical purposes", the company is given a lot of freedom on *why it needs to store the individual's data* for an unlimited time frame.

The previous examples of disclosing health data in this research illustrated the common act of processing personal data. The processing of specific personal data - such as health data - is defined by art. 9, paragraph 1, which clearly states that processing personal data is not allowed in case of possibly revealing health data[2]. Nonetheless, exemptions exist, for example:

- the data subject gives permission (art. 9, paragraph 2(a))
- *"processing is necessary to protect the vital interests of the data subject or of another natural person"* (art. 9, paragraph 2(c)), which could be applied to track the outbreak of a disease as discussed in the current COVID-19 pandemic by the German Robert Koch Institute;
- *"processing is necessary for the purposes of preventive or occupational medicine, [...] medical diagnosis, the provision of health or social care or treatment or the management of health or social care systems and services on the basis of Union or Member State law or pursuant to contract with a health professional and subject"* (art. 9, paragraph 2(h)), which can again be applied to survey the outbreak of the present COVID-19 pandemic and "improve the quality of life for a number of people" (paragraph 57; L119/29).

In this case, the processing of personal (health) data has to be undertaken under the *"responsibility of a professional subject to the obligation of professional secrecy"* (art. 9, paragraph 3).

[2] Art. 9, paragraph 1: *"Processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation shall be prohibited."*

2.3 Geomasking Techniques

Researchers have faced confidentiality challenges by developing three main strategies for preserving spatial privacy: policy-based changes, anonymization, and geomasking or data obfuscation (Ardagna et al., 2008). Among all spatial privacy protection strategies, geomasking has been the most commonly applied and probed method (Ajayakumar et al., 2019). As depicted in figure 4, geomasking techniques and their masking degree can be generally categorized in various classes, which will be further elaborated in this chapter.

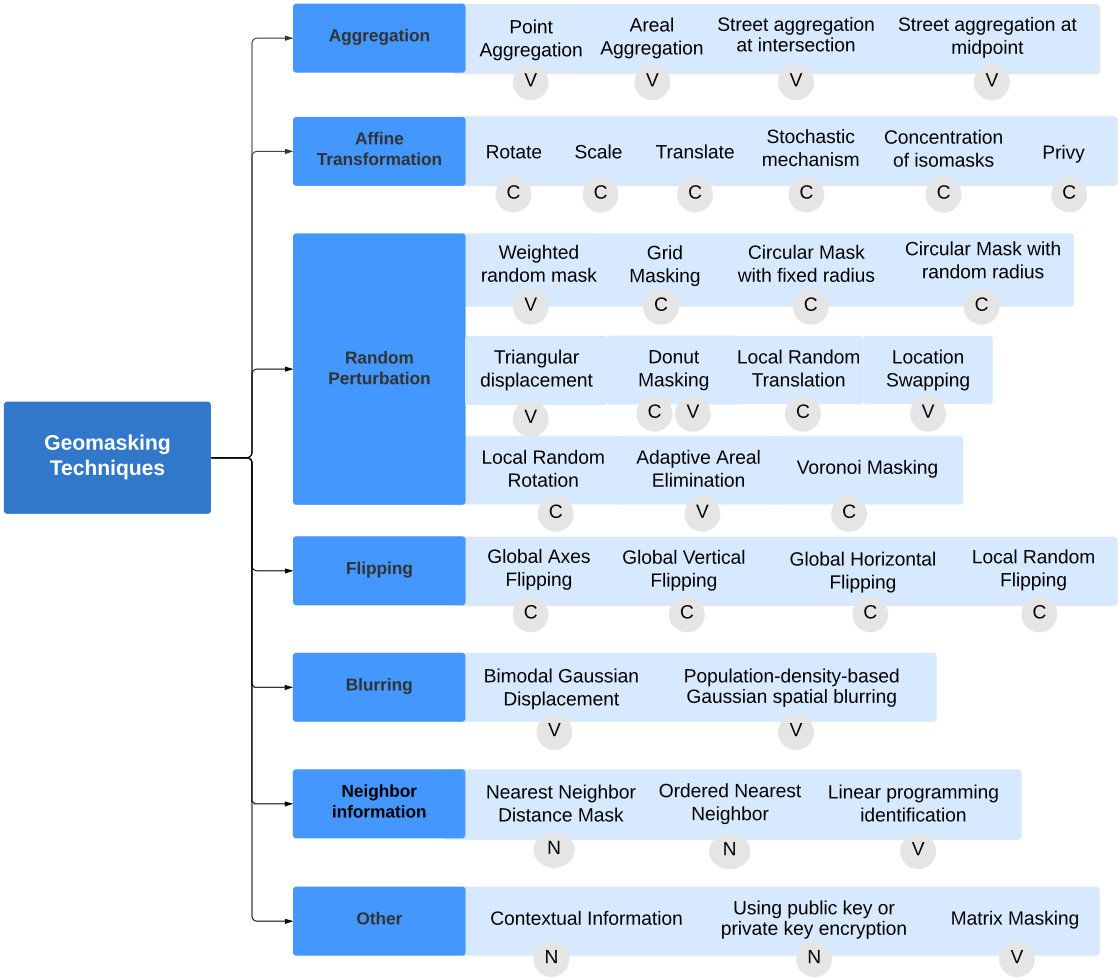
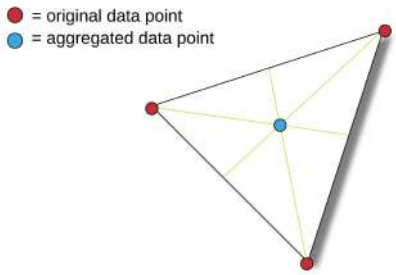
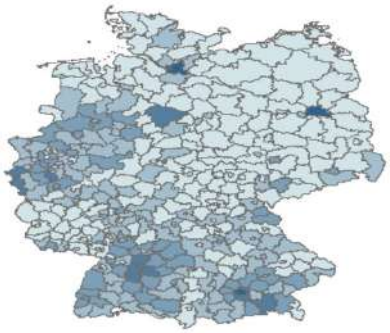


Figure 4: The different classes of geomasking methods. The grey circles represent the masking degree (C = constant; V = variable; N = not applicable) (adapted from Gupta and Rao (2020) and Kounadi and Leitner, 2015b (2015)).

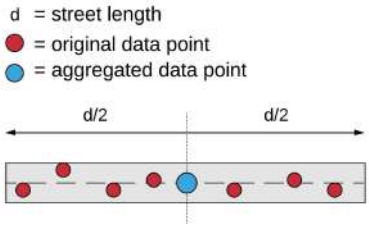
The first masking methods were based on attribute transforming, record transforming, or displacing to preserve the confidentiality of non-spatial and tabular databases (Duncan, Pearson, et al., 1991). These were further evolved by Armstrong et al. (1999), who took the individual's location into consideration (Armstrong et al., 1999). The methods created by Armstrong et al. (1999) can be differentiated between *spatial aggregation* and *point aggregation*. Generally said, **aggregation** represents the basic method of aggregating data points to specific units before releasing data. In point aggregation, geographically similar points are combined on the map and are replaced by one point illustrating an individual phenomenon, as illustrated in figure 5a. Contrary to point aggregation, spatial aggregation combines all points into one area (Armstrong et al., 1999). A well-known output of this is the so-called choropleth-maps (Zandbergen, 2014) (see figure 5b). Two other methods of aggregating data points, which are more flexible than the aforementioned techniques, are aggregating points either in the middle of a street segment or at the intersection of a street (Leitner and Curtis, 2004; see figures 5c and 5d).



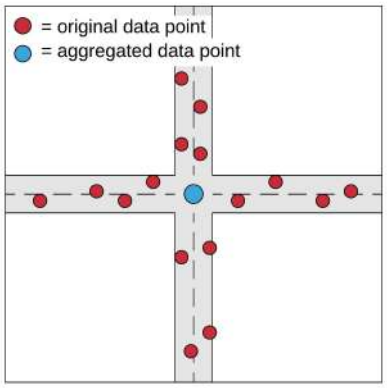
(a) Point aggregation: Original data points are replaced by one point (adapted from Gupta and Rao (2020)).



(b) An instance of spatial aggregation: COVID-19 cases in Germany on April 7th, 2020 (ESRI, 2020f; ODS, 2020).



(c) Data aggregation in the middle of a street segment (adapted from Gupta and Rao (2020)).



(d) Data aggregation at the intersection (adapted from Gupta and Rao (2020)).

Figure 5: Data aggregation. The red dots symbolize the original locations while the blue dots represent one of the many potential obfuscated locations.

As is evident, the decreased spatial resolution and the aggregated points demonstrate a great protection regarding spatial confidentiality (Ajayakumar et al., 2019; A. Curtis et al., 2011). However, various authors have criticized the techniques' influence on the ability to distinguish spatial relations or clusters (Armstrong et al., 1999; Zandbergen, 2014; Kwan et al., 2004; Hampton et al., 2010) and the loss of a similar point pattern to that of the ODP (Kwan et al., 2004). Especially in the field of epidemiology or public health, it is apparent that health data aggregated to an administrative unit or political area becomes inadequate to observe epidemiological and geographical trends (Boulos et al., 2006), as depicted in figure 5b. Furthermore, spatial units do not fulfill the same for all findings (i.e., the amount of addresses can vary) (Krieger et al., 2002). Moreover, the spatial units cannot correspond to significant spatial or social divisions (Oakes and Johnson, 2006). Thus, it is indisputable that geographical analyses have to examine relations in areas which are not only described in a more flexible way than by using administrative, political units (Armstrong et al., 1999) or streets, but also maintain a similar point pattern. Regarding point aggregation, error in spatial analyses can be introduced (i.e., incorrect cluster detection (Hampton et al., 2010). Finally, the masking degree of all aggregation techniques is variable (see figure 4).

The **affine transformations techniques** - also referred to as *isomasks* (Ajayakumar et al., 2019) - techniques apply basic geometric transformations to move spatial point data (Armstrong et al., 1999). Here, it can be differentiated between *rotate, scale* and *translate*, *stochastic mechanism*, and *concentration of isomasks*. Affine transformations change the original location by either twisting every point by a set angle (referred to as rotation), or moving every point by a set increment (referred to as translation) or by moving every point by scaling constant (referred to as

scale) (see figure 6) (Armstrong et al., 1999).

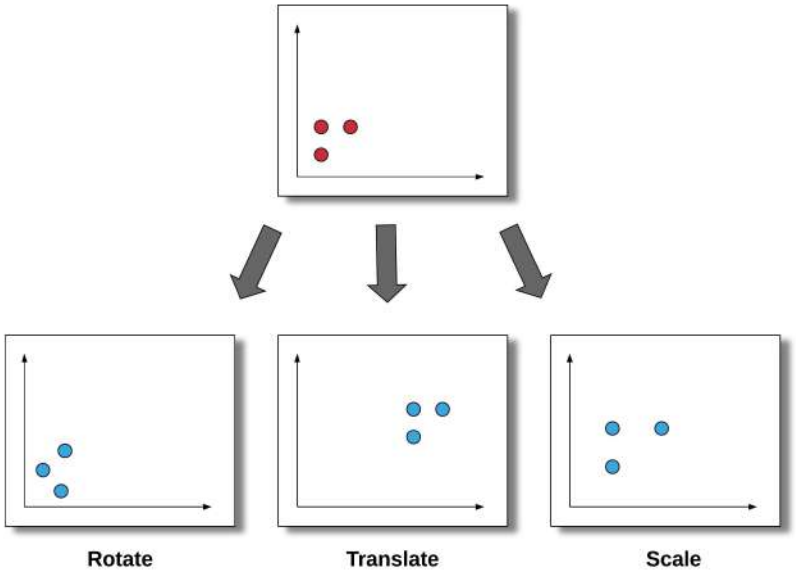


Figure 6: Affine transformation by rotating, translating, and scaling (adapted from Armstrong et al. (1999)).

Applying a stochastic mechanism is choosing a random scaling constant, increment or angle in a predefined range as depicted in figure 7 (Kounadi and Leitner, 2015b). The concentration of isomasks is the combination of rotation, scale, and translation illustrated in figure 8 (Kounadi and Leitner, 2015b).

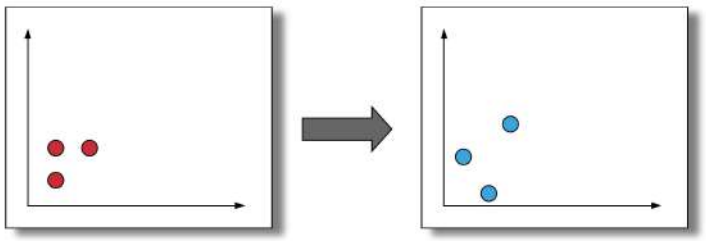


Figure 7: The stochastic mechanism.

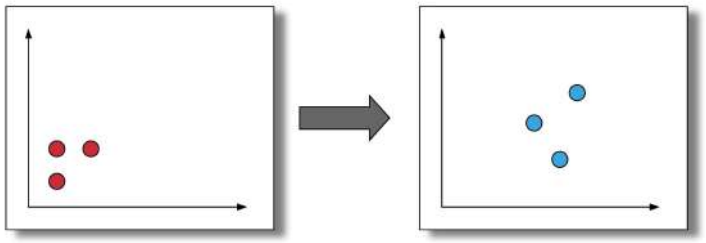


Figure 8: Concentration of isomasks.

The last method in this category - the *Privy* method - includes a random rotation and spatial translation of the ODP. The distance offset is obtained from a random number to assure a minimum distance from the original location (see figure 9) (Ajayakumar et al., 2019).

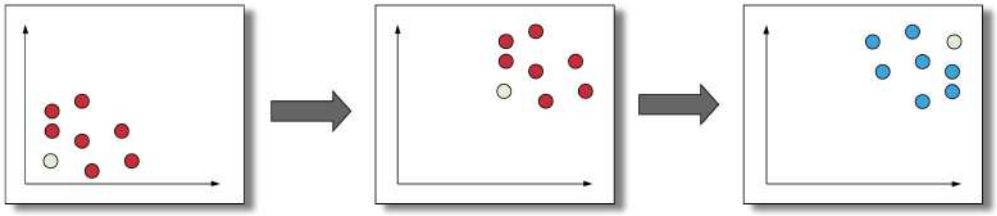
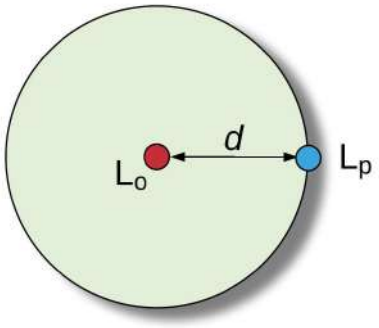


Figure 9: The privy method. The lightly colored dot resembles a dislocated point and facilitates the comprehension of this method (adapted from Ajayakumar et al. (2019)).

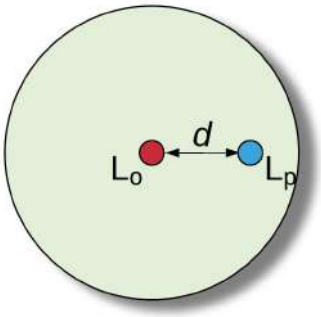
The masking degree of such methods is always constant and they do not consider underlying population density (Gupta and Rao, 2020) as utilized in other methods, making the instability range equal for all data points (Gupta and Rao, 2020). A strength of all affine information methods is that they maintain directional information (Haley et al., 2016), while rotate, translate and the privy technique also provide an overall point density as well as a relative point density. On the other hand, the masked data sets by all affine transformation methods except for translate have a vague or even no resemblance to the original pattern, and thus, are useless for research and analysis due to the loss of spatial attributes (Haley et al., 2016). Another drawback of the methods can be observed by insufficient anonymity (Wieland et al., 2008): The first three methods transfer all ODP by either a set increment, angle, or rotation. Through that, a re-identification of individuals is quite simple in contrast to the stochastic mechanism, concentration of isomasks, or the privy method which underlie a more complex obfuscation technique. Furthermore, the methods do not consider the underlying local geography or, for instance, the number of patients, too. As a consequence, the possibility of re-identifying an individual depends on these two variables and the missing underlying population density. For example, when applying rotation on a data set, it is evident that the original location must be within proximity. Supposing an uninhabited area such as a body of water or a cliff intersects this region, then the area where the original location developed from, would be reduced (Wieland et al., 2008; Gambs et al., 2010). Further, all ODP can be transferred to a location that is incorrectly linked with the research resulting in false identification. Finally, the displacement constants are not able to be exchanged (Haley et al., 2016). As evident, the affine transformation techniques have various drawbacks which is why they are not preferred by scientists (Kounadi and Leitner, 2015a).

The next class of geomasking techniques is represented by **RP** - sometimes referred to as *adaptive geographical masking* (Kounadi and Leitner, 2016) -, which is a technique that changes the location of the data points without aggregating (Armstrong et al., 1999; Kwan et al., 2004) but by transferring every data point by an arbitrarily determined direction and increment (Armstrong et al., 1999; Zandbergen, 2014). Various mechanisms were developed by, i.e., Hampton et al. (2010), Leitner and Curtis (2004), and Kwan et al. (2004): For instance, the *weighted random mask* aimlessly rotates every data point on a circle applying the variable radius r, where r depends on the underlying population density (Kwan et al., 2004). When applying the *grid masking* method, every original data point is moved to uniform grid cells (A. Curtis et al., 2011; Leitner and Curtis, 2004). The *circular mask with a fixed radius* moves the ODP in a random direction (0°- 360°) within a circle based on a set displacement as can be seen on figure 10a (Kwan et al., 2004; Kounadi and Leitner, 2015a), while the *circular mask with a random radius* is a similar technique to the former but differentiates due to random radius displacement which is between $l_o \leq d_i \leq d$. Here, d_i represents the selected random displacement, l_o resembles the original point location, and d symbolizes the upper bound of the radius (Kwan et al., 2004) (see figure 10b).

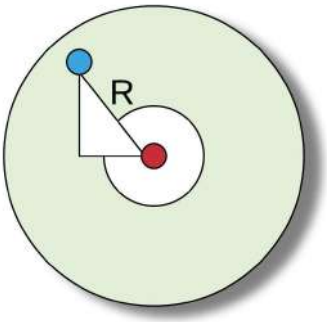
The rarely applied *triangular displacement* moves the ODP based on the Pythagorean Equation as depicted in figure 10c. Contrary to other techniques, this approach is not only considering the population density to dislocate ODP but also data sensitivity, research type, quasi-indicator availability, temporal aspects, and end-user type (Murad et al., 2014). *DM* aimlessly displaces the perturbed points outside of a buffer zone around the original point location as illustrated in figure 10d (Hampton et al., 2010). Furthermore, each location is displaced a range inversely proportional to the population density underneath. Through that, the spatial error is decreased while privacy protection is supported (Hampton et al., 2010). This technique has greatly increased privacy protection with an insignificant impact on the specificity and sensitivity of encountering



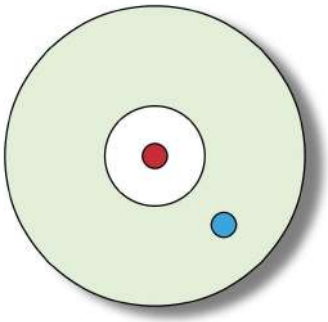
(a) Circular mask with a fixed radius (adapted from Gupta and Rao (2020)).



(b) Circular mask with a random radius (adapted from Gupta and Rao (2020)).



(c) Triangular displacement (adapted from Murad et al. (2014)).

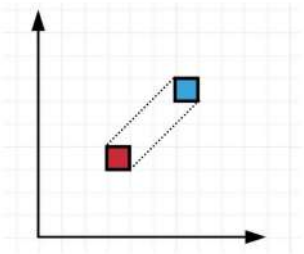


(d) Donut Masking (adapted from Zandbergen (2014)).

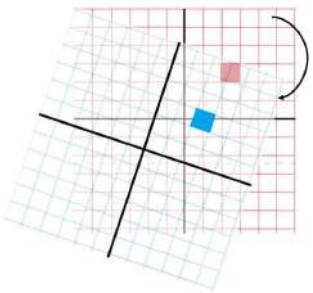
Figure 10: The two circular masks (a,b), triangular displacement (c), and DM (d).

disease clusters (Allshouse et al., 2010). Contrary to the other geomasking techniques, DM is able to have both a variable and a constant masking degree: due to its methodology, DM can be simply utilized with a fixed maximum distance, consequently converting the variable masking degree into a constant one (Kounadi and Leitner, 2015b).

When applying *local random translation*, a random translation factor δ translates every data point within a grid cell (Leitner and Curtis, 2006) (see figure 11a) while implementing the *local random rotation technique*, the center of every grid cell with an ODP is rotated (see figure 11b) (Leitner and Curtis, 2004).



(a)



(b)

Figure 11: Local random translation (a) and local random rotation (b) (adapted by Gupta and Rao (2020)).

Unlike many geomasking techniques, the *location swapping* method considers the surrounding geography. This mask switches an original data point with a masked data point based on all possible locations with alike geographic characteristics (Zhang et al., 2017) and assures that the masked points are transferred to predefined regions. In addition, research has shown that the masked point pattern is very similar to that of the original point pattern (Zhang et al., 2017).

Weighted random mask, DM, triangular displacement as well as AAE and VM employ the underlying population density to calculate the span an individual needs to be dislocated to provide locational privacy to a specific level. As a reminder, one location cannot be distinguished by at least K-1 other locations within a specific area (Ghinita et al., 2010). Yet, this technique almost always counts on the presupposition that the underlying population density is homogeneously dispersed. However, this precondition is barely met and increases the possibility of under-protecting individuals when mapping discrete data (Allshouse et al., 2010). Adding to that, as already criticized about the affine transformation techniques, all of the RP-methods do not consider the local geography or number of patients or victims apart from the underlying population density. Considering applying the circular mask with a random radius, the original data point is dislocated a specific distance *d*. Through that, it is evident that the original position has to be located within the circle with distance *d*. In the case of this region being part of an abandoned area or a body of water, the possible area for the original location can be easier identified, as depicted in figure 12.

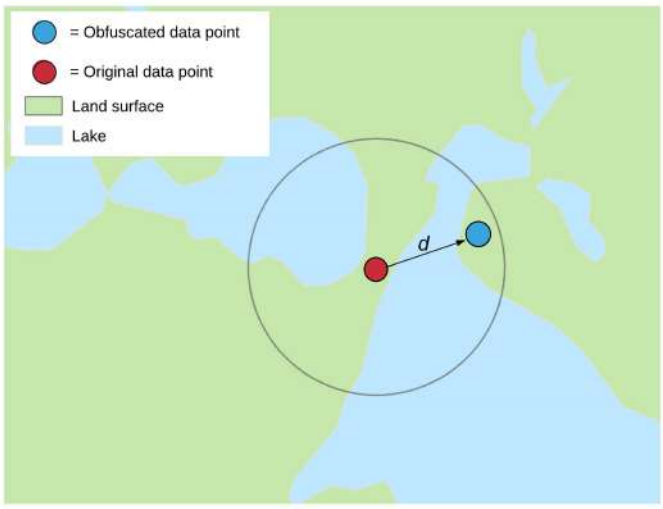


Figure 12: The circular mask with a random radius applied in Northeastern Germany. The re-identification is simple since the original location must be located within a circle with the distance *d*. Adding to that, the possible area of the original location is narrowed down due to lakes, facilitating the re-identification (data retrieved from DIVA-GIS).[3]

Finally, all geomasks within the RP-class are confronted with the same issue described before: data points can be transferred to residences that participate in the research already and seek protection or become a part of the research incorrectly by gaining a data point leading to false identification (Haley et al., 2016). An advantage of RP-techniques is that the displacement of the ODP can be limited by geographic boundaries, i.e. administrative units or census tracts (Haley et al., 2016).

The **flipping methodology** was developed by Leitner and Curtis (2004). These mechanisms reverse the original coordinates about both central axes or the horizontal or vertical axis. *Global axes flipping* flips the vertical and horizontal axes of the map to obscure, *global vertical flipping* flips the only the vertical central axis to disguise, *global horizontal flipping* flips the horizontal central axis to conceal, and last but not least, *local random flipping* flips vertical, horizontal, or randomly both axes of each grid cell to shield the data points as depicted in figure 13. (Leitner and Curtis, 2004; Leitner and Curtis, 2006). The methods depict the same problems as discussed before: a wrong re-identification can occur or debarred places of living become part of the research all the sudden since the area, the points are transferred to, is identical for all data points (Kounadi and

[3]http://www.diva-gis.org/gdata. Last accessed on May 8th, 2020.

Leitner, 2015b). Adding to this, the point pattern changes (Kounadi and Leitner, 2015a) heavily as depicted in figure 13 making the obfuscated data set useless for taking right decisions (Gupta and Rao, 2020). Finally, the flipping methodology techniques are as unpopular as the techniques including scaling, translation, and rotation, and are, therefore, not preferred by scientists neither (Kounadi and Leitner, 2015a).

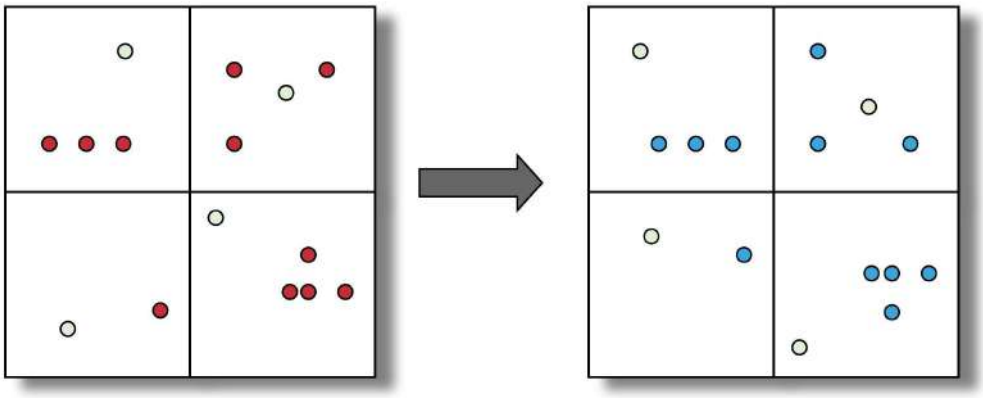


Figure 13: An instance of the local random flipping methodology by Leitner and Curtis (2004). The light green dots symbolize a dislocated point to enhance the understanding of this mechanism.

The category **blurring** represents a method to make data points barely perceivable. Through this, the risk of re-identification is reduced (Gupta and Rao, 2020; Haley et al., 2016). The blurring mechanisms can be differentiated between *Bimodal Gaussian displacement* (A. Curtis et al., 2011) - sometimes referred to as *randomized skew* (Haley et al., 2016) - and *population-density-based Gaussian spatial blurring* (Cassa et al., 2006), also called Gaussian skew (Haley et al., 2016). The former is implementing a bimodal Gaussian distribution to calculate a random displacement (A. Curtis et al., 2011) within a square (Cassa et al., 2006) while the latter also applies a Gaussian distribution for irregular displacement of data points which is contrary based on the underlying population density (Cassa et al., 2006;) (see figures 14a and 14b).

(a) Bimodal Gaussian Blurring.

(b) Population-density-based Gaussian Blurring.

Figure 14: Blurring mechanism (adapted by Gupta and Rao (2020)).

An advantage of considering the population density is that points are displaced greater distances in rural areas than in urban areas (Cassa et al., 2008) which does not apply for bimodal Gaussian blurring since this method does not consider the underlying population density (Cassa et al., 2008).

A strong detriment is that a release of multiple data sets that are obfuscated by the population-density-based Gaussian spatial blurring can facilitate restoring the ODP (Cassa et al., 2008). Besides, the masker must have knowledge of the re-identification risk which makes this mechanism more complicated to implement (Haley et al., 2016).

The **neighbor information**-category distinguishes between the *nearest neighbor distance mask, ordered nearest neighbor*, and *linear-programming identification*. The former determines the space information of every point to the two nearest neighbors. The second resembles approximately the relative positions of the points. ODP resemble set A, while masked points resemble set B. The latter applies linear programming. Hereby, the discrete locations of A are dislocated to some locations B (Kounadi and Leitner, 2015b).

The last category illustrates diverse masking techniques: *Contextual information* implements contextual information to change geographic identifiers (Kounadi and Leitner, 2015b). By *Using public key* or *private key encryption*, data can be also encrypted (Weiser and Scheider, 2014). Finally, *Matrix masking based on spatial smoothing techniques* illustrates a varying level and form of the masking, which depends on the disclosure risk and spatial pattern, deriving in a utility-risk profile (Kounadi and Leitner, 2015b).

Figure 4 also represents the masking degree which is further explained in the following paragraph:

In the previous paragraphs, it has become clear that the majority of masks move points within an area predetermined by the "masker". This area is called the "uncertainty area" and is determined by the masking degree and describes the disclosure degree of the data points (Kounadi and Leitner, 2015b). For example, when applying the circular mask with a fixed radius to protect the location of 50 individuals, the masker can choose a radius. In the case of a selected radius = 15 m, the masking degree is characterized by 15 m resulting in an uncertainty area of ca. 707 m^2. Within this uncertainty area, the disclosure level of one of 50 individuals is 2% (Kounadi and Leitner, 2015b). It is necessary to clarify that applying a higher masking degree will result in a new point pattern which is less spatially similar from the original pattern (Kwan et al., 2004; Cassa et al., 2006; Kounadi and Leitner, 2015b).

As depicted in figure 4, all geomasks apply a masking degree except for the methods contextual information, ordered nearest neighbor information, and nearest neighbor distance mask. These mechanisms either eliminate geographic identifiers or support spatial information. Hence, a masking degree is not applicable to these techniques. The other geomasks provide either a variable or constant masking degree. A variable masking degree declares that the extend of the uncertainty area differentiates based on a certain parameter which is expressed by the underlying population density in most cases (Kounadi and Leitner, 2015b). For instance, applying a mask with a variable masking degree on one data point within a region without any residents, the masker will need to expand the size of the region to reach a minimum amount of residents to provide privacy. Contrary to that, a constant masking degree describes the extent of the uncertainty area is identical for all data points (Kounadi and Leitner, 2015b).

It appears that most geomasking techniques are not able to maintain the original point pattern which can result in ineffective data sets from a research and analysis point of view (Gupta and Rao, 2020). Another problem is represented by the re-identification or false identification by transferring data points to prior excluded residences. Zimmerman and Pavlik (2008) have discovered that disclosing the obfuscated metadata leads to a decrease in confidentiality, too. In other words, there is not only the potential of re-identifying individuals by unmasked data but also by masked data. Thus, the quantity of disclosed information, the masking methodology, as well as its parameters must be considered before publishing the confidential data (Kounadi and Resch, 2018).

Not one of the introduced mechanisms has brought forth an unmitigated solution to provide spatial confidentiality, still, many of the discussed techniques offer efficiency by ensuring privacy in various scenarios (Shu et al., 2018; Wang et al., 2018). For instance, some of the aforementioned masks are appropriate for masking data points while others are efficient in maintaining the original point pattern (i.e., donut masking, circular mask with a fixed radius) (Gupta and Rao, 2020). Since the proposed method AVM will be developed by merging VM and AAE, the following two subchapters will give a more thorough insight into these two techniques.

2.3.1 Adaptive Areal Elimination

The AAE-geomasking method was created based on the concepts of geographical isomasks and adaptive geographical masking. The former assures privacy by moving the original locations within

uncertainty areas created by the masks as seen elaborated in chapter 2.3. The so-called uncertainty areas describe an area, where the obfuscated points are displaced in, e.g. torus or circle (Kounadi and Leitner, 2016). For instance, DM moves the ODP within an uncertainty area selected from a uniform distribution (Hampton et al., 2010) while the population-density-based Gaussian spatial blurring dislocates confidential points within a circle in a distance and direction based on a normal distribution (Cassa et al., 2006).

The latter, the adaptive geographical masking methods, move the ODP within uncertainty areas, whose sizes are based on the underlying population density, proposing SKA (Kounadi and Resch, 2018). As explained in chapter 1.2, offering SKA, each data point cannot be identified among k-1 locations. However, as already criticized in chapter 2.3, the methods proposed by Cassa et al. (2006) and Hampton et al. (2010), consider the population information within administrative boundaries, which influences the resolution applied for the disclosure information. Resulting from that, when calculating the displacement error, the methods assume that the underlying population density is homogeneously distributed - which is not the case in most instances. This assumption can result in obfuscated data points with a lower actual k-anonymity ($Kact$) than the estimated k-anonymity ($Kest$). Especially within areas with a great population distribution heterogeneity, this can occur (Allshouse et al., 2010).

Hence, AAE is aiming to ensure K-anonymity when the obfuscation method and its parameters are known, plus, the original locations cannot be re-identified among the k-cases. K-anonymity can be measured precisely when uncertainty areas do not lie on top of each other and when k-anonymity is applied at a lower or equal level of the available resolution (Kounadi and Leitner, 2016). To accomplish this, AAE can be used to a) either dislocate original data points randomly within k-anonymized areas or b) dislocate original data points to a position in the centroids of the k-anonymized areas (Kounadi and Leitner, 2016). Ergo, random perturbation based on AAE is named adaptive random perturbation (ARP) while point aggregation in AAE called adaptive point aggregation (APA). In this context, centroids are to be understood as the center of gravity, which are calculated by the polygon's n set of geographic coordinates X_i, Y_i, called vertices.

Centroid of X coordinate:

$$C_x = \frac{1}{6A} \sum_{i=0}^{n-1} (x_i + x_{i+1})(x_i y_{i+1} + x_{i+1} y_i)$$

Centroid of Y coordinate:

$$C_y = \frac{1}{6A} \sum_{i=0}^{n-1} (y_i + y_{i+1})(x_i y_{i+1} + x_{i+1} y_i)$$

Here, *"A"* describes the area of the polygon (Waller and Gotway, 2004; Bourke, 1988).
In that regard, random displacement describes the fact that discrete data points have the same possibility of being displaced anywhere within their k-anonymized area.

To execute the AAE-technique, two data sets are needed: a) discrete point file and b) a spatial data set that either includes an attribute with discrete information (e.g., as administrative units containing an attribute field with addresses) or represents discrete information (e.g., point data representing addresses). This attribute is called RoRi (risk of re-identification). Generally, RoRi can contain information such as addresses, households, or population.

Firstly, the data is pre-processed. If applying a point data set, the points are accumulated into polygons (Kounadi and Leitner, 2016). Secondly, a disclosure value (DV) for the RoRi-field will be delineated. Hereby, the DV describes the minimum K-anonymity which is used to obscure confidential information (Kounadi and Leitner, 2016). In the next step, the process of merging polygons starts: Depending on the prior defined DV, every polygon containing a lower RoRi-value than the DV, is merged with its neighboring polygon or polygons until each polygon has RoRi-values that are either greater than or equivalent with the DV. Hereby, the general spatial rule is considered: Every polygon is dissolved with the adjacent polygon that has the longest joint boundary expect for equilateral polygons. In that case, the merging process performs with all neighboring polygons or with the equilateral polygon. Through this third step, the K-anonymized areas are established assuring that the areas are reusable for the DV and the same data set (Kounadi and Leitner, 2016). In the fourth and last step, it can be chosen whether the discrete data points

are aggregated to the centroids of the merged polygons or randomly dislocated within the merged polygons. Figure 15 gives a detailed description of the implementation of AAE.

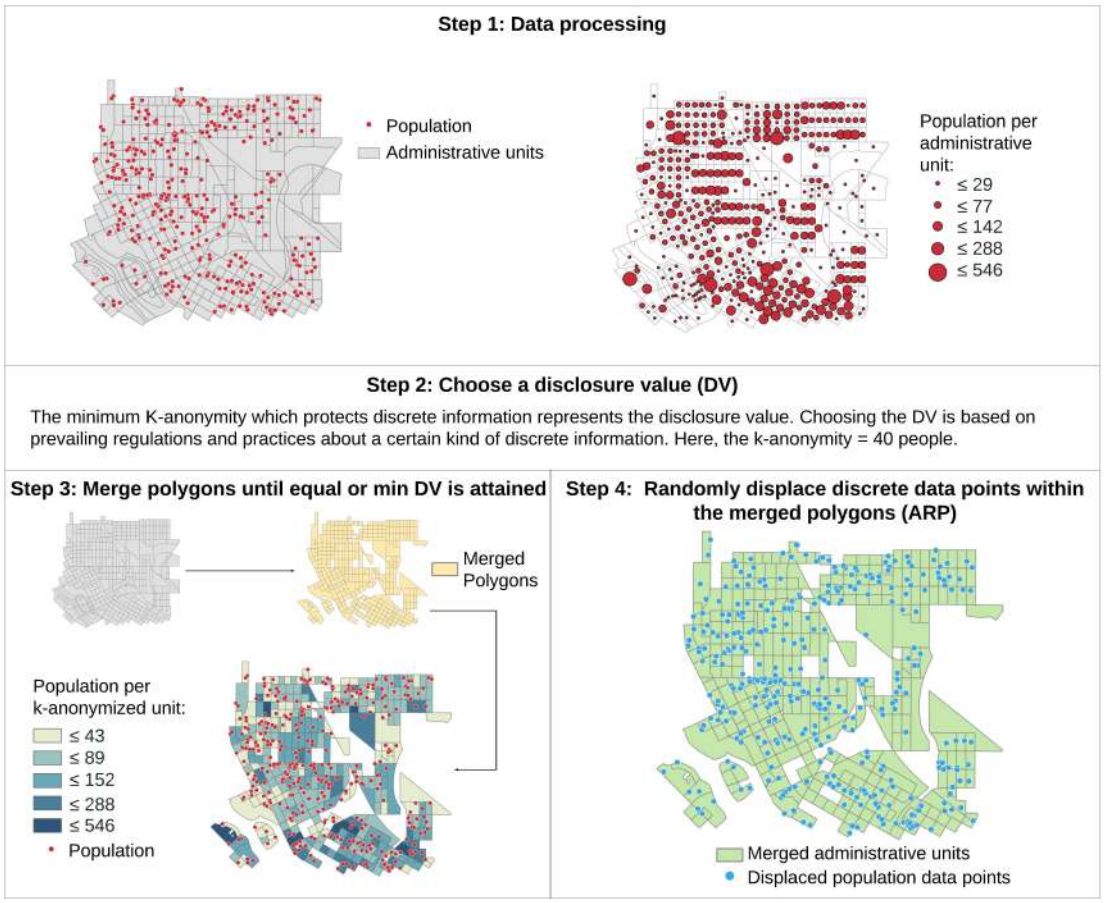


Figure 15: Illustration of obfuscating data points applying the AAE technique (adapted from Kounadi and Leitner (2016)) (data retrieved from ESRI).[4]

At the first sight, this technique seems to keep the spatial pattern of the ODP. However, when studying the merged polygons and the obfuscated points more closely, some masked data points are moved further distances than necessary. This can be explained by considering the process of merging polygons: For instance, some polygons with a small population might be merged with polygons consisting of a bigger population ensuring k-anonymity already. As a consequence, some data points from the second polygon might be moved to the first polygon with the smaller population although dislocating them in their original polygon shape would have been suitable to ensure k-anonymity from the beginning. This issue is illustrated in figure 16 which is based on the data applied in this research. In this research, ARP will be implemented which will be referred to as AAE for the remainder of this research.

[4] https://www.arcgis.com/home/item.html?id=0578ba603f0c47b99ba8e2abdb19efd7. Last accessed on February 26th, 2020.

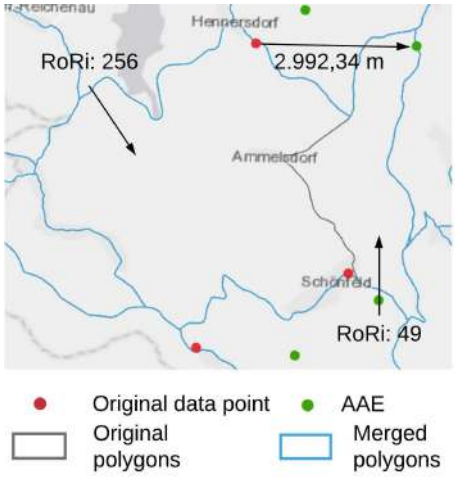


Figure 16: A data point is moved by AAE from its original polygon (RoRi = 256) to an area of the newly merged polygon based on a K-anonymity = 50. Nonetheless, this part consists of a polygon with a RoRi = 49. Furthermore, the displacement distance is at almost 3.000 m.

.

2.3.2 Voronoi Masking

The VM-method creates Voronoi polygons around the to be masked data points. Hereby, Voronoi polygons describe areas where the outer limits are equidistant between the neighboring data points or "where inside the polygons is closer to the corresponding point than to any other point" (Aurenhammer and Klein, 2000; Voronoi, 1908 in Seidl et al., p. 256). Afterwards, data points are displaced to the closest segment of its corresponding Voronoi polygon (Seidl et al., 2015; Gupta and Rao, 2020) (see figure 18).

One advantage of this method is that points in neighboring polygons are displaced to the same position, enhancing their K-anonymity (Seidl et al., 2015). Another asset of VM is that a higher point density will result in smaller distances between the original and obfuscated location resulting in a pattern that is very similar to that of the ODP set (Seidl et al., 2015; Gupta and Rao, 2020). Regarding a small scale area or an area with a minimum of two households, VM dislocates the ODP a lesser distance than compared to other masking methods (Gupta and Rao, 2020; Seidl et al., 2015) (refer to figures 17a and 17b) which do not consider the underlying settlement patterns (Seidl et al., 2015). Again, a similar point pattern can be assured which is valuable for further research (Gupta and Rao, 2020). Finally, Seidl et al. (2015) praise that in case of applying a data set that is including all residences within the area of interest, no displaced point will be located on an actual residence. This means, that none of the displaced data points stays in its original positions or at the centroids of other places of living. Through that, a false identification of residences is not possible (see figure 17a) (Seidl et al., 2015). However, this can be refuted for areas with estates of terraced houses which are commonly seen in bigger cities as depicted in figure 17b. Here, the data points are still located on the original building albeit not on their original house number. Therefore, VM cannot decrease the risk of false identification in this regard.

Another disadvantage of this technique can be demonstrated by applying VM in areas with scattered residences: Here, some data points will be dislocated large distances, which will change spatial patterns (Seidl et al., 2015). Adding to that, figure 18 depicts a smaller amount of masked data points than the ODP although the same number is present. As described before, this can assure a higher K-anonymity but the map viewer will not be aware that some data points represent at least two addresses, increasing the risk of spatially analyzing the output differently. It can be concluded, that the proposed VM-technique is an efficient approach regarding the preservation of the spatial point pattern. This has also been proved by Seidl et al. (2015) who implemented various statistical methods to evaluate the performance of VM. Nevertheless, the MDP can also result in a lesser number of *visualized* obfuscated data points as illustrated in figure 18 - although more are present in reality.

The algorithm for this obfuscating mask was automated in Python for ArcGIS Pro and ArcMap and is attached in appendix A.

(a) VM applied on an area with detached houses including all residences.

(b) VM applied on an area with semi-detached houses including all residences.

Figure 17: VM applied on address points within a neighborhood of detached houses (a) and within a neighborhood of semi-detached houses (b).

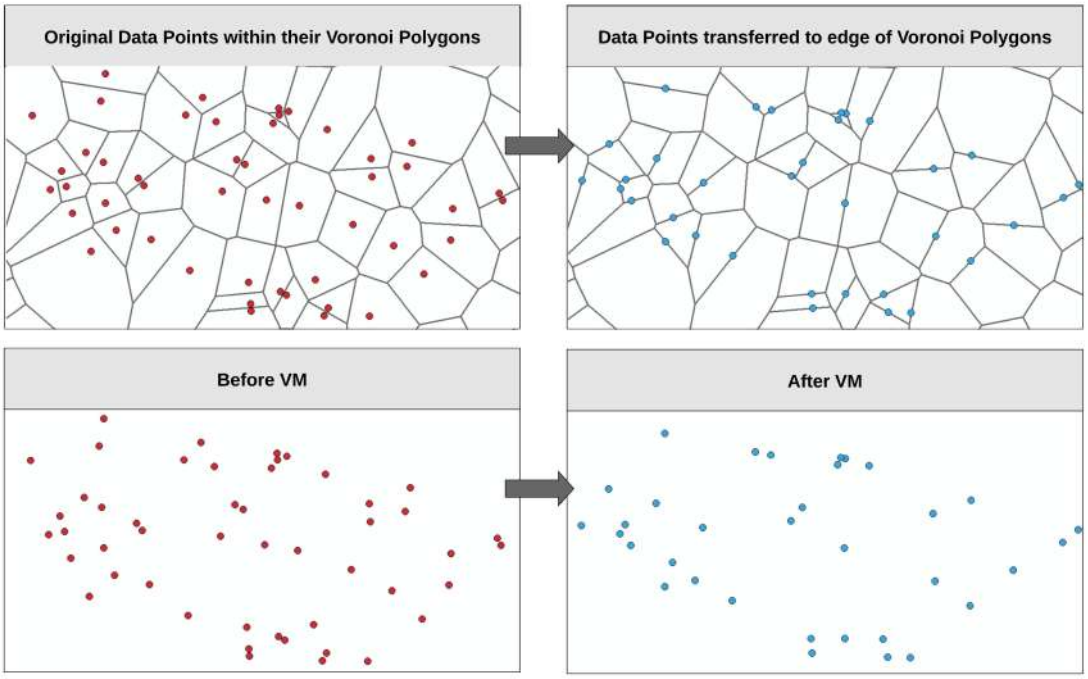


Figure 18: Graphical illustration of Voronoi Masking (adapted from Seidl et al., 2015).

.

2.4 Exploratory Spatial Data Analysis

To analyze whether the proposed method AVM is achieving the research motives as elaborated in chapter 1.3.2, this subchapter provides an outline of the methods of *exploratory spatial data analysis* (ESDA). ESDA supports the identification and characterization of locations, shapes, and magnitudes of statistically substantial patterns within an area of interest (Fu et al., 2014). Several of the aforementioned researches have implemented ESDA methods on ODP and MDP to investigate and compare the performance of obfuscating techniques. For instance, during the early stages of geomasking, Armstrong et al. (1999) scrutinized their masked output with that of the ODP by exploring the *dimension* of spatial information (Kounadi and Leitner, 2015b). For this, the authors grouped the dimension into four classes:

- pair-wise relations,
- "event-geography relations" (Kounadi and Leitner, 2015b, p. 743),
- anisotropies, and
- trends.

Thereby, the first class calculates the orientation and distances between the MDP and other spatial features and compares those with the ODP (Kounadi and Leitner, 2015b). For example, the distance between an obfuscated cholera case and a water body can be collated to the distance between the original cholera cases and the same water body. The second class is used to distinguish whether a geomask maintains the relative and actual distances as well as the relative orientation between the ODP and MDP (Armstrong et al. (1999) in Kounadi and Leitner, 2015b). Class three describes "the change of properties in different directions" (Kounadi and Leitner, 2015b, p. 743). This class as well as the last class test whether the mask sustain the directional pattern and the existence of anisotropies and trends (Kounadi and Leitner, 2015b).

Other authors, such as Seidl et al. (2015) applied the *kernel density difference*, *global Moran's I*, *distance to k-nearest neighbor*, the cross K-function - also known as *Ripley's K-function* (ArcGIS PRO, 2020) -, and the *nearest neighbor hierarchical cluster analysis* (NNHCA) on household-level data in Vienna, Austria, to juxtapose the effectiveness of grid masking, random weighted perturbation, RP, and their newly developed obfuscating technique VM. Kwan et al. (2004) scrutinized the performance of three RP-masks on death cases caused by lung cancer in Ohio, USA. The evaluation was based on the Ripley's K-function - as utilized in Seidl et al. (2015) -, "the visualization of point pattern" (Kwan et al., 2004, p. 18), and the kernel estimation of density surface. Leitner and Curtis, 2004 (2004) also analyzed the visualization of the point pattern.

As evident, several approaches exist to analyze the efficiency of geomasks. To not exceed the frame of this thesis and due to the predominant interested in the effects of spatial distribution between the ODP and MDP, this research will visualize the ODP and MDP, scrutinize the mean and median centers, the Ripley's K-function, and the NNHCA to analyze the dissimilarities regarding spatial information loss and the preservation of the same data granularity (Seidl et al., 2019) and, hence, assisting the finding of the answers for research questions one, two, three, and four (see chapter 1.3.2). The statistical methods are introduced in the following subchapters.

2.4.1 Visualization of Point Pattern

The visualization of the point pattern was implemented by Seidl et al. (2015) as well as Kwan et al. (2004). The former used this technique to scrutinize the extent of the ODP and compare it with that of the MDP whilst the latter implemented this method to study the derived output. This technique will be applied to investigate whether the MDP are displaced on other residencies increasing the risk of false re-identification or are transferred to void locations such as forests or lakes.

2.4.2 Mean and median centers

As a second method, the mean and median centers of the ODP and the MDP will be studied. This method has been applied by Seidl et al. (2015), Hampton et al. (2010), and Gupta and Rao (2020) to compare the center of the point distributions per data set (mean) (Burt et al., 2009; Ebdon, 1988) and to calculate the central tendency of outliers (median) to model the location-allocation of the data sets (Wong, 1999). When comparing the output of the MDP and ODP, the smaller the displacement distance between the ODP and MDP, the more effective is the geomask in terms of having a similar location-allocation like ODP. In conclusion, the analysis of the mean and median centers is a valuable measurement to summarize the overall locations of the ODP and MDP.

2.4.3 Ripley's K-function

The third technique to be utilized is the point similarity analysis (PSA) of Ripley's K-function that describes a spatial analysis to investigate point patterns. This function allows us to evaluate whether the obfuscated points are clustered, dispersed, or randomly distributed like the ODP (Kwan et al., 2004, Seidl et al., 2015, ArcGIS PRO, 2020) and whether the point distribution

remains linked or not. In the case of linked point distribution, the geomasks perform spatially dependent on the ODP (Seidl et al., 2015). Complementary to other ESDA such as the Nearest Neighbor Analysis or the well-known Moran's I, the Ripley's K-function conflates spatial dependence regarding point feature scattering or aggregation *over a variety of distances* (ArcGIS PRO, 2020; Dixon, 2014) which returns a more detailed output. The other aforementioned methods do not include this attribute (Dixon, 2014). In recent research, this method has been applied by Seidl et al. (2015) to test whether the MDP are gathered around the ODP while Kwan et al. (2004) employed this technique to testify whether the dissimilarities between the point patterns are considerably distinct or alike in respect of the random simulations (Seidl et al., 2015).

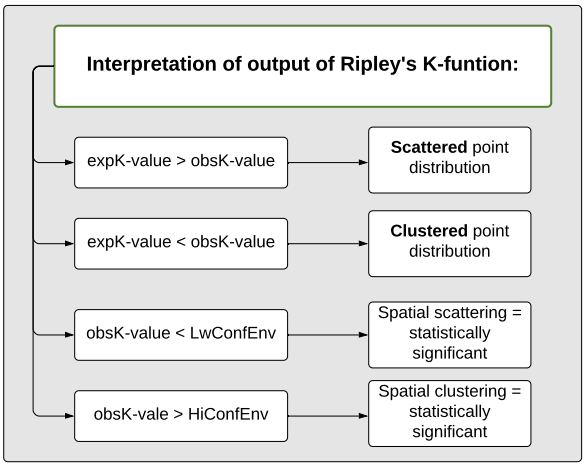


Figure 19: Interpretation of the returned values of the Ripley's K-function.

Ripley's K-function has also been recommended by Kounadi and Resch (2018) to provide the dissimilarities between the ODP and MDP. Hence, this function can be interpreted as a valuable tool for this study.

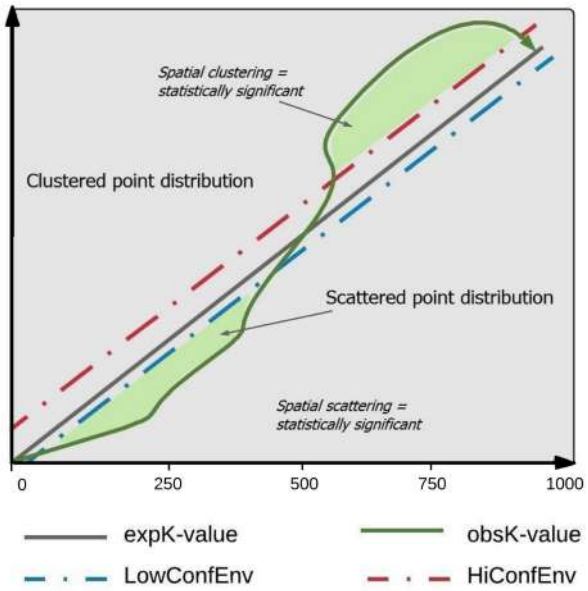


Figure 20: Interpretation of the returned values of the Ripley's K-function (adapted from ArcGIS PRO (2020).

By analyzing the spatial patterns over several lengths as well as spatial scales, the point patterns alter. Through that, the impact of spatial processes is demonstrated and it can be reflected in how the spatial scattering or aggregating of point centroids shifts when the size of the neighborhood

varies (ArcGIS PRO, 2020). This interpretation is based on the derived outputs *expected K-value* (expK-value), *observed K-value* (obsK-value), *upper confidence envelope value* (HiConfEnv), and *lower confidence envelope value* (LwConfEnv). For instance, assuming the expK-value is **larger** than the obsK-value, the point distribution is more scattered than a random distribution. In the event of the expK-value being **lower** than the obsK-value, the point distribution is more aggregated than a random distribution. Supposing the obsK-value is **lower** than the LwConfEnv, the spatial scattering is statistically significant. Finally, in case of a **higher** obsK-value than the HiConfEnv, the spatial aggregation is statistically significant, too (ArcGIS PRO, 2020) (see figures 20 and 19). The confidence envelope is built on random distribution. This so-called *permutation* can be selected by either 9, 99, or 999. In the researches by Kwan et al. (2004) and Seidl et al. (2015), the permutation of 99 was selected, which describes a point distribution of 99 times (ArcGIS PRO, 2020). Afterward, for each distance, the k-value with the biggest quantity beneath and above the expK-value becomes the confidence interval and can be analyzed.

For this study, only the expK-values and obsK-values will be investigated.

2.4.4 Nearest Neighbor Hierarchical Cluster Analysis

In almost every prior mentioned research, the impact of obfuscating methods on original hot spots have been probed. This is of great significance since clustering detecting plays a vital role in spatial analysis (Kounadi and Leitner, 2015b). For instance, by detecting hot spots, high concentrations of crime incidents can be explored and predicted for future scenarios (Levine, 2008; Chainey et al., 2008 in Kounadi and Leitner (2015)). NNHCA has been often applied in crime analysis, but can be used in every point distribution (Kounadi and Leitner, 2015b) presenting a beneficial instrument for this research.

Many statistical approaches exist to identify hot spots (Everitt, 1974) and are usually known as *"cluster analysis"*. These statistical methods intend to group "cases into relatively coherent clusters" (Levine, 2004, p. 2). Hereby, hierarchical clustering can be defined as the grouping of data built on characteristics they obtain (Grubesic and Murray, 2001; Xu and Wunsch, 2008). This is done by calculating a (n x n) matrix and dissimilarities between each data point (D), which is mostly done by an Euclidean metric. D defines the groups and connects the most comparable observations into clusters. The nearest neighbor distance is commonly used as the dissimilarity component (Bailey and Gatrell, 1995), as seen in Seidl et al. (2015). It compares the average width between all data points with the space between two data points. In case of the distance matching the *a priori* criterion (normally the determined possibility of a threshold distance between groups of points or between two points), the observation is connected and forms a new cluster. This method is repeated until all data points have been grouped into the first-order cluster. Afterward, the first-order cluster is then examined for the second-order cluster repeating the procedure (Bailey and Gatrell, 1995). This process can be described as a *dendogram* also known as "an inverted tree diagram" (Levine, 2004, p. 2) and is depicted in figure 21.

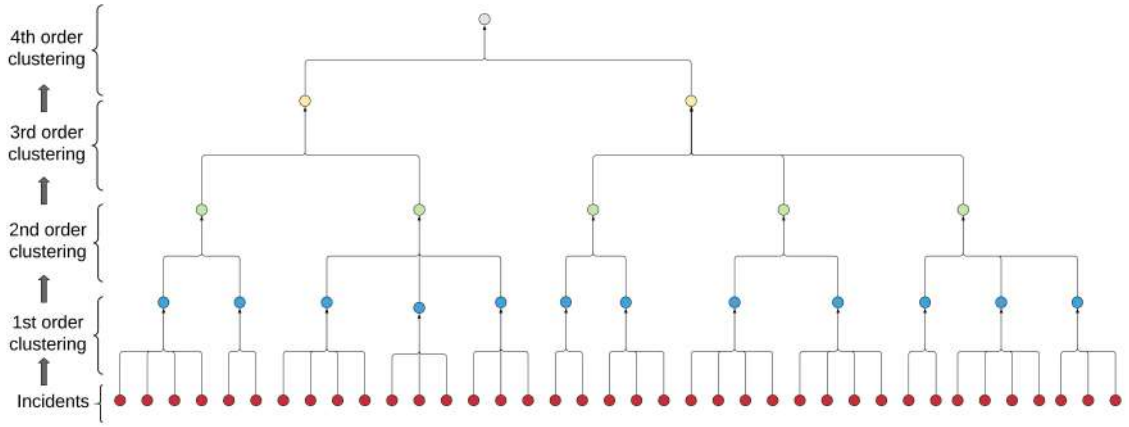


Figure 21: Illustration of the Nearest Neighbor Hierarchical Cluster Analysis (adapted from Levine (2004)).

Implementing this method allows not only the examination of the first-order cluster but also

the cluster density and the standard deviational ellipses (Seidl et al., 2015). In sum, this statistical method is applied to compare the cluster of the ODP with the MDP regarding size, orientation, density, and amount of cluster ellipsoids as illustrated in figure 2.4.4.

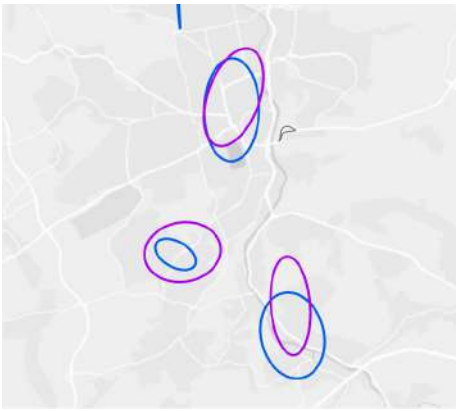


Figure 22: Hypothetical comparison of standard deviational ellipses of two obfuscating techniques.

The NNHCA specifies groups of cases that are locationally close and groups all data points into one single cluster or the method does not succeed (Levine, 2004). Hereby, within the program CrimeStat, a threshold distance is defined that is collated to "the threshold to the distances for all pairs of points" (Levine, p. 14). Now, only the data points which are nearer to one of more data points than the set threshold distance will be chosen for the clustering. Moreover, a minimum number of data points that be contained by the cluster, can be selected. The first order cluster is based on both criteria (being part of a group including the minimum number of data points and being closer than the selected threshold value) (Levine, 2004) (see figure 23).

Figure 23: Hypothetical illustration of selecting points for clustering.

Afterward, as elucidated above, the program keeps clustering and develops a hierarchy of clusters until all clusters are joined into one single cluster or the clustering deteriorates - which is highly possible (Levine, 2004).

3 Methodology

The main objective of this study is *to combine the two geomasking techniques VM and AAE to create the mechanism AVM through which the individual's privacy is protected while the false re-identification risk is decreased and data utility maintained.* To achieve this, this chapter presents an outline of the methods applied as depicted in figure 24. In the first step, the data will be processed (see chapter 3.2.2). As a second step, the AVM-method is developed. For this, chapter 3.3 indicates a workflow to explain the performance of this algorithm. Thirdly, the original data set created in the first step will be masked by applying the mechanisms of AAE, AVM, DM, and VM followed by calculating the mean and median centers and by applying the statistical methods Ripley's K-function and Nearest Neighbor Hierarchical Cluster Analysis (see section 3.4). At last, it is elaborated how the data sets and work environments are prepared for the to be applied statistical methods (refer to chapter 3.4).

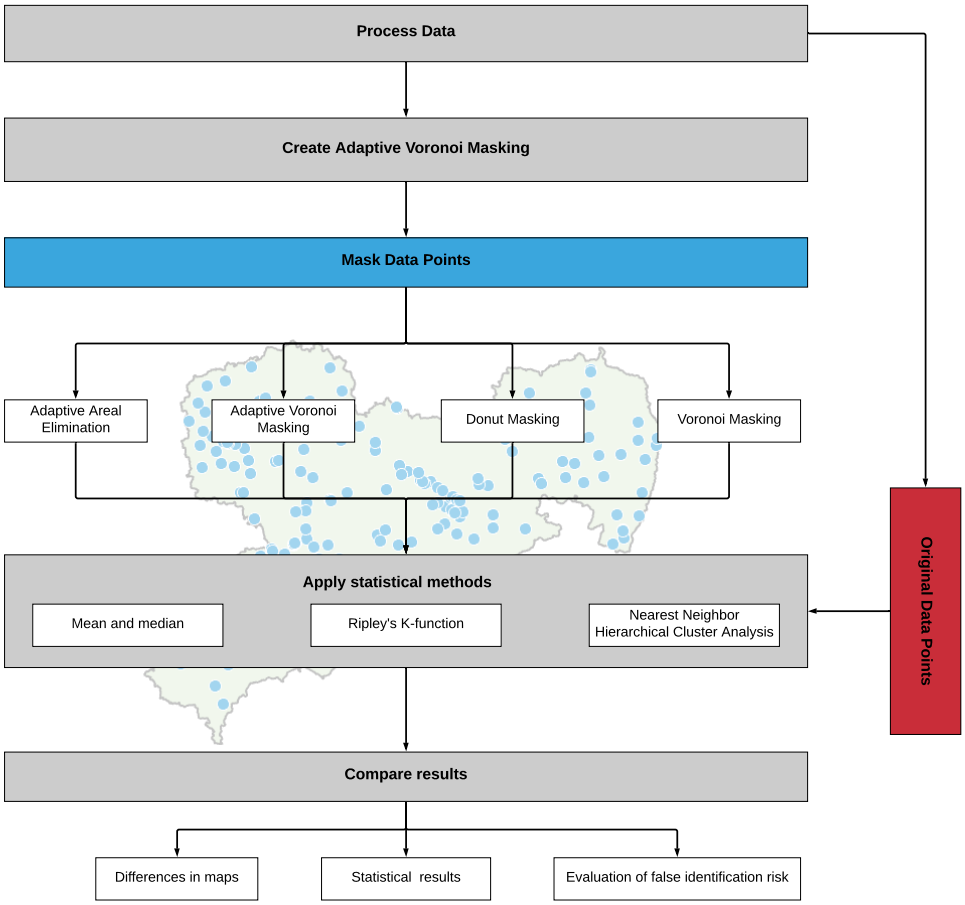


Figure 24: Overview of methodology applied to analyze the performance of geomasks.

3.1 Study Area

The research area is taken place in the Free State of Saxony in Eastern Germany. It is adjacent to the federal states Brandenburg, Saxony-Anhalt, Thuringia, and Bavaria as well as the countries the Czech Republic and Poland (from the north, counterclockwise). In total, Saxony counts 419 municipalities and 13 districts as illustrated in figure 25 (Bundesamt für Kartographie und Geodäsie, 2019).

Regarding population, Saxony has more than 4 million inhabitants (Statistisches Landesamt Sachsen, 2020) of which more than 563.000 were registered in the state capital Dresden in the end of 2019 making Dresden the 12^{th} biggest city in Germany (Dresden, 2020).

Yet, the highest population and population density are found in the city of Leipzig with a total of 587.857 people and 1.974 inhabitants per km^2. Contrary to that, the district Nordsachsen has only 97 inhabitants per km^2 - the lowest in Saxony (Sachsen, 2018). Hence, the State of Saxony is

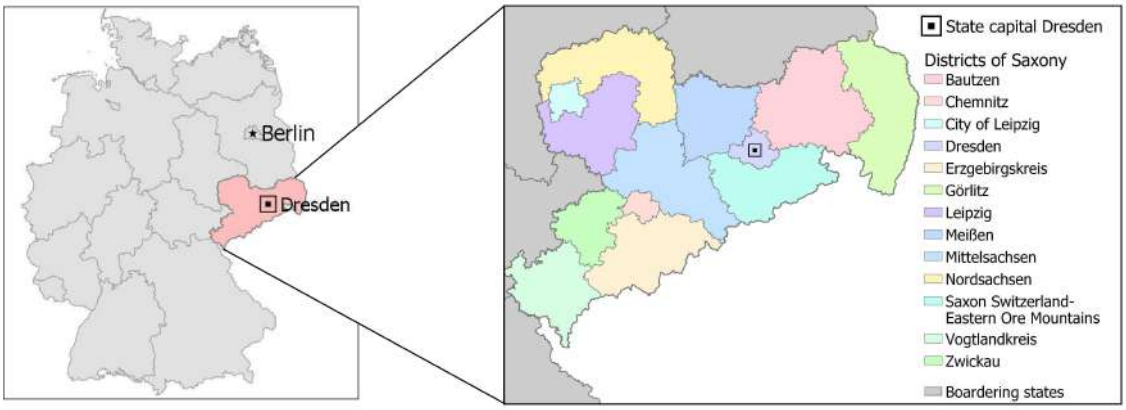


Figure 25: The state of Saxony (depicted in rosé) with a close-up of its districts (Bundesamt für Kartographie und Geodäsie, 2019).

an explicit choice to investigate the performance of various geomasking techniques because it has highly populated as well as rural areas. Thus, the geomasking methods will be applied on the State of Saxony, the city of Leipzig - since it has the most inhabitants and the highest population density - as well as on the district of Zwickau. Zwickau was chosen because when calculating the average inhabitants (ca. 313.685) and population density (ca. 493/km^2) per district in Saxony, Zwickau has the closest values (inhabitants: 317.531; population density: 334/km^2) (Sachsen, 2018).

3.2 Technical Implementation

3.2.1 Software

The software being used for this research are ArcGIS Pro 2.5 by the international GIS-software developer ESRI (ESRI, 2020a) and CrimeStat 3.3. by Levine & Associates (2020). ArcGIS Pro is used for data exploration, visualization, for running the AAE- and DM-algorithm, and for the creation of the AVM- and VM-algorithms. Hereby, the embedded ArcPy-package is utilized. Moreover, the statistical methods introduced in chapter 2.4 - with the exception of NNHCA - will be operated in ArcGIS Pro. NNHCA will be performed in CrimeStat 3.3 which represents a program of spatial statistics for exploring locations of crime incidents (Levine, 2007). The software ArcGIS Pro runs under license, which is provided by Utrecht University. CrimeStat can be downloaded for free online.[5]

3.2.2 Data

For this study, two data sets are required: A point data set representing sensitive or confidential discrete data to employ the various geomasking techniques on, and a polygon file including the amount of addresses or population in each polygon. The choice of the study area is based on the availability of processed, free, and ready-to-use data. Moreover, the data must allow different levels of spatial granularity and population density.

For this, a line shapefile representing the street network in Saxony was derived online[6], and used to create streets blocks. Through that, it is aimed to develop blocks which are not too coarse but also not too small. Since the original road network file also included several street classes such as *"footway", "path",* or *"cycleway"*, all street duplicates as well as all street classes except for *"primary", "secondary"* and *"tertiary"* were deleted off the shapefile to generalize the road network. Thus, an enormous amount of small blocks is avoided resulting in shorter processing times of the obfuscating techniques. Also, only single-line road features in place of matched pairs of divided road lanes are maintained. Small and open configurations of roads are removed. Now, the tool *"Construct Polygons"* was used to create a polygon feature from the remaining polylines. Any polygons that developed outside the study were removed manually. Since the to be applied geomasks require several variables, further attribute fields were included:

[5] https://www.icpsr.umich.edu/CrimeStat/files/CrimeStat-3.3.zip. Last accessed on July 26th, 2020.
[6] https://download.geofabrik.de/europe/germany/sachsen.html. Last accessed on May 20th, 2020.

Firstly, and most importantly, the underlying population density has to be considered for the to be applied obfuscating mechanisms. For this, a point data set symbolizing population, addresses, or disease is demanded. Hence, a point data set demonstrating addresses in Saxony from 2018 was chosen and can be downloaded directly from ESRI[7]. Originally, the point data set consists of 947.164 data points. Despite the fact that this data set is not representing confidential data such as a specific disease, this data set is still suitable for this research since addresses represent a population distribution which is similar to that of a health distribution in most cases. Therefore, the points will be counted per block as the RoRi-information and, subsequently, reduced in size creating three random subsets of address point data. This was achieved by adding a column named *"Count"* to the point data set setting the data type to *"short integer"*. Then, by right-clicking on *"Calculate field"* of the newly created column, the "Count" is set to "1". Eventually, the tool *"Summarize within"* is applied to count the number of address points in each polygon. As a result, a new polygon shapefile was produced with two new columns *"Count of points"* and *"Sum Count"* specifying how many addresses lie within each polygon. Last but not least, a new column *"Rori"* was created and the values of "Sum Count" were added. Finally, the columns "Sum Count" and "Count of points" were deleted. Figure 27 gives a detailed description of this approach.

Secondly, a new attribute field *"id"* was established assigned with values enumerating the blocks. Thirdly, a string field describing the name of the polygon was produced and named *"id_text"*. The values from "id" were copied to "id_text"; this step is necessary to execute the algorithm of AAE. Thirdly, a string field (*"area"*) was produced by calculating the geometry of each polygon unit.

Finally, the in ArcGIS Pro embedded tool *"Clip"* is implemented to generate the study areas of Zwickau and the city of Leipzig. This is done by claiming polygon shapefiles depicting the municipalities and district from the *Federal Agency for Cartography and Geodesy* [8]. Resulting from the prior elucidated steps, the four geomasks will be applied on three data sets:

- The Free State of Saxony with 2735 polygons,
- The district of Zwickau with 257 polygons,
- The city of Leipzig with 223 polygons

In the previous paragraph, it was stated that three random point data sets need to be formed. Again, the tool "Clip" was applied to ensure that only data points within the three study areas are considered. Thereafter, the X and Y coordinates were assigned to each data point since the geomask DM requires point data sets including coordinates. This was achieved by utilizing the tool *"Add XY Coordinates"* in ArcGIS Pro. Subsequently, the tool *"Create Random Points"* was operated on a point data set including 885.209 addresses in Saxony to create random data sets consisting of 20.000, 2.000, and 200 points for each of the aforementioned data sets mimicking different population densities. Solely a single point data set per area would not comply with the scope of this research (Kounadi and Leitner, 2015b) because of the efficiency of each obfuscating method, and particularly of AVM, is examined on different population densities. Although the data points are selected randomly by the tool "Create Random Points", they represent the population distribution, hence, they serve as realistic data sets. Throughout the remaining study, the three point data sets are called "original point data". In case of referring to a specific study area with a certain number of points, the data sample is named *study area + number of points* (i.e., Saxony 200).

However, due to the computational power of geomasking techniques - predominantly DM which results in growing processing time with an increasing amount of data points - the data point set consisting of 20.000 data points was removed from this study.

The data sets are appropriate for this study for several reasons: Firstly, the point data set only contains addresses and no further information about names, patients' locations, burglaries locations, diseases, age, or income. Hence, in case of transferring data points to a residence, false identification or a re-identification regarding one of the former attributes cannot jeopardize an individual (Kounadi and Leitner, 2015b).

[7] https://opendata-esri-de.opendata.arcgis.com/datasets/esri-de-content::adressen-sachsen?geometry=6.577%2C49.700%2C20.452%2C52.124&selectedAttribute=REGBEZ. Last accessed on April 14th, 2020.

[8] https://gdz.bkg.bund.de/index.php/default/open-data/verwaltungsgebiete-1-250-000-ebenen-stand-01-01-vg250-ebenen-01-01.html Last accessed on May 28th, 2020.

Furthermore, the address data set for Saxony was originally received from a governmental land-surveying and geoinformation institution in Saxony (*Staatsbetrieb Geobasisinformation und Vermessung Sachsen*) (ESRI Deutschland, 2020) promising an up-to-date and accurate data set, too. The data sets representing the municipalities and districts of Saxony were retrieved from a German governmental institution as well, ergo, regular data maintenance can be expected. Besides, the last update of this particular data set was in 2019, making the data current and valuable for this work. The data set resembling the road network in Saxony is OpenStreetMap-data (OSM). It only contains street names and was created in 2018, presenting a contemporary data set as well.

Yet, it must be elaborated that the clipped study areas do not completely correspond in size with the original district of Zwickau, City of Leipzig, or the Free State of Saxony (see figures 26a-26c). This is due to the removal of smaller streets creating quietly different sizes and shapes of the study areas. Nevertheless, this does not affect the purpose of this research which is the combination of VM and AAE developing AVM and comparing its performance with that of AAE, DM, and VM on two different scale point data sets in three different study areas.

Due to these facts, it can be concluded that the data sets are effective to compare the performance of various geomasking mechanisms regarding data utility and the risk of re-identification. Furthermore, since the original point data sets represent different population densities, the AVM-method can be utilized on high and low population density. Through that, it can be concluded whether the method is pertinent for different levels of population density.

(a) Leipzig. (b) Zwickau. (c) Saxony.

Figure 26: The three study areas compared to their original polygon units.

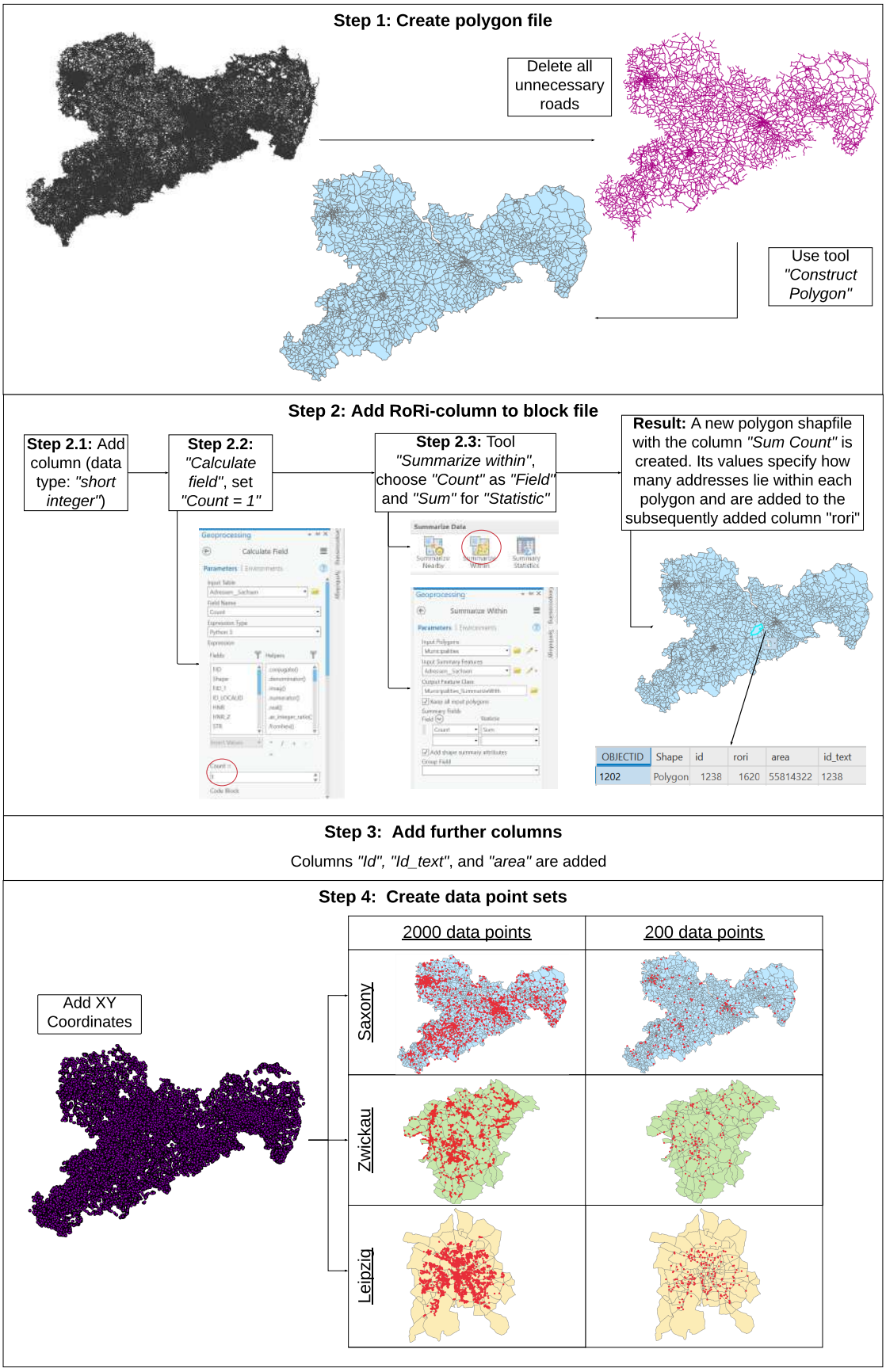


Figure 27: Work flow of data preprocessing

3.3 New Geomasking Technique: Adaptive Voronoi Masking

This research presents the new geomasking technique AVM which is based on the concepts of AAE and VM. As presented in chapter 1.2, the AVM algorithm is the first anonymization method that is constructed from two obfuscating methods and which aims to

- preserve data utility;
- consider topological polygon relationships;
- offer a high degree of spatial k-anonymity to attenuate the level of re-identification;
- considers the underlying topography.

Furthermore, it aims to extract the advantages as well as the disadvantages of VM and AAE:

For instance, an exceptionally great asset of the AAE mask is depicted by the consideration of the underlying population density while VM expresses an advantage regarding the displacement of the ODP. Exemplary, VM shifts the ODP smaller distances in case of a high point density creating a masked point pattern which is very similar to that of the ODP. Regarding a small scale area with few data points, VM displaces the ODP lesser distances (see chapter 2.3.2). Contrary to that, AAE has a tendency to shift data points further distances than necessary (refer to chapter 2.3.1) since it considers the outer segments of the merged polygons which present a greater possible displacement area than the polygon the data point was originally located in. Adding to this, data points can be moved to a part within the new dissolved polygon which used to have a smaller population density before it was merged with its neighboring polygon (see chapter 2.3.1). Another asset of VM is that it transfers ODP in neighboring Voronoi polygons to the same position which is praised by Seidl et al. (2015) since it improves the K-anonymity.

However, a detriment of both geomasks is that they do not consider the underlying topography. Data points can be displaced on lakes or in forests symbolizing illogical new locations (refer to figures 28a and 28c). Besides, when applying AAE as well as VM, data points can be displaced to residences resulting in false re-identification as depicted in figure 17b in chapter 2.3.2 or as illustrated in the figure below (figure 28b).

Figure 28: Graphical illustrations of AAE and VM moving data points to illogical locations (a, c) or increasing the risk of false re-identification (b).

Based on these remarks, AVM extracts the asset of considering the underlying population density by joining polygons as AAE does and displaces the ODP based on the concept of VM. In respect thereof, the ODP will be moved to the closest segment of their corresponding Voronoi polygon which is laying within their merged AAE-polygon as illustrated in figure 31. In case a Voronoi segment lies *outside* its dissolved polygon, the point is transferred to the boundary of the merged polygon and not to the edge of the Voronoi polygon. Through that, AVM intends to circumvent the predicament of moving data points to a polygon containing a different population threshold preserving the SKA. Further, the underlying topography is considered by moving the data points to the closest street intersection. Through that, AVM avoids shifting the ODP to

another residence causing false re-identification or to invalid locations such as water bodies or forests. Besides, a lower risk of re-identification can be assured by moving the ODP to a street intersection than by displacing the ODP to the nearest street segment due to a higher amount of surrounding buildings. The workflow of AVM is documented in figure 29.

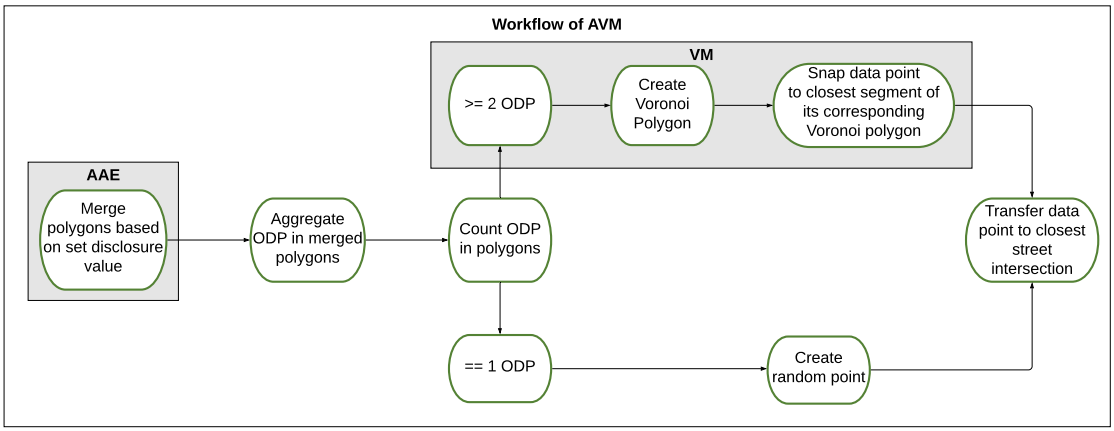


Figure 29: The performance steps of AVM.

Hence, to execute AVM, the following data sets are required:

- a discrete point file (as needed in VM and AAE)
- a spatial data set including RoRi (i.e., an attribute field with the amount of households) (as required in AAE)
- a line data set depicting streets.

Regarding the line data set illustrating the street network, it is noteworthy that the file is based on the same open street file as the one used for the artificial polygons in chapter 3.2.2. However, the street network for the intersection displacement is more detailed because it contains not only the primary, secondary, and tertiary classes but also the *"residential class"*. Thus, a smaller but yet meaningful displacement of the data points can be achieved. More street classes such as *"footway"* or *"cycleway"* were not included to prevent false re-identification.

Firstly, the data is pre-processed as done for AAE. The data points must be assembled within the polygons. Subsequently, a disclosure threshold for the RoRi-field is selected and polygons with a smaller value than the chosen DV are merged with its adjacent polygon until all polygons receive a RoRi-value that is greater or equal to the set DV. Here, the general spatial rule is applied defining that every polygon is combined with the bordering "polygon that has the longest shared border" (Kounadi and Leitner, 2016, p. 62).

Secondly, every data point that is laying within a polygon with at least two data points is transferred by the concept of the VM technique. It is guaranteed that the data points are replaced to the closest segment of their corresponding Voronoi polygon *within* their dissolved polygon.

Thirdly, the polygons containing only one data point randomly transfer the data point within their merged polygon.

Afterward, all newly displaced data points are shifted to the closest street intersection inside their united polygon.

Based on the data introduced in chapter 3.2.2, the following figure represents the results of AVM's steps. This example is located in the city center of Dresden. The code for establishing SKA-polygons is attached to appendix B. The algorithm of AVM was implemented in Python for this study (refer to appendix C) and can be executed in ArcGIS Pro as well as ArcMap.

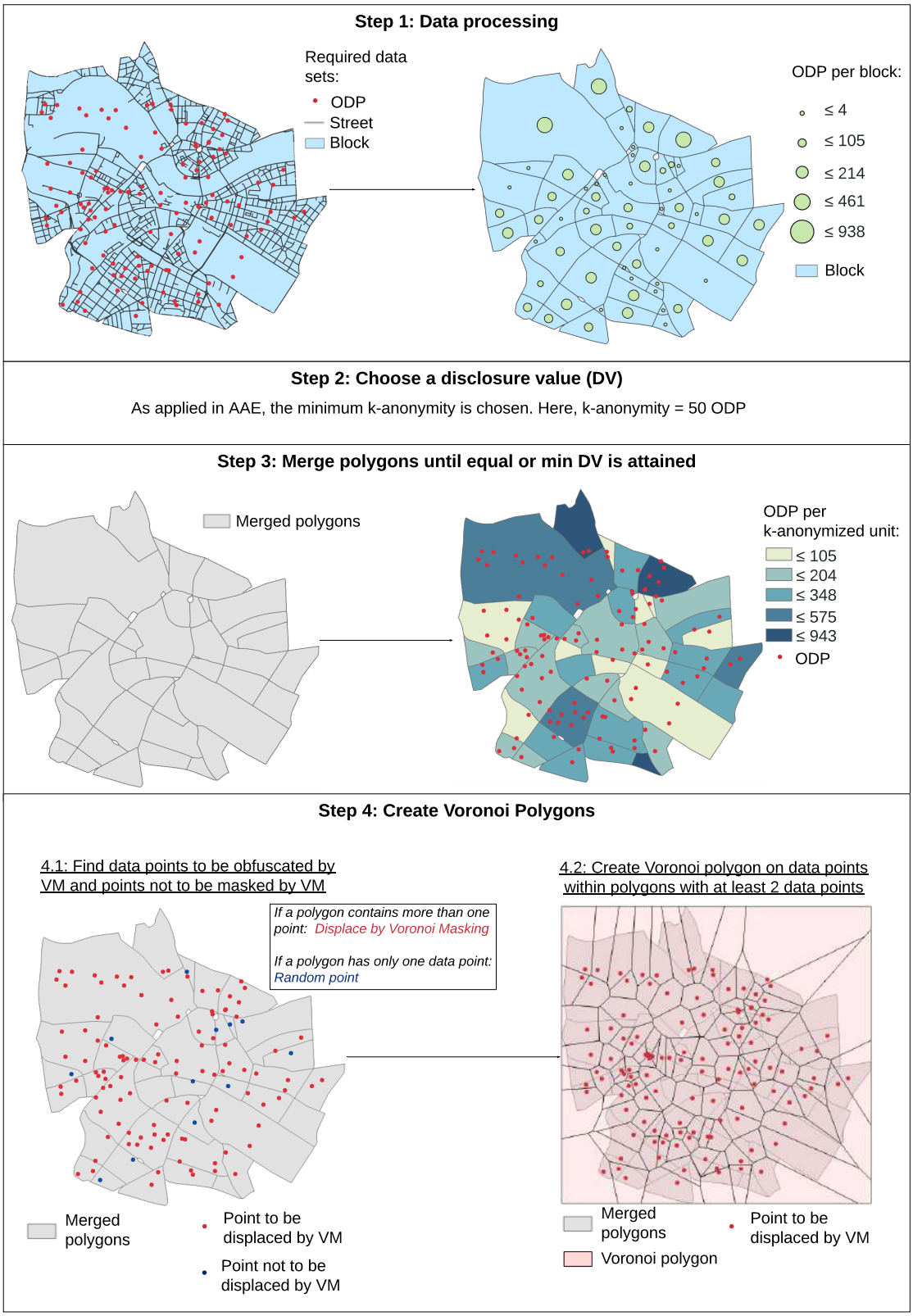


Figure 30: A detailed visualization of the performance of Adaptive Voronoi Masking in the city centre of Dresden (steps 1-4).

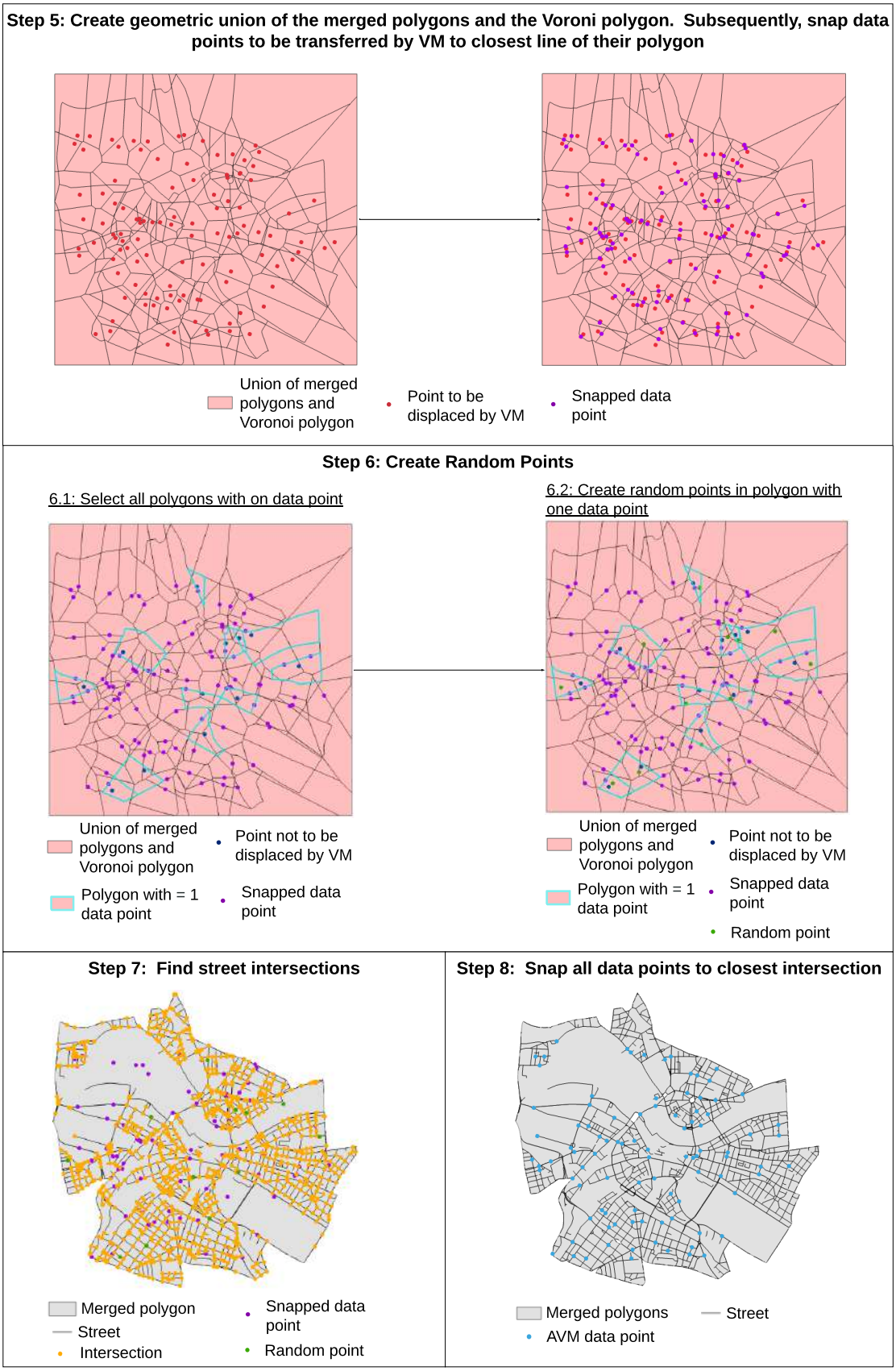


Figure 31: A detailed visualization of the performance of Adaptive Voronoi Masking in the city centre of Dresden (steps 5-8).

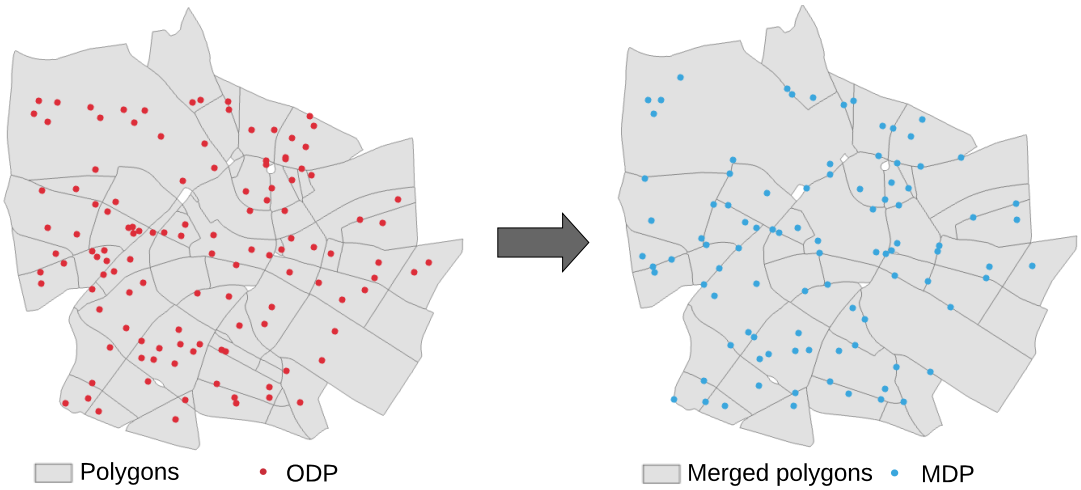


Figure 32: Data points before (left) and after the displacement by AVM (right).

A comparison of the ODP and the MDP by AVM is depicted in figure 32. For an improved visibility, the line feature class illustrating the streets was removed.

3.4 Methods To Evaluate Geomasking Performance

3.4.1 Visualization of Point Pattern

To study the MDP and compare them with the ODP regarding their extent of the point pattern and whether data points are moved to illogical locations or other domiciles, the MDP and ODP are visualized in ArcGIS Pro with a base map depicting the underlying topography. Moreover, the extent of the ODP will be visualized and compared with that of the MDP. For this, no further data pre-processing is required.

3.4.2 Mean and median centers

The tool "Mean Center" in ArcGIS Pro (ESRI, 2020b) is applied for calculating the overall position (Wong, 1999) of the ODP and the MDP whereas the tool "Median Center" (ESRI, 2020c) is implemented to model the location-allocation of the data sets (Wong, 1999). By doing so, the overall locations of the ODP and masked data sets are summarized and can be compared. No further data pre-processing is demanded calculating the mean and median centers of the ODP and MDP.

3.4.3 Ripley's K-function

To explore the Ripley's K-function, the tool *"Multi-Distance Spatial Cluster Analysis (Ripley's K-Function)"* in ArcGIS Pro is applied (ESRI, 2020d). As seen in former research (Seidl et al., 2015), the tool is implemented on each masking technique with a 99% confidence, thus, the tool is run with 99 simulations. The number of bands is set to 15 to explore the point pattern of the ODP and MDP changing over various lengths. The *"distance increment"* is set to an arbitrarily 50 m for the smaller areas Leipzig and Zwickau. Due to the vast size of Saxony, the distance increment is 300 m. As done in the previous study by Seidl et al. (2015), the border correction is ensured by employing the administrative boundary of the created block files for either Leipzig, Zwickau, or the state of Saxony depending on the point data set. For this, the blocks of each region were united by employing the tool *"Dissolve"* in ArcGIS Pro (ArcGIS Pro, 2020).

3.4.4 Nearest Neighbor Hierarchical Cluster Analysis

Unlike the previous statistical methods, NNHCA is employed in CrimeStat. To implement this tool, the data setup has to be arranged correctly: First, the coordinates from each study area are manually added in the field *"Grid area"* in the tab *"Reference File"*. For *"Area"* in tab *"Measurement Parameters"*, the area of each study area is calculated in square kilometers (km^2)

in ArcGIS Pro. Every data point file of the ODP as well as MDP is selected as the *"Primary File"*. For the *"spatial description"* of the actual NNHCA, following values are elected:

At first, the *threshold distance* - ten - is set as the default value which describes a random nearest neighbor distance (Levine, 2004). This value is chosen due to the varying extents of the point data sets. Only points that are closer than the threshold distance are picked for a cluster. For the second parameter, the *minimum number of points*, value five was selected to assure the finding of clusters since the previous simulation runs for this research using higher values (i.e., ten or 15) had resulted in zero clusters. Thirdly, the NNHCA is run with 99 simulations. The chosen simulation parameter assigns the routine to arbitrarily impute N cases to a square of the same shape as the set study area measuring the total clusters based on the other selected parameters (e.g., the minimum number of points, threshold distance). This simulation is repeated 99 times, valuating the "approximate confidence intervals for the particular first-order Nnh" (Levine, 2004, p. 6.30).

4 Results

This chapter demonstrates the results of the ESDA. The statistical approaches are applied to the ODP as well as on the MDP. Through that, the effects of the obfuscating techniques on the ODP and its pattern can be examined (Kounadi and Leitner, 2015b). The evaluation starts with a visualization of the MDP (see chapter 4.1), followed by a broad analysis by examining the mean and median centers (refer to chapter 4.2). Thereafter, the Ripley's K-function is utilized (see chapter 4.3) offering a statistical analysis of the spatial distribution between the ODP and MDP (Seidl et al., 2015, Kounadi and Resch, 2018). The last method is the more detailed NNHCA (refer to chapter 4.4) that is also considering the local point pattern to identify the cluster density, "number of first order clusters", and the "standard deviational ellipses" (Seidl et al., 2015, p. 255).

As a reminder, the different methods are employed to conduct dissimilarities regarding spatial information loss and the preservation of original data granularity (Seidl et al., 2019). In the ideal case, the spatial analysis of the AVM-masked data points will be equal to that of the ODP (Hampton et al., 2010).

Another fact to be reminded of is that from all aforementioned and applicable geomasking techniques with a variable masking degree, donut masking was chosen as a comparative geomask since it has a small effect on the geographical characteristics of the original point pattern as highlighted in academic literature (Hampton et al., 2010, Zandbergen, 2014, Allshouse et al., 2010, Lu et al., 2012). The algorithm for this method was retrieved online[9].

4.1 Visualized Outcome of the Geomasks

The figures 33a-38e depict the ODP and the results of the MDP by AAE, AVM, DM, and VM. The outer shape - or *extent* - of the ODP is illustrated with a purple square for an enhanced orientation. For output of the MDP and ODP including a background map, please refer to appendix D.3.

When comparing the figures 33a-38e, the resulted outcome by AAE is striking: AAE preserves the spatial extent of the ODP the least. For Leipzig 200, five data points (figure 33b), for Leipzig 2000 53 points (figure 34b), for Zwickau 200 six points, and for Zwickau 2000, 21 points (figure 36b) were dislocated outside of the original extent.

DM performs more successfully than the aforediscussed method: For Leipzig 200, only one point (refer to figure 33d), for Leipzig 2000 five points (see figure 34d), for Zwickau 200 as well as for Zwickau 2000 one point each (figures 35d and 36d) are not located within the extent of the ODP.

The new technique AVM retains the spatial extent more effectively than AAE and DM. For Leipzig 200 (refer to figure 33c) and Zwickau 200 (figure 35c) only three points are not contained within the original extents. Regarding the point data set Zwickau 2000, only one data point is transferred outside the extent of the ODP (figure 36c). The displacement outside the original extent can be justified on the creation of random points to the other end of an underlying merged polygon and the subsequent movement to the closest intersection.

The method VM is clearly outperforming the other geomasks regarding the preservation of the original extent. Only one data set (Leipzig 200) has one data point outside the original extent (see figure 33e).

Regarding the preservation of the point pattern, it is perceivable that AAE seems to abandon the pattern the most: Particularly figure 34b shows that the strongly visible floodplain forest (Stadt Leipzig, 2020) - that is intersecting the city of Leipzig from northwest to southwest - is not kept by AAE. Unlike that, AVM, DM, and VM maintain this meandering space in most regard (see figures 34c, 34d, and 34e) while AAE blurs this region completely. This is also noticeable in figure 36b: AAE does not sustain the several free spaces between the various point accumulations as presented by the ODP (figure 36a), AVM (figure 36c), DM (figure 36d), and VM (figure 36e).

Another factor to discuss is the *new* location of the MDP. This study has already discussed the problem of geomasks transferring data point to other residencies causing the risk of false identification or to locations which are obviously uninhabited (e.g., forest or lake). To examine whether the here applied geomasks move any data points to illogical locations or other domiciles creating false re-identification, the MDP and ODP were further inspected by underlying an OpenStreetMap as

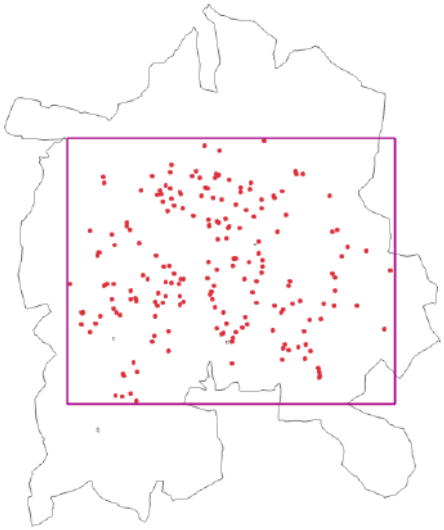
[9] https://mserre.sph.unc.edu/BMElab_web/donutGeomask/pyDonutGeomaskUserInstructions1.0.pdf. Last accessed on August 8th, 2020.

a base map depicting natural reserves, buildings, and infrastructures as well as water bodies (see figures 39a-39e).

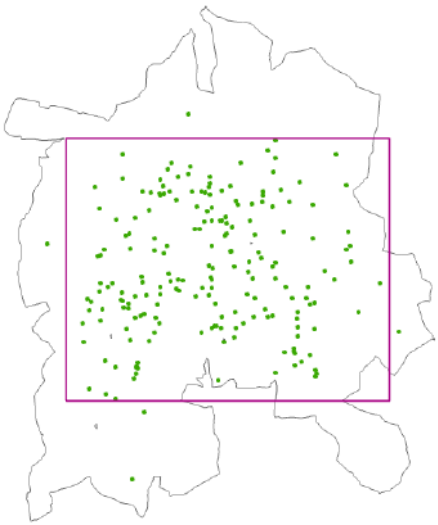
Figure 39a is illustrating a park in the city center of Leipzig. It can be noticed, that all ODP originate on buildings. Every geomask except for AVM displaced data points to either buildings or an uninhabited area as here, the park. Only one VM point had been transferred to a street. Contrary to that, AVM moved all data points to a street intersection decreasing the risk of false re-identification immensely. Figure 39b is depicting an area with family homes, a central park, and a lake with a displaced VM data point. Another DM data point had been moved on a house enhancing the risk of false re-identification. Figure 39c is portrays the aforementioned floodplain forest in Leipzig (Stadt Leipzig, 2020). It is noticeable that many points transferred by AAE are laying within this natural habitat which is certainly invalid for a displacement location. Finally, the last two figures 39d and 39e show rural areas within Saxony with a low point density. As in the previous examples, AVM points remain on street intersections decreasing the risk of false identification. Again, AAE, DM, and VM moved data points to uninhabited areas, i.e., a forest or grassland.

The generated output with the underlying base map has demonstrated that all applied obfuscating methods do not consider the surrounding environment of the study areas except for the new geomask AVM. Data points were moved by AAE, DM, and VM to lakes, forests, other residencies, or grassland which are either illogical locations or result in false re-identification. Contrary to that, AVM considers the infrastructure, and, hence, does not displace any data points to another residency circumventing the risk of false re-identification. Yet, an AVM data point on an intersection can still lead to assumptions about its origin, but the possibility of finding the right descent is a lot lower due to the accumulation of possible buildings around the intersection. Furthermore, the AVM data point might have been randomly placed to this part of the study area and might have its original location on the other side of the dissolved polygon.

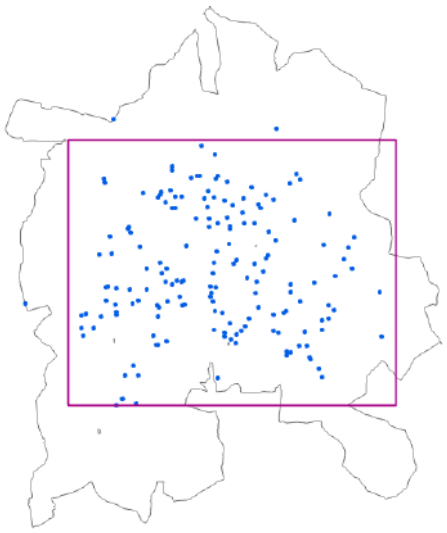
It can be inferred that regarding the visualized outcome, AVM succeeds the other geomasks. It outperforms all methods regarding the risk of false re-identification and scored the second strongest results regarding the preservation of the outer shape of the ODP. Just as observed with DM and VM, AVM demonstrates effective preservation of the point pattern. 34b is also a good examples for the preservation of the point pattern as well. AAE seems to abandon the pattern the most: Particularly figure 34b shows that the strongly visible floodplain forest (Stadt Leipzig, 2020) - that is intersecting the city of Leipzig from northwest to southwest - is not kept by AAE. Unlike that, AVM, DM, and VM maintain this meandering space in most regard (see figures 34c, 34d, and 34e). This is also noticeable in figure 36b: AAE does not sustain the several free spaces between the various point accumulations as presented by the ODP (figure 36a), AVM (figure 36c), DM (figure 36d), and VM (figure 36e).



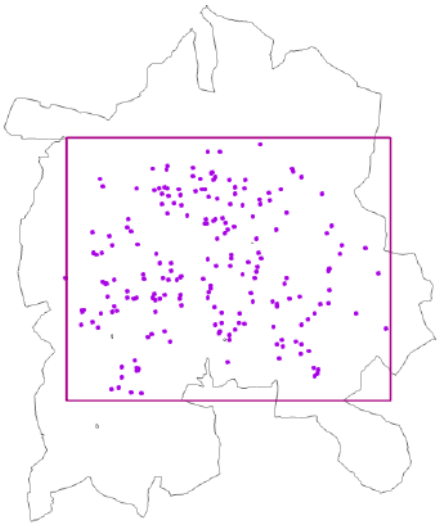
(a) ODP.



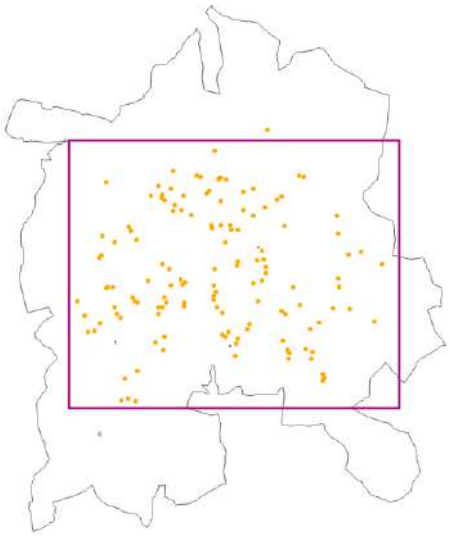
(b) AAE.



(c) AVM.

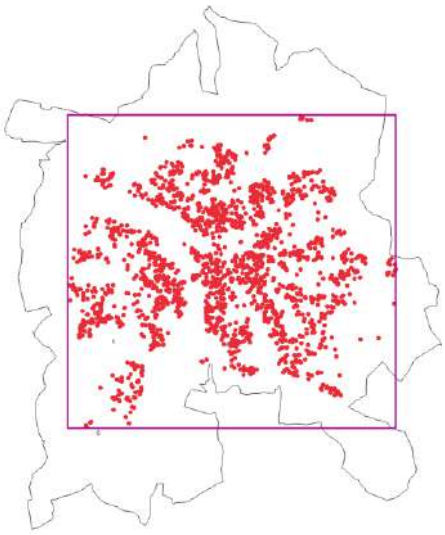


(d) DM.

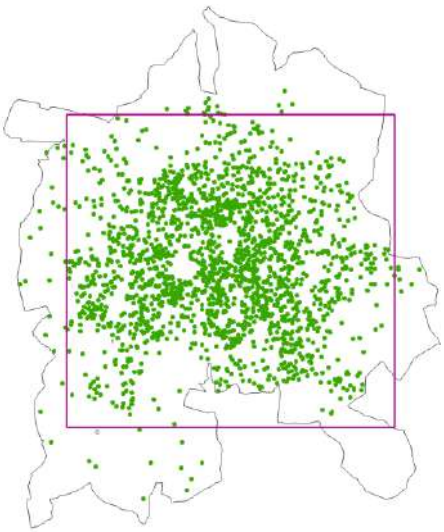


(e) VM.

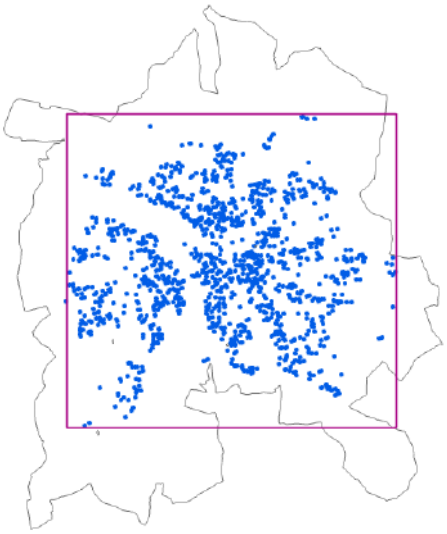
Figure 33: The results of the obfuscating techniques compared to the ODP in Leipzig (200 data points).



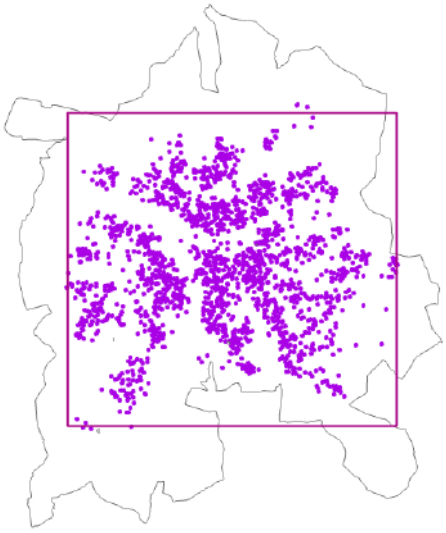
(a) ODP.



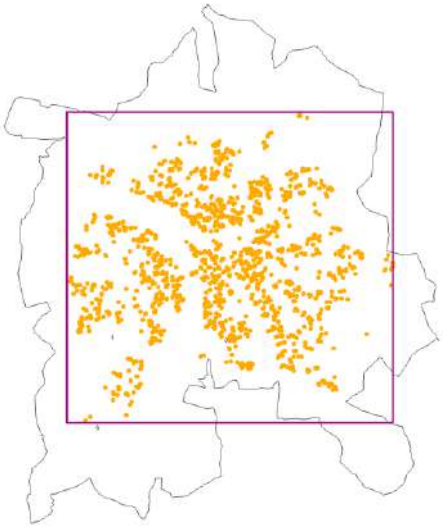
(b) AAE.



(c) AVM.

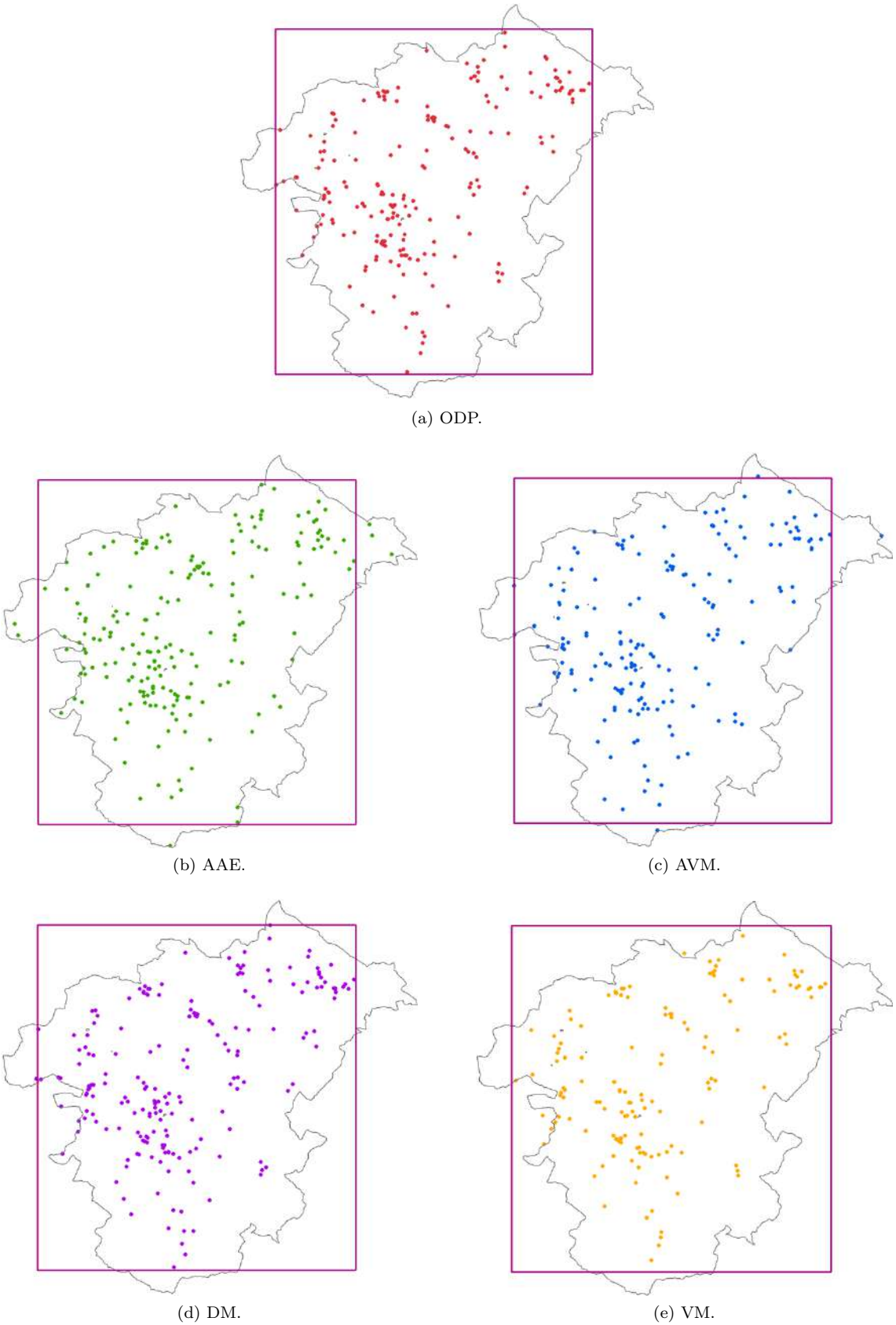


(d) DM.



(e) VM.

Figure 34: The results of the obfuscating techniques compared to the ODP in Leipzig (2000 data points).



(a) ODP.

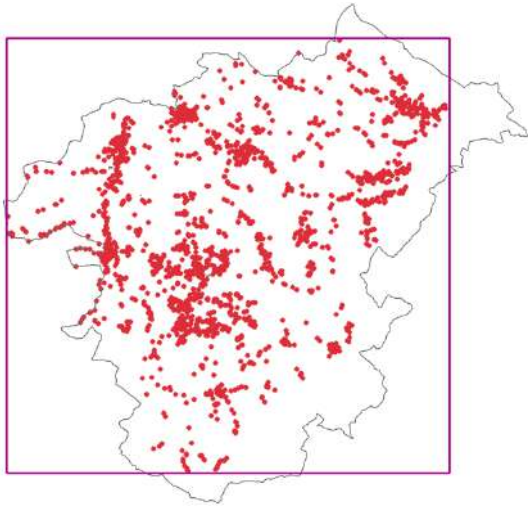
(b) AAE.

(c) AVM.

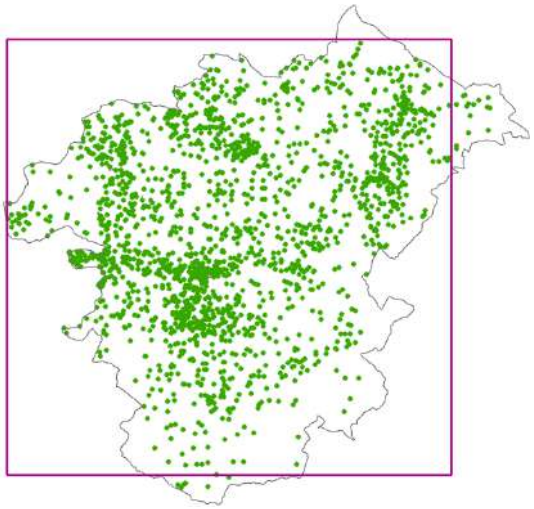
(d) DM.

(e) VM.

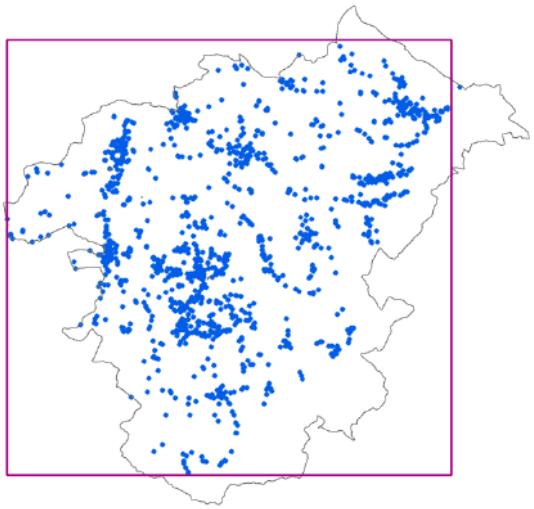
Figure 35: The results of the obfuscating techniques compared to the ODP in Zwickau (200 data points).



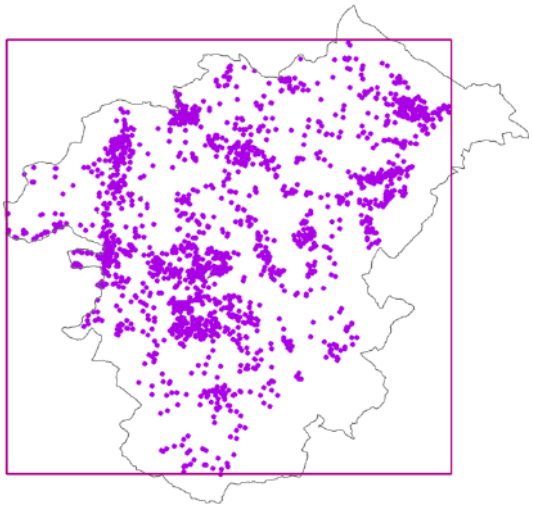
(a) ODP.



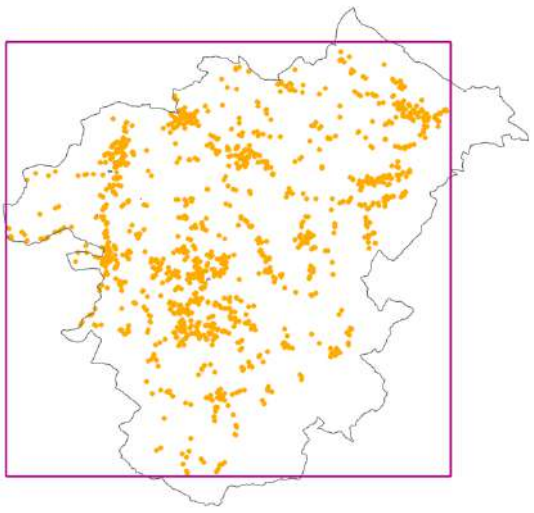
(b) AAE.



(c) AVM.

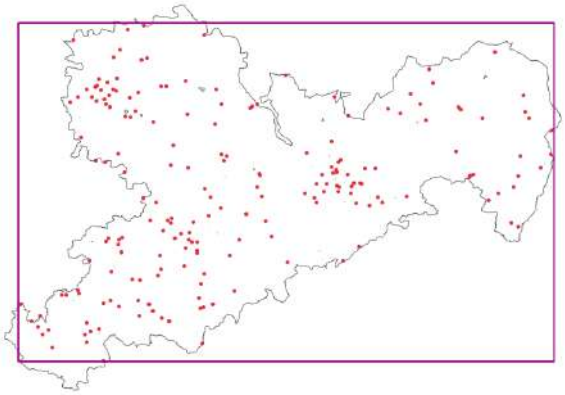


(d) DM.

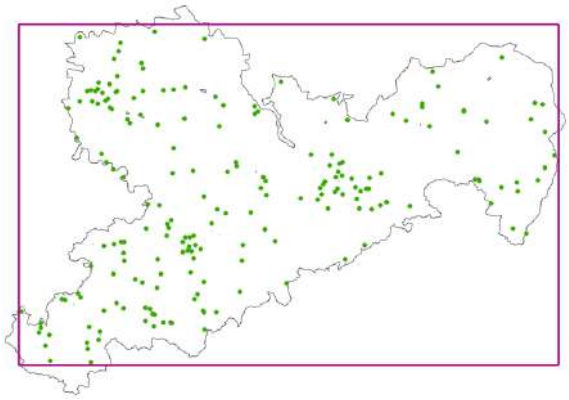


(e) VM.

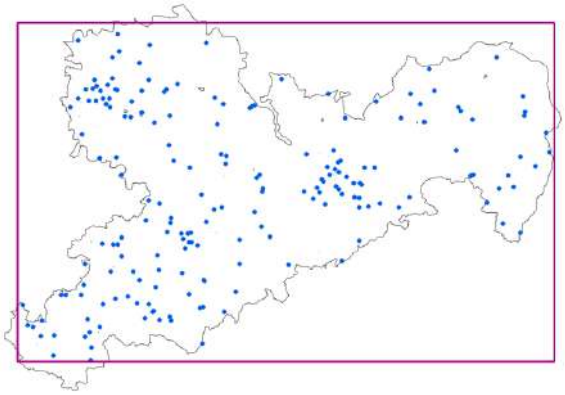
Figure 36: The results of the obfuscating techniques compared to the ODP in Zwickau (2000 data points).



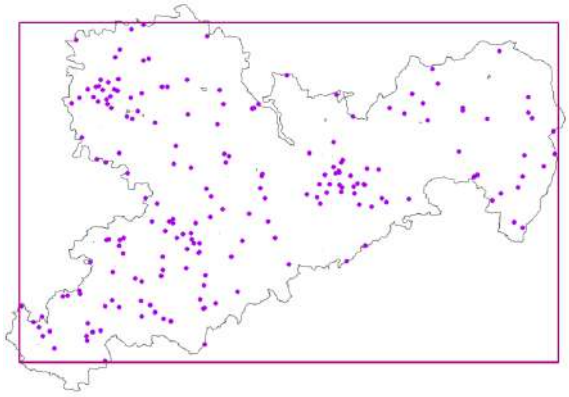
(a) ODP.



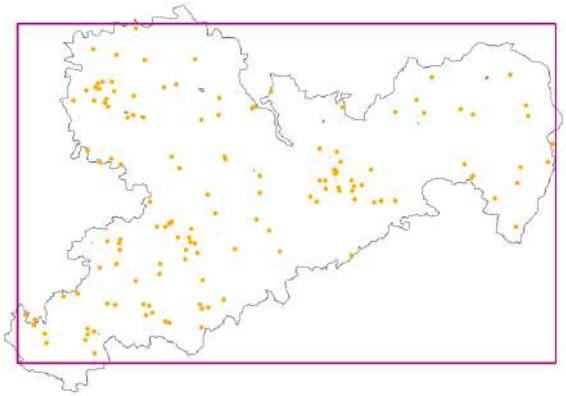
(b) AAE.



(c) AVM.

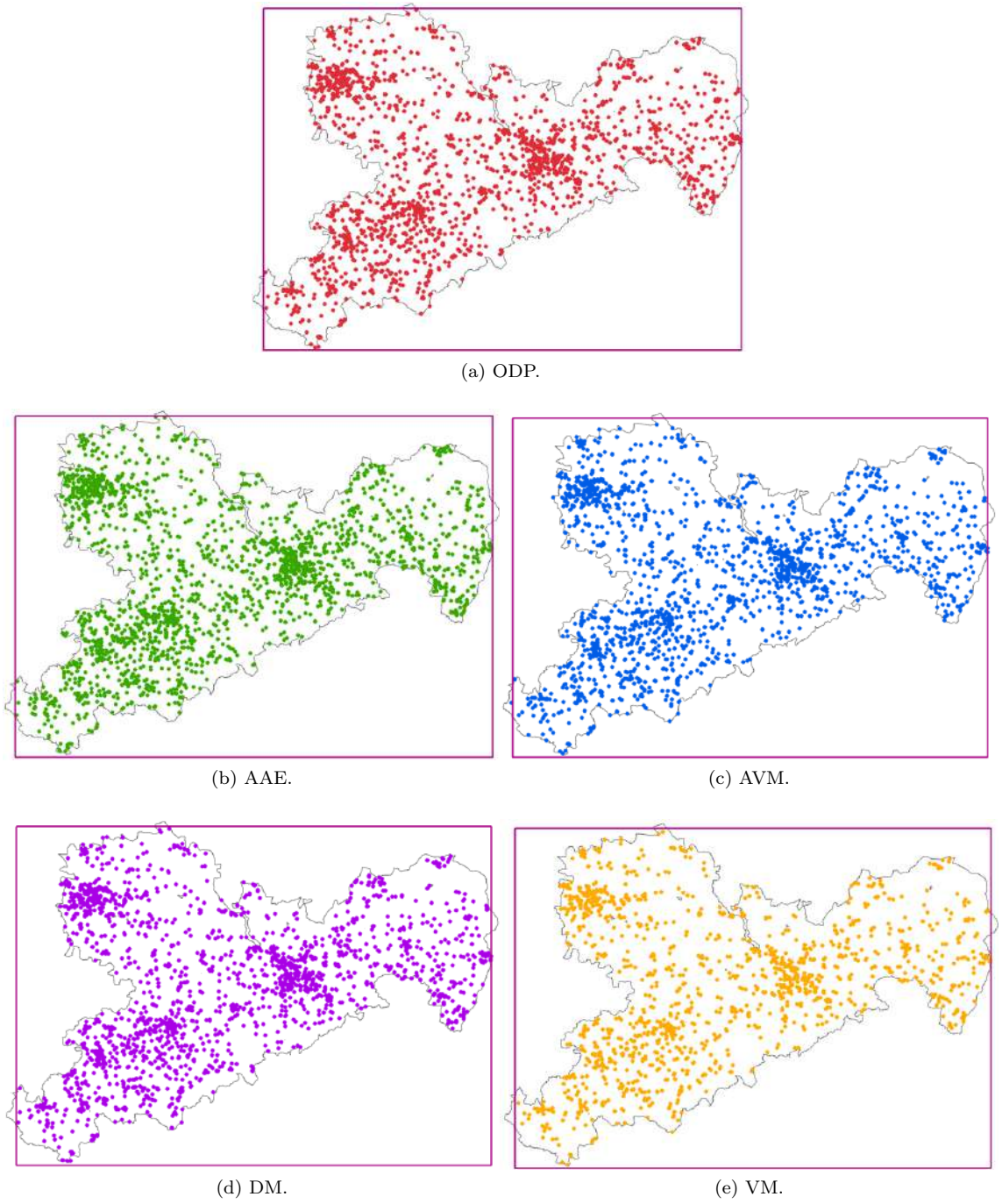


(d) DM.



(e) VM.

Figure 37: The results of the obfuscating techniques compared to the ODP in Saxony (200 data points).



(a) ODP.

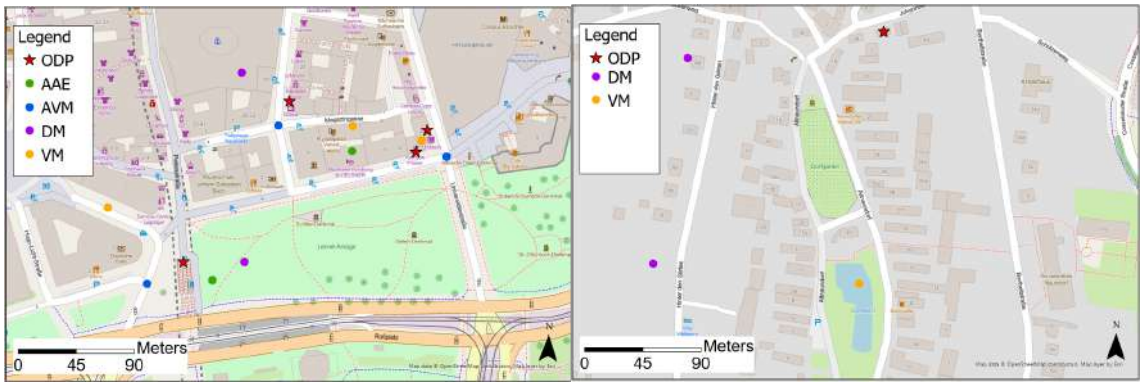
(b) AAE.

(c) AVM.

(d) DM.

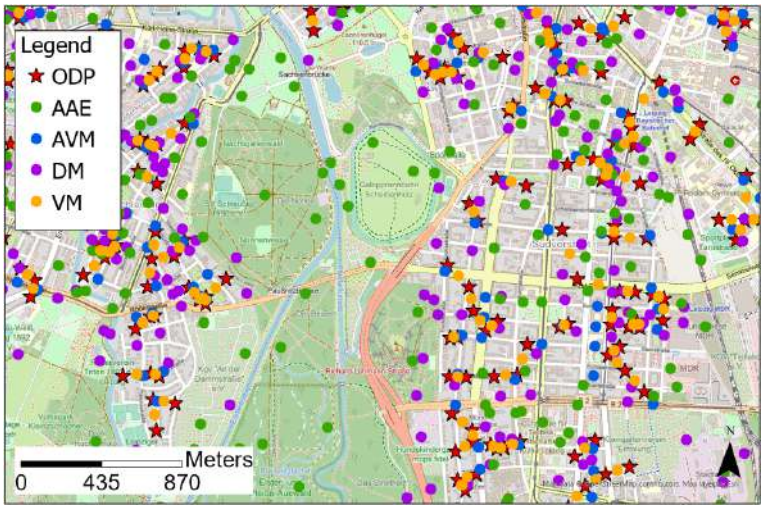
(e) VM.

Figure 38: The results of the obfuscating techniques compared to the ODP in Saxony (2000 data points).

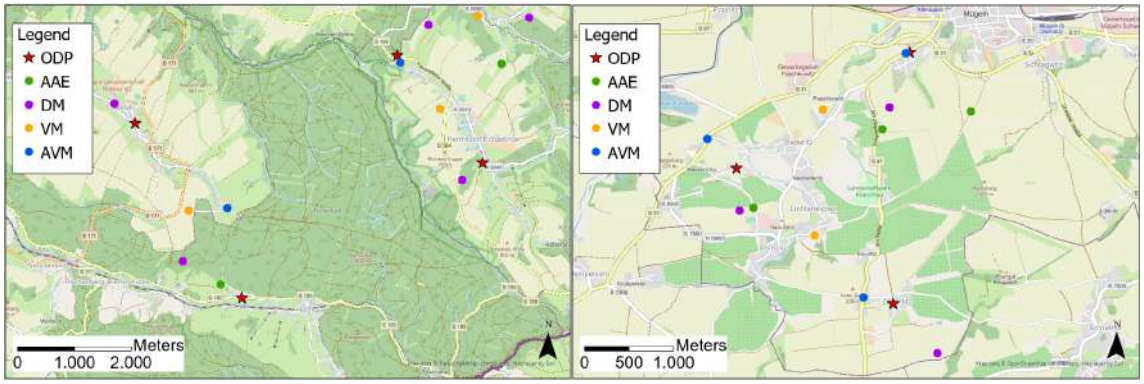


(a) ODP and MDP in the city centre of Leipzig.

(b) VM transferred on a lake.



(c) ODP and MDP around the floodplain forest in Leipzig.



(d) ODP and MDP in a rural area in Saxony.

(e) ODP and MDP in a rural area in Saxony.

Figure 39: The results of the obfuscating techniques compared to the ODP in various areas.

4.2 Mean and median centers

The results of the mean and median centers are depicted in the figures 40a-45b. The distances of displacement between the obscured and original mean and median centers are listed in table 2 and 3.

Regarding the mean centers, it can be noted that VM and DM outperform AAE and AVM in most obfuscation settings: DM has the smallest displacement distances for Leipzig 200 (12,71 m), Zwickau 200 (4,53 m) (see figure 42a), Zwickau 2000 (3,97 m) (as shown in figure 43a), and Saxony 200 (98,91 m) (refer to figure 44a). VM has the smallest displacement distances for Leipzig

2000 (1,18 m) (as depicted in figure 41a) and Saxony 2000 (9,61 m) (illustrated in figure 45a). The newly developed method AVM has AAE encompasses the furthest displacement distances for all data sets. For Leipzig 2000, the displacement distances between the MDP by AVM, DM, and VM vary between a maximum of two meters (refer to table 2). A small variation of displacement can also be seen for Leipzig 200, Zwickau 2000, and Saxony 2000. Unlike this, AAE demonstrates the greatest divergences towards the other displacement distances: AAE's displacement distance differs by up to 668,06 m to the best score for Saxony 200 (see table 2).

Mean center					
Area	**Data set**	**AAE**	**AVM**	**DM**	**VM**
Leipzig	200 data points	129,72 m	28,86	**12,71** m	12,72 m
	2000 data points	44,24 m	3,57 m	3,02 m	**1,18 m**
Zwickau	200 data points	177,18 m	85,71 m	**4,53 m**	23,48 m
	2000 data points	39,46 m	10,11 m	**3,97** m	6,14 m
Saxony	200 data points	766,97 m	616,41 m	**98,91** m	182,45 m
	2000 data points	39,83 m	27,46 m	14,39 m	**9,61** m

Table 2: The derived displaced distances (in meters) from the mean center of the ODP to the mean centers of the masked data points (the lowest displacement distance per data set is depicted in bold).

Regarding the median centers, VM surpasses the other geomasks: It represents a displacement distance of 1,23 m to the original median center of Leipzig 2000 (see figure 41b), 42,33 m to the original median center of Zwickau 200 (as seen in figure 42b), and 13,93 m to the original median center of Saxony 2000 (as illustrated in figure 45b). DM presents the lowest displacement distances for Leipzig 200 (1,63 m) (figure 40b) and Saxony 200 (90,82 m) (refer to figure 44b) while the new geomasking technique AVM has the smallest displacement for Zwickau 2000 (1,61 m) (as depicted in figure 43b). Again, AAE has much greater displacement distances than the other three geomasking techniques (except for Saxony 200) (refer table 3). Here, AVM performs poorly with 1528,40 m (see table 3).

Median center					
Area	**Data set**	**AAE**	**AVM**	**DM**	**VM**
Leipzig	200 data points	158,17 m	28,11 m	**1,63** m	40,30 m
	2000 data points	31,58 m	5,51 m	3,12 m	**1,23 m**
Zwickau	200 data points	562,56 m	55,22 m	100,68 m	**42,33 m**
	2000 data points	241,07 m	**1,61 m**	8,27 m	3,39 m
Saxony	200 data points	1357,64 m	1528,40 m	**90,82** m	231,68 m
	2000 data points	66,85 m	65,52 m	29,08 m	**13,93 m**

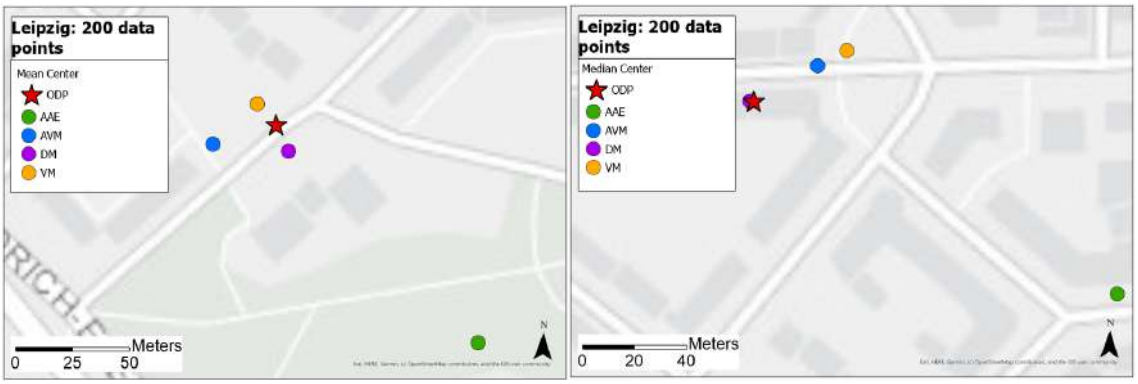
Table 3: The derived displaced distances (in meters) from the median center of the ODP to the median centers of the masked data points (the lowest displacement distance per data set is depicted in bold).

The values of the mean and median centers do not differ strongly for AVM, DM, and VM for Leipzig 2000 (less than two meters). Therefore, a strong discrepancy between the ODP and MDP can be highlighted (Seidl et al., 2015).

Finally, it can be seen that the lower the point density, the stronger the variation of the outcome between the geomasking techniques: For instance, the 200 data points in Saxony represent the strongest variations between the geomasking techniques and the highest displacement distances between the masked and original mean and median centers. Contrary to that, Leipzig has the highest population density with 2000 data points and demonstrates the smallest displacement distances and lowest variations between the mean and median centers of the MDP and ODP (see 2, 3 and figures 41a and 41b).

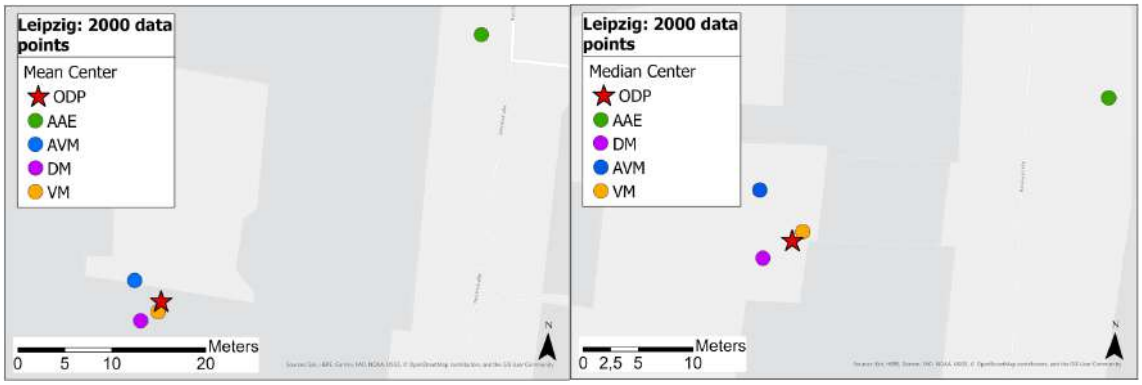
Due to the previous examinations, it can be concluded that DM fares the best in terms of the mean center. Concerning the median center, VM derived the best results. The new method AVM succeeds the other methods only for one data set (median center for Zwickau 2000). This leads to the assumption that AVM is an effective method for data sets with a high population density. However, this can also be concluded for DM and VM which reach even smaller displacement

distances and, above all, attain these for almost all data sets. Unlike this, AAE performed the worst for almost all data sets except for the median center of Saxony 200. It is noteworthy though, that AAE derives fewer displacement distances for lower population densities than for high densities. The outcome of the various mean and median centers can be justified by the fact that AAE completely and AVM partly disperse ODP whilst VM and DM transfer data points based on their original location.



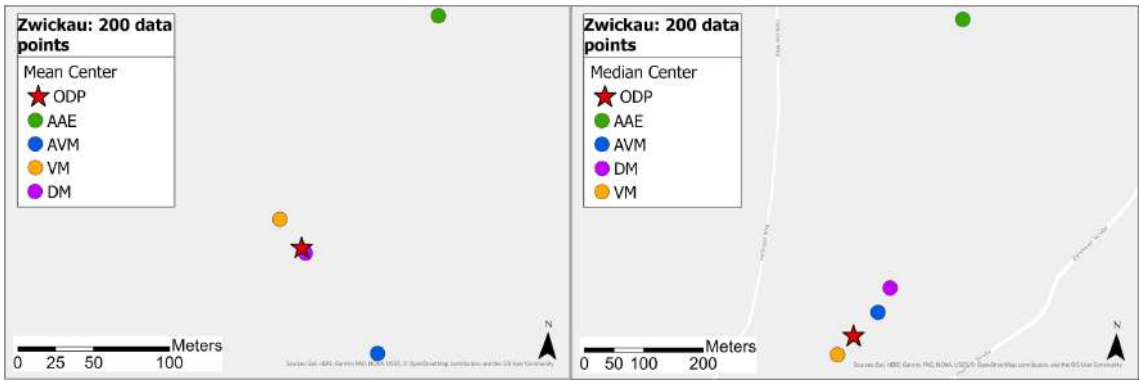
(a) Leipzig: 200 data points. (b) Leipzig: 200 data points.

Figure 40: The mean (a) and median (b) centers of the 200 masked and original data points in Leipzig.



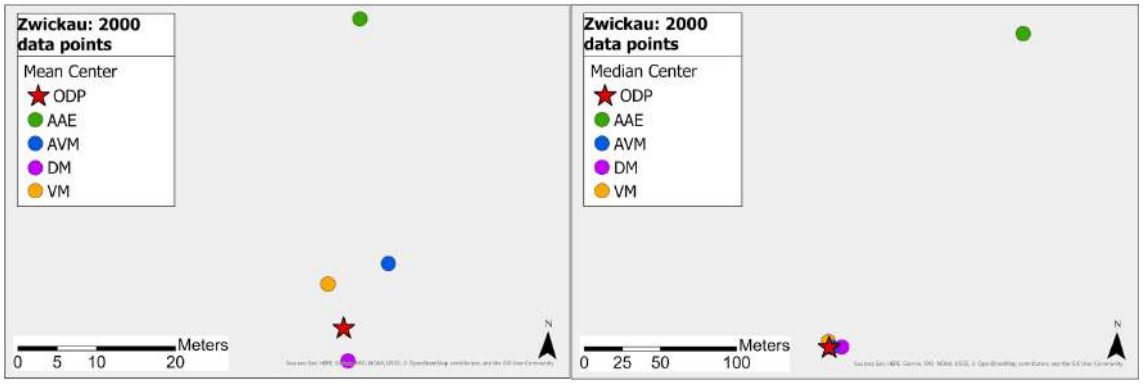
(a) Leipzig: 2000 data points. (b) Leipzig: 2000 data points.

Figure 41: The mean and median centers of the 2000 masked and original data points in Leipzig.



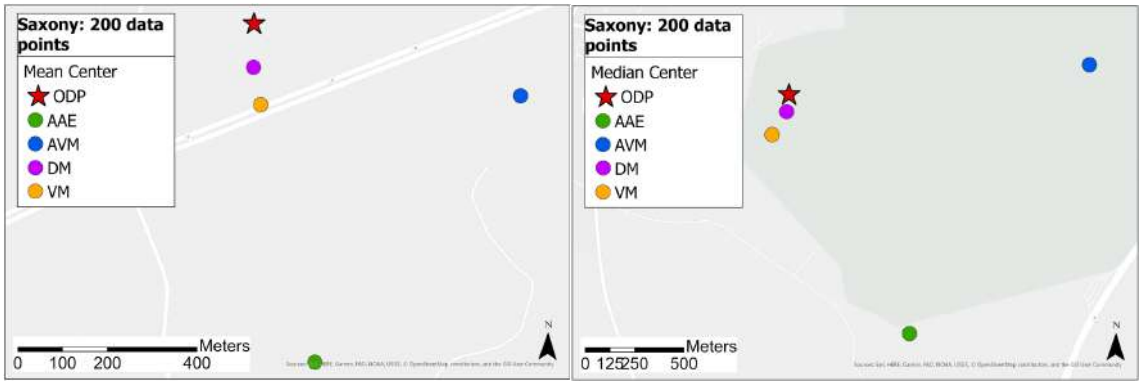
(a) Zwickau: 200 data points (b) Zwickau: 200 data points

Figure 42: The mean and median centers of the 200 masked and original data points in Zwickau.



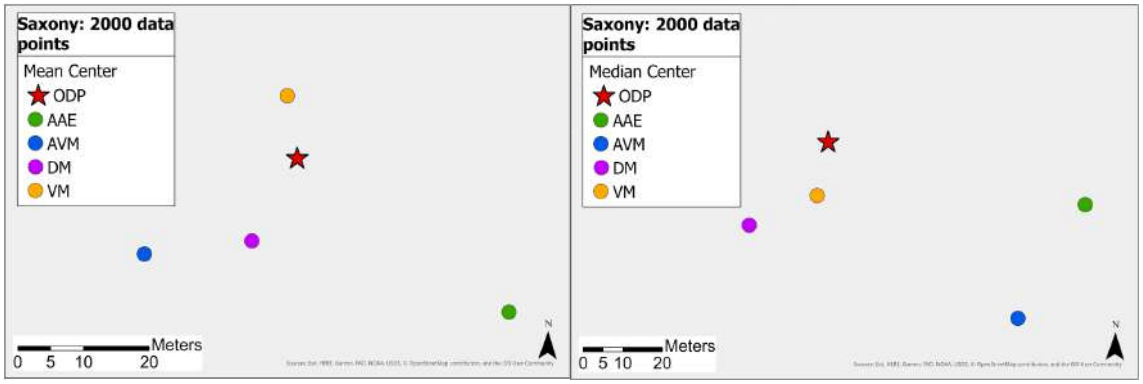
(a) Zwickau: 2000 data points (b) Zwickau: 2000 data points

Figure 43: The mean and median centers of the 2000 masked and original data points in Zwickau.



(a) Saxony: 200 data points (b) Saxony: 200 data points

Figure 44: The mean and median centers of the 200 masked and original data points in Saxony.



(a) Saxony: 2000 data points (b) Saxony: 2000 data points

Figure 45: The mean and median centers of the 2000 masked and original data points in Saxony.

4.3 Ripley's K-function

To achieve a concise overview, the results of the expK-values and obsK-values of the ODP and MDP of every five bands were summed up and, hence, allow a straightforward comparison in the histograms below (see figures 46-51). The entire output of the Ripley's K-function has been attached within a supplementary file (refer to appendix D.2.4).

Figure 46 depicts the Ripley's K-function for the ODP and MPD of Leipzig 200. It is observable that the ODP and MPD attain higher obsK-values than the expK-value. This interprets that their point distribution is more clustered than it would be in a random distribution (refer to chapter 2.4.3). It is visible, that across all bands, AVM maintains a higher obsK-value than the ODP representing a stronger point aggregation than the original data set whereas AAE and DM are less clustered as can be seen by their lower obsK-value. AVM demonstrates higher obsK-values for the first ten bands indicating a more clustered point pattern and slightly lower obsK-values for the last five bands describing a more disperse pattern (see figure 46). It is of great significance that DM has the most similar values to those of the ODP throughout all distances while AVM has very similar ones for all bands as well. Ergo, both methods describe a strong linkage to the ODP. On the contrary case, AAE demonstrates much lower obsK-values throughout all distances concluding that AAE's data points are scattered around the ODP and do not remain linked.

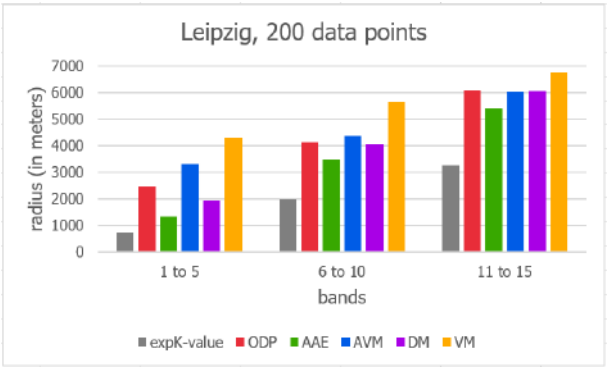


Figure 46: Ripley's K-function graph depicting the obsK-values of the ODP and MDP compared with the expK-value from 99 simulations for Leipzig 200 data points.

Similar observations can be made for the Ripley's K-function for Leipzig 2000 (refer to figure 47): Again, AAE has much lower obsK-values (bands 1-5: ca. 1258,77 m, bands 6-10: ca. 3294,10, and bands 11-15: ca. 5208,80) than the ODP (ca. 1801,64 m, ca. 4152,11 m, and ca. 6186,17 m; refer to appendix D.2.4) appointing a strong scattering and an absence of point linkage between AAE and ODP. As noticed in the prior histogram, DM increases the linkage of the point patterns as well as the point clustering with a growing distance compared to the ODP. The output of VM maintains a higher aggregation than the ODP for the first five bands and the last five bands. The bands 6-10 prove a deficiency in point linkage as well as a point dispersion. Finally, the new

obfuscating method AVM exceeds all other techniques for this data set: Throughout all distances, the data points remain highly linked and are only a bit more accumulated than the ODP, presenting very similar osbK-values as the ODP.

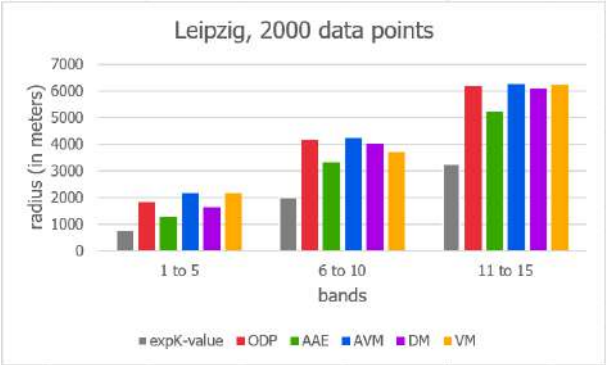


Figure 47: Ripley's K-function graph depicting the obsK-values of the ODP and MDP compared with the expK-value from 99 simulations for Leipzig 2000 data points.

The next two histograms describe the Ripley's K-function for the district of Zwickau (see figures 48 and 49). For Zwickau 200 (refer to figure 48), the mask AAE does not maintain the point linkage and scatters its masked points more than the ODP. DM has also smaller obsK-values than the ODP, too, but the values (bands 1-5: ca. 1807 m, bands 6-10: ca. 4333 m, and bands 11-15: ca. 6714 m (refer to appendix D.2.4)) remain more similar to those of the ODP than the AAE (bands 1-5: ca. 2177 m, bands 6-10: ca. 4776 m, and bands 11-15: 6986 m (see appendix D.2.4)). Still, DM demonstrates a point dispersion and a low point linkage to the ODP as well. VM and AVM result in higher osbK-values throughout all distances describing a stronger aggregation than the ODP. Lastly, the obsK-values of AVM are again the closest to those of the ODP during all bands demonstrating a great point linkage.

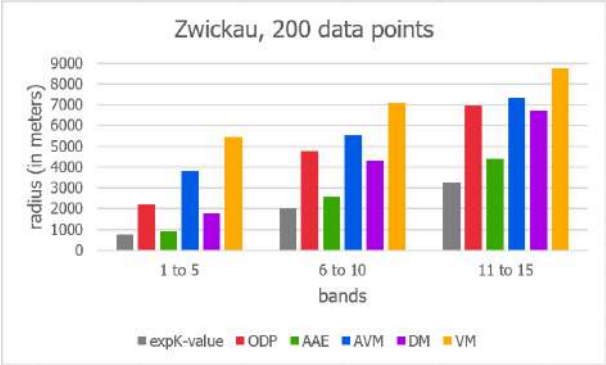


Figure 48: Ripley's K-function graph depicting the obsK-values of the ODP and MDP compared with the expK-value from 99 simulations for Zwickau 200 data points.

For Zwickau 2000, the same remarks can be applied: AAE and DM have more scattered while AVM and VM have more clustered data points than the ODP. The latter also demonstrates very similar obsK-values as well as a persistent point linkage to the ODP.

In terms of Saxony 200, VM shows very different obsK-values than the ODP as well as a much stronger point accumulation. This characteristic of VM was explained in chapter 2.3.2, where VM tends to change the point pattern within areas with scattered domiciles. Unlike this, all other geomasks have very similar obsK-values to those of the ODP: AAE demonstrates a dispersed point distribution for the first five bands but, contrary to the prior observed histograms, receives a clustered point pattern for the last ten bands. Reversely, AVM and DM have an aggregated pattern for the first five bands while depicting a scattered pattern for the last ten bands (see figure 49). For this data set, DM has the most similar values as the ODP (refer to appendix D.2.4).

As examined for Zwickau 200 and Zwickau 2000, AAE and DM represented a more scattered while AVM and VM have a more clustered point pattern than the ODP again. AVM has the most

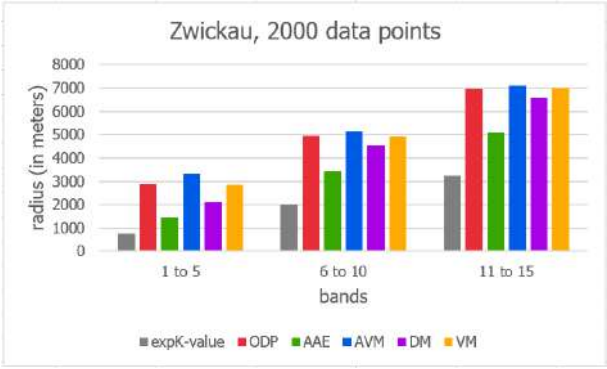


Figure 49: Ripley's K-function graph depicting the obsK-values of the ODP and MDP compared with the expK-value from 99 simulations for Zwickau 2000 data points.

similar obsK-values for the first five bands whereas DM represents the more alike values for the last ten bands. Both methods show a great point linkage to the ODP for all distances.

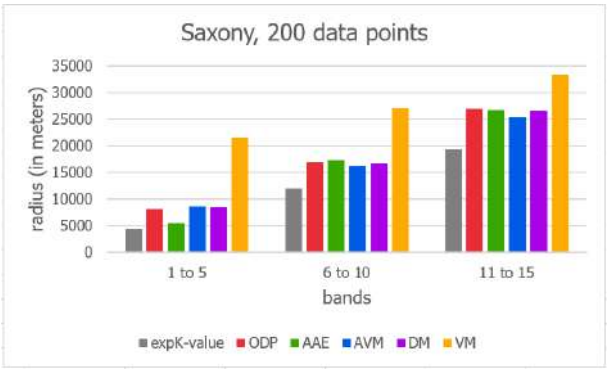


Figure 50: Ripley's K-function graph depicting the obsK-values of the ODP and MDP compared with the expK-value from 99 simulations for Saxony 200 data points.

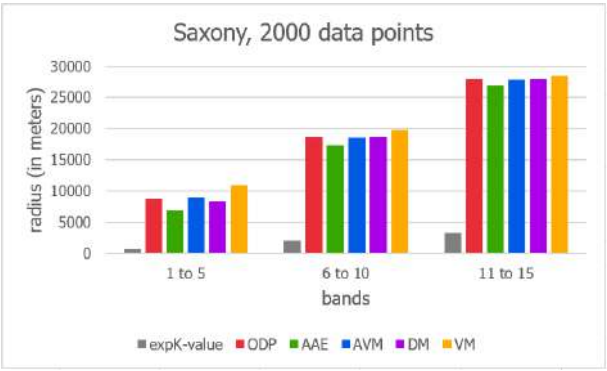


Figure 51: Ripley's K-function graph depicting the obsK-values of the ODP and MDP compared with the expK-value from 99 simulations for Saxony 2000 data points.

The Ripley's K-function was implemented to evaluate whether the obfuscated points are more aggregated or dispersed than the ODP and whether the MDP retain similar obsK-values like the ODP. From the histograms depicting the output of this statistical analysis, it can be concluded that the obfuscating method DM and AVM significantly outperform the other techniques:

DM demonstrates the most similar obsK-values for Leipzig 200 (all bands), Saxony 200 (all bands), Saxony 2000 (bands six to 15), and Zwickau 200 (all bands). Due to the similar values, an efficient point linkage to the ODP can be concluded as well as stable maintenance of the original point pattern. Still, this method has more scattered point patterns than the ODP's for all data sets

excluding Saxony 200 (bands one to five) and Saxony 2000 (bands 11-15). The mask's tendency to disperse data points as well as the similar obsK-values can be explained by its property to randomly transfer data points within a buffer zone around the original location (refer to chapter 2.3).

Similar observations can be made for AVM: Within all data sets, AVM retains similar obsK-values as the ODP's exemplifying a strong point linkage throughout all distances. It even scores the most alike obsK-values for Leipzig 2000 (bands one to ten), Saxony 2000 (bands one to five), and Zwickau 2000 (all bands). Additionally, AVM has a greater point aggregation than the ODP in almost all cases (exempted are Saxony 200 and Saxony 2000) which can be justified by the displacement of data points to the same street intersection. Since Saxony 200 and Saxony 2000 represent a low point density due to the great size of the study area, the by AVM masked point patterns of these two data sets are more scattered than in previous observations. The results confide in the new masking technique that the masked data points maintain a spatial dependency on the ODP (Seidl et al., 2015) and retain the point pattern in areas with a low *and* high point density.

The next technique, VM, displays a bigger point accumulation than the ODP for all data sets with a low point density (Leipzig 200, Zwickau 200, and Saxony 200). Here, it not only demonstrates the highest but also most dissimilar obsK-values than the ODP which can be elucidated by the technique's character to alter the point pattern in scattered areas due to greater displacement distances based on the established Voronoi polygon. Contrary to that, VM reaches very similar obsK-values to those of the ODP for the data sets with a higher point density (i.e., Leipzig 2000 (first and last five bands), Zwickau 2000 (all bands), and Saxony 2000 (bands six to 15)) and even the most similar one for Leipzig 2000 (bands 11-15) which can be explained due to the small point displacement resulting in a very similar point pattern to that of the ODP (as elaborated in chapter 2.3.2). Hence, it can be concluded that VM retains a similar point pattern for data sets *with a high population density* - in this case, for Leipzig 2000, Zwickau 2000, and Saxony 2000 - but not for low-density point data sets.

Last but not least, AAE fares the weakest in this statistical method: Throughout all distances and data sets, the AAE points are dispersed and do not resemble a point linkage to the ODP. This is due to the fact that AAE has the opportunity to transfer data points within an entire area presenting a merged polygon, abandoning the point linkage to the ODP.

Finally, it must be noted that the Ripley's K-function is perhaps not an excellent choice to highlight "finer-scale difference[s]" (Seidl et al., 2015, p. 260).

4.4 Nearest Neighbor Hierarchical Cluster Analysis

The derived output of the NNHCA for each ODP and MDP are listed in the tables 4-7 comparing the number of clusters found, the mean points per cluster, and mean cluster density per m^2 (Seidl et al., 2015). The complete output of the NNHCA is found in the appendix D.2.3. The resulted ellipsoids of the first-order NNh are depicted in figures 52-55 and are examined for the size, location, and orientation of the ODP and MDP clusters.

For Leipzig 200, the ODP produced six clusters, implying an averagely 7,33 points per cluster. Regarding the amount of generated clusters, AAE, AVM, and DM are the nearest of the derived ODP-value with seven clusters. AVM has the closest number of mean points per cluster at 7,50; AAE the most dissimilar (6,43). VM created the most ellipsoids at nine. Regarding the last parameter, the mean cluster density, VM has the highest cluster density whereas AVM has the most similar one (refer to table 4).

Leipzig 200			
Geomask	**Clusters found**	**Mean Points**	**Mean cluster density (per m^2)**
ODP	6	7,33	0,0000130
AAE	7	6,43	0,0000169
AVM	7	7,50	0,0000113
DM	7	6,86	0,0000207
VM	9	6,67	0,0000221

Table 4: Output for NNHCA of Leipzig 200.

Figure 52 illustrates that the clusters of DM align most successfully with the shape and location of the cluster ellipses of the ODP. Only one DM outlier can be depicted in the north of Leipzig. VM depicts two absent clusters in the east of the study area as well in the center. The new method AVM has one outlier in the north of Leipzig sharing this specific location with DM and AAE. AAE fulfills the alignment and matching of the ODP the least as three outliers are located in the north, southeast, and west of the study area.

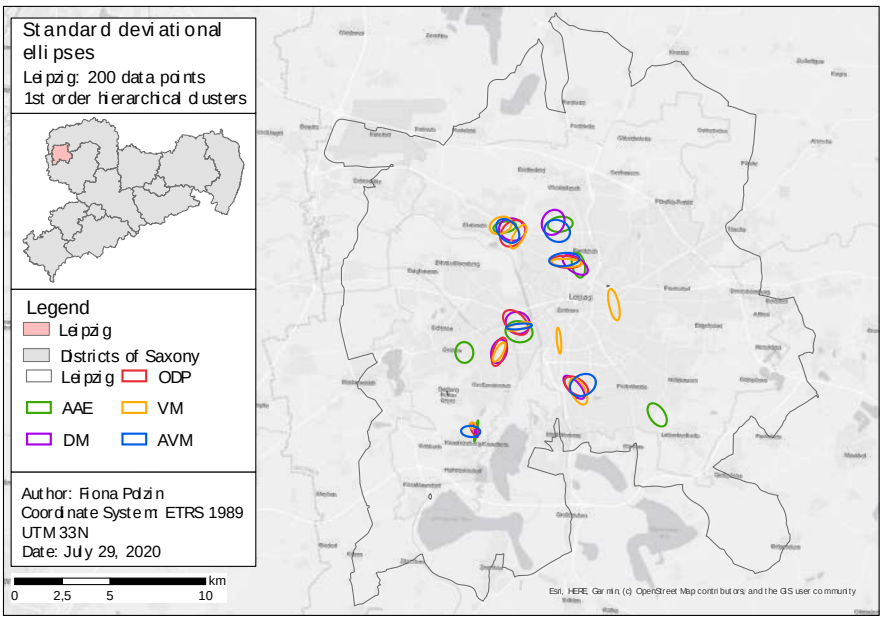


Figure 52: First hierarchical clusters for Leipzig 200.

For Leipzig 2000, VM almost receives the same number of clusters as the ODP (VM: 139; ODP: 136). AVM is second at 140 while AAE accomplishes only 57 and, hence, demonstrates the lowest performance. Regarding the mean points, ODP contains 7,73 points; DM reaches the nearest measure at 7,83. Again, AAE has the most dissimilar value at 7,02. Finally, DM has the most alike mean cluster density at 0,0000979 m^2 (ODP: 0,0001144 m^2). Contrary to the prior described comparisons, AAE accomplishes the second closest rate at 0,0000835 m^2 while AVM has the least comparable one at 0,0008718 m^2 (see table 5).

Leipzig 2000			
Geomask	**Clusters found**	**Mean Points**	**Mean cluster density (per m^2)**
ODP	136	7,73	0,0001144
AAE	57	7,02	0,0000835
AVM	140	8,08	0,0008718
DM	109	7,83	0,0000979
VM	139	8,02	0,0001964

Table 5: Output for NNHCA of Leipzig 2000.

Figure 53 exemplifies that DM has two cluster ellipsoids that are not situated with those of the ODP. VM represents four that are not matching whereas AVM contains nine and AAE 21. Significantly, most of these outliers by location are not positioned in the center of the study area. Only AAE has a tendency to place clusters closer to the center.

About Zwickau 200, the ODP have found eight clusters, a mean point at 7,75, and a density of 0.0000055 per m^2. With respect to the first parameter, AVM has reached the closest value. In terms of mean points, VM outperforms the other masking methods at 7,82 whilst the mean cluster density, DM shows the nearest value at 0,0000060 per m^2 (see table 6).

The next map shows the cluster ellipsoids for Zwickau 200 (refer to figure 54). VM illustrates one outlier in the west and one north of the city of Zwickau (situated centrally within the study

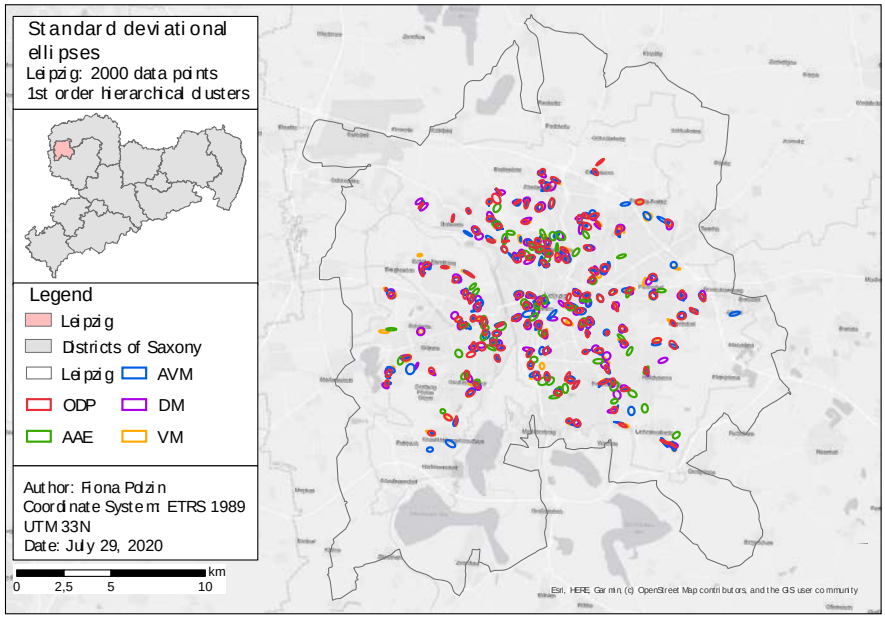


Figure 53: First hierarchical clusters for Leipzig 2000.

Zwickau 200			
Geomask	**Clusters found**	**Mean Points**	**Mean cluster density (per m^2)**
ODP	8	7,75	0,0000055
AAE	5	6,60	0,0000028
AVM	9	7,44	0,0000177
DM	8	7,63	0,0000060
VM	11	7,82	0,000077

Table 6: Output for NNHCA of Zwickau 200.

area). Within the same region is also an AVM outlier located. In the northeast, two VM and two AVM clusters gather around one ODP cluster. AAE does not have outliers but did not build any clusters at three locations with an ODP ellipsoid which corresponds with the previous table (see table 6). DM does not have any outliers neither. Regarding the size and orientation of the ellipsoids, AAE clusters do not correlate with any of the ODP clusters. Three AVM clusters approximate well with ODP clusters particularly in the center around the city of Zwickau. DM represents four clusters that align perfectly with ODP ellipsoids situated around the city of Zwickau whereas VM has the only cluster that matches up acceptably with one ODP ellipsoid.

Finally, for Zwickau 2000, the ODP contain 110 clusters, mean points of 8,70, and a mean cluster density of 0,0000655 per m^2. VM has 111 clusters followed by AVM with 112 outperforming the other obfuscating techniques at 89 (DM) and 44 (AAE). The closest mean point value was obtained by DM at 8,66 while AAE reaches the most similar value for mean cluster density at 0.0000511 (per m^2) (refer to table 7).

The corresponding map (see figure 55) shows that no outliers exist. Five VM clusters approximate fully with ODP clusters. However, north of the city of Zwickau two VM ellipsoids aggregate around one ODP cluster. Three AVM clusters match wholly with ODP clusters in the center of the study area whereas. AAE has only three ellipsoids that only intersect the ODP clusters but are very dissimilar due to their orientation and size.

In the next comparison - Saxony 200 -, four clusters were generated by the ODP with mean points at 6,5 and a mean cluster density of zero. DM succeeded the same values as ODP whereas VM indicates the most distinct values: it has found six clusters, has a mean point at 6,7, and a cluster density of 0,000002 per m^2 (see table 8).

Figure 56 exhibits that the cluster ellipsoids of VM and AVM have one outlier in the southwest of Saxony and do not have ellipsoids in the east of the study area (around the city of Chemnitz)

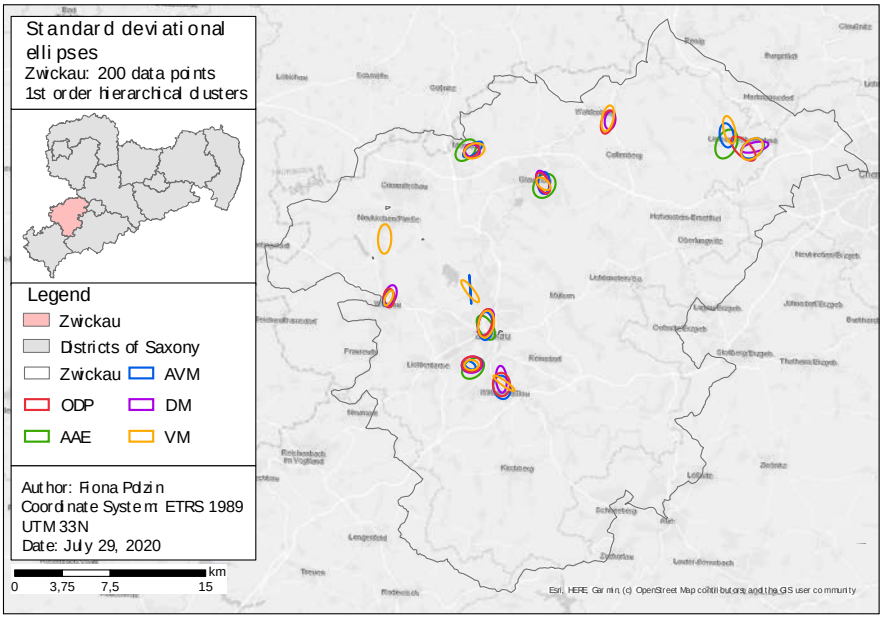


Figure 54: First hierarchical clusters for Zwickau 200.

Zwickau 2000			
Geomask	**Clusters found**	**Mean Points**	**Mean cluster density (per m²)**
ODP	110	8,70	0,0000655
AAE	44	8,25	0,0000511
AVM	112	8,98	0,0004652
DM	89	8,66	0,0000598
VM	111	8,94	0,00010002

Table 7: Output for NNHCA of Zwickau 2000.

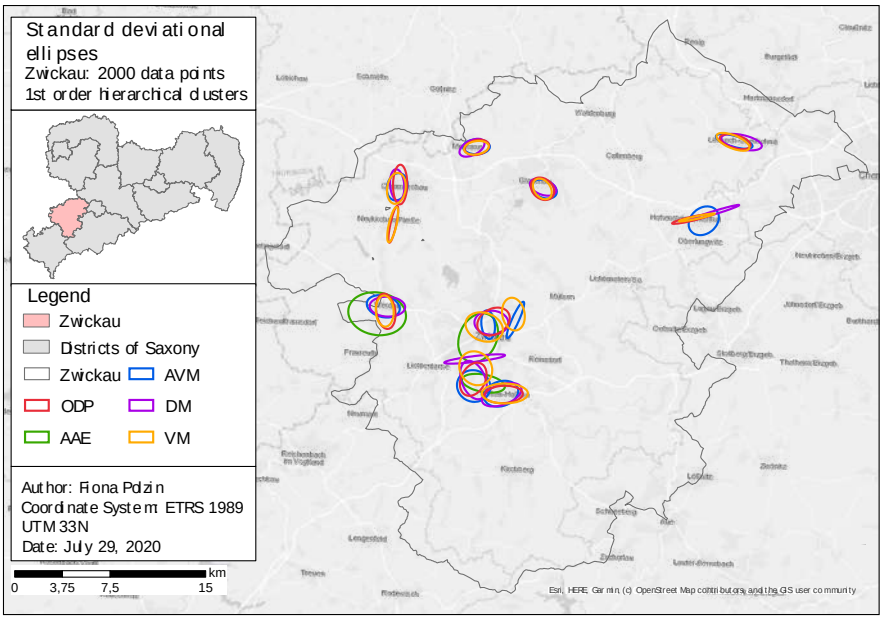


Figure 55: First hierarchical clusters for Zwickau 2000.

Saxony 200			
Geomask	**Clusters found**	**Mean Points**	**Mean cluster density (per m²)**
ODP	4	6,5	0
AAE	3	7	0
AVM	5	6,4	0
DM	4	6,5	0
VM	6	6,7	0,0000002

Table 8: Output for NNHCA of Saxony 200.

where ODP does. It is noteworthy that both techniques developed two clusters at Leipzig in the northwest of the study area where only one ODP ellipsoid is located. With the size and orientation of the cluster ellipsoids, the clusters of DM align thoroughly in Leipzig and Chemnitz while one VM cluster approximates the ODP cluster in the west of the study area (state capital Dresden). The ellipsoids of AVM do not match the orientation or size of the ODP clusters neither in Leipzig nor around Dresden. The AAE ellipsoids in Leipzig and Chemnitz as well as in Dresden do not match perfectly with those of the ODP but are more similar than the AVM clusters. Regarding this data set, it can be concluded that DM fares the best in terms of all analyzed parameters.

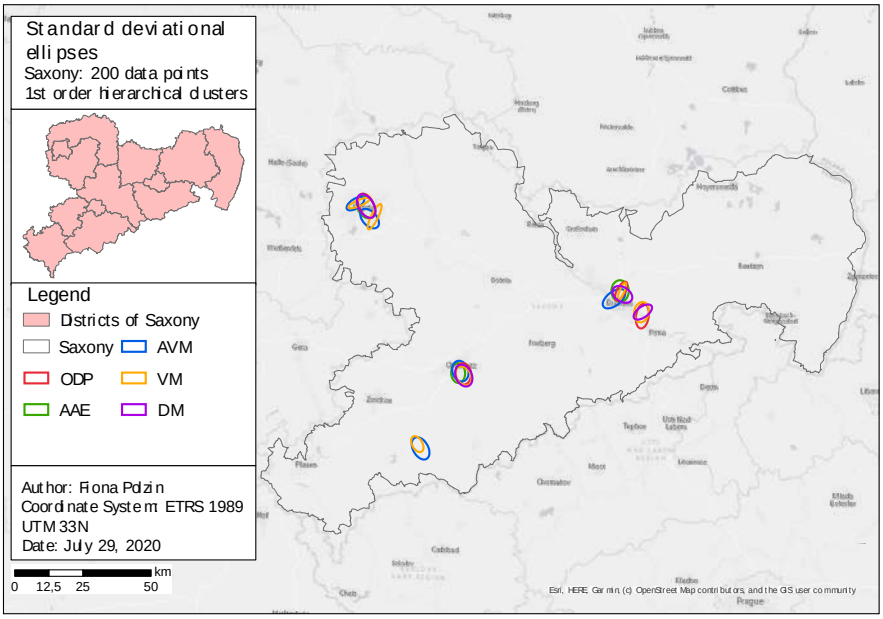


Figure 56: First hierarchical clusters for Saxony 200.

For Saxony 2000, the ODP demonstrates 66 clusters, mean points at 7,86, and a mean cluster density at 0,0000042 per m^2. Regarding the first parameter, AVM outperforms the other methods at 63 clusters. Concerning the second parameter, DM reaches the closest mean points at 7,80. Finally, the most alike mean cluster density has been obtained by AVM at 0,0000044 per m^2. AAE fares the worst with regard to the number of clusters (36) and the mean cluster density (0,0000030 per m^2) while VM demonstrates the least efficiency for mean points (8,07) (refer to table 9).

Figure 57 highlights eleven outliers each for DM and VM. Unlike in the previously studied maps, AVM and AAE have fewer outliers at seven (AVM) and ten (AAE). Yet, AVM and DM have many cluster ellipsoids that are extremely similar in terms of orientation and size while AAE performs the worst since it has 30 clusters less than the ODP.

The observations from the previous examinations can be summarized as followed: Regarding the number of clusters found, no obfuscating method encompasses the same number of clusters as those of the ODP for all data sets. However, DM scores the same amount for Saxony 200 and Zwickau 200 and is, ergo, the only method that reached the same number of found clusters for two data samples. The new geomasking technique AVM succeeds similar values as those of the ODP:

Saxony 2000			
Geomask	**Clusters found**	**Mean Points**	**Mean cluster density (per m²)**
ODP	66	7,86	0,0000042
AAE	36	7,58	0,0000030
AVM	63	7,68	0,0000044
DM	61	7,80	0,0000037
VM	71	8,07	0,0000055

Table 9: Output for NNHCA of Saxony 2000.

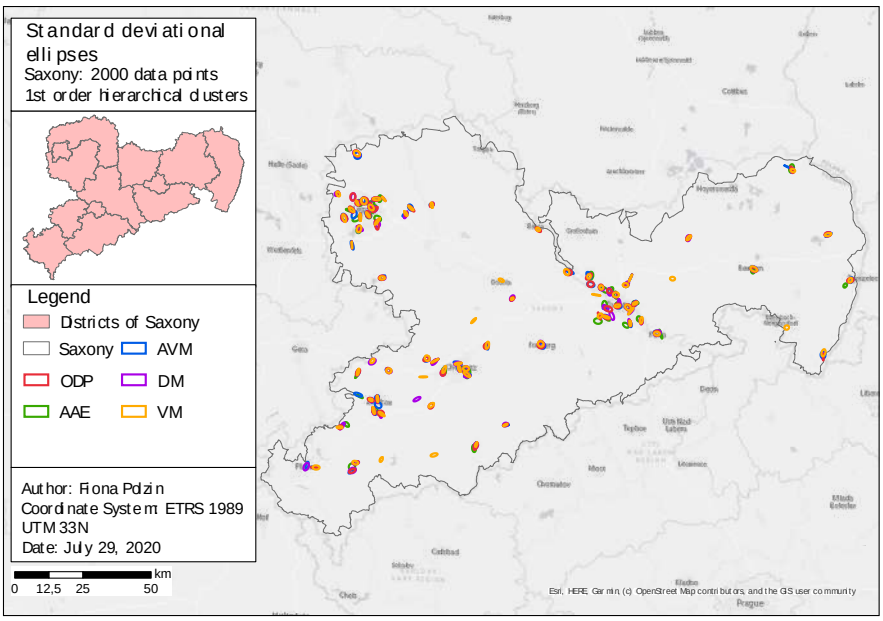


Figure 57: First hierarchical clusters for Saxony 2000.

For instance, for Leipzig 200, AVM receives one cluster more (see table 4) and for Zwickau 2000 only two more (refer to table 7). VM reaches a close number to that of the ODP regarding data sets of 2000 data points (refer to tables 5, 9, and 7). Therefore, it can be summarized that DM has the best performance regarding the generating of clusters while VM has the most polarising one depending on the size of the data set. This is plausible since VM tends to move data points to the same location: In case of greater point data sets, more points are located to the same location.

Speaking of location: the figures 52-55 evinced that DM has the strongest efficiency in terms of orientation and size followed by VM which also establishes various cluster ellipsoids which are similarly shaped and orientated. AAE has clearly the weakest ability since many of these clusters varied in size and orientation.

With regard to the mean points, again, no geomask reaches the same values as the ones of the ODP for all samples. Here, an observation of varying masks regarding the best performance can be obtained: DM demonstrated the best efficiency for all data sets except Leipzig 200 where AVM scored the closest value for this data sample (see table 4). VM always attains the second or third closest rate to those of the ODP while AAE indicates the least efficiency concerning all data sets because it represents the most dissimilar values for all data sets.

Despite the afore-evaluated parameters, with respect to the mean cluster density, AAE shows strikingly good performance: It reaches the best efficiency for the data sets Leipzig 200, Leipzig 2000, Saxony 200, as well as for Zwickau 2000. For Saxony 200 and Saxony 2000, the new mask AVM demonstrates the most similar rates while for Leipzig 200, Leipzig 2000, and Zwickau 200, AVM has a much higher mean cluster density than the ODP and succeeds the worst here. Similar observations can be made for Saxony 200, Saxony 2000, and Zwickau 2000 where VM represents the lowest efficiency. This can be explained by the methods' characteristics of transferring data points to the same location. While VM snaps data points to the edge of a Voronoi polygon which

is located between two data points, AVM moves data points to the same intersection. Unlike this, DM has a comparably low cluster density compared to those of VM and AVM. This is due to the smaller displacement distances and the chosen random direction of this method, which has a smaller tendency of transferring data points to the same location as VM and AVM. Contrary to that, AAE has the lowest cluster density for so many data sets which can be justified due to the random displacement within a particular area and the low possibility of randomly moving points to the same location. The higher mean cluster density of AVM and VM prove the tendency of geomasks to enhance the coherence of existing cluster ellipsoids (Seidl et al., 2015).

For the NNHCA, it can be concluded that DM demonstrates the strongest performance in terms of its clusters' orientation and size, the number of clusters as well as for the mean points. Contrary to the methods demonstrating a very strong or very weak performance in terms of a specific parameter, DM has not demonstrated any strong penchants. Quite the converse, DM seems to even out itself by always obtaining rather similar values to that of the ODP.

For the interest of this thesis, the new method AVM has presented a great ability regarding the number of clusters as well as for the mean cluster density. VM also proved a strong liability in terms of mean cluster density and of size and orientation with the ODP clusters whereas AAE has shown the worst efficiency concerning the number of clusters.

5 Discussion

After presenting the results of the statistical methods, this chapter collects the information revealed and answers the five research questions of this study.

Research question 1: *How do commonly used geomasking methods affect the false re-identification risk?*

To revisit the definition of false re-identification, false re-identification is described as the incorrect concatenating of a person or household to a specific data point and has been identified as an ordinary problem of obfuscating techniques. Chapter 2.3 introduced commonly applied geomasks including their risk of false and correct re-identification. Whereas the obfuscating techniques part of the *aggregation*-class do not demonstrate a risk of false re-identification due to the clustering of data points, the obfuscating techniques of the other groups do not preclude this obstacle: Any method may transfer an ODP to an address location which either has been part of the research or has been excluded from the research and, hence, has unwillingly become part of it by receiving a data point. This can be explained by two factors: Firstly, and most importantly of all, many methods - except for the aggregation methods *aggregation to a street intersection or segment* and the *location swapping method* - neglect the underlying geography. Due to that, points can be moved to other residences but also to illogical locations such as forests, lakes, and agricultural land. Predominantly, this problem has been reported for the methods DM, AAE, and VM in the figures 17b, 28c, or 33, however, this issue also applies to other methods such as Grid Masking, Bimodal Gaussian Displacement or the Privy method. Secondly, many methods do not consider the underlying population density. These methods are particularly the masks of the group *affine transformation* such as the Privy method or stochastic mechanism but also some methods of the *RP-*, *Flipping-*, or *Blurring*-group. Nevertheless, methods estimating the population density within their study areas such as AAE or DM, enable the masker to determine a level of K-anonymity, describing the process of a data point not being distinguishable from at least k-1 other data points. Thus, it is more challenging for the map viewer to identify the original location and, thus, identify an individual correctly or wrongly. Finally, the risk of false identification always exists for all methods not aggregating data points. This is due to the fact that a map viewer may try to trace back the original location of data points. A data point situated on a street surrounded by, for example, ten houses, has a chance of 10% being correctly re-identified on the one hand, on the other hand, it also has a 10% possibility of being mistakenly identified. Still, a correct or false identification can lead to discrimination, harassment or stigmatization which should be prevented as much as possible and has been highlighted in academic literature (Seidl et al., 2015, Duckham and Kulik, 2006, Schilit et al., 2003, Schwab et al., 2011).

Research question 2: *To what degree can the proposed geomasking technique reduce the false re-identification risk?*

Research showed that geomasks have been considering the underlying population density to decrease the displacement error. Further, as discussed for research question one and in chapter 2.3, the underlying topography as well as the population density play vital roles regarding the risk of false re-identification. The new method AVM tackles these issues by, firstly, considering the underlying population density to assure that the data points cannot be distinguished from k-1 data points. Furthermore, the technique affirms that points stay within their merged polygon avoiding an impact of the displacement error (for example, moving a data point to an abandoned area without any other residence). Secondly, AVM displaces all data points to their nearest street intersection. Through that, it is ensured that no point is situated on a domicile or a void location (i.e., forest, river, etc.). Furthermore, AVM may snap multiple data points to the same intersection increasing their K-anonymity once again. Hence, in terms of wrongly re-identifying an individual by transferring the MDP to a residency, the risk is not existent. As explained for research question one, a map viewer can still try to re-identify a location by assuming the point's origin. However, the amount of buildings around an intersection is generally higher decreasing the risk of false re-identification. Moreover, the map viewer cannot distinguish how many points are situated on the intersection. Ergo, it can be concluded that the proposed method AVM demonstrates an extraordinarily low level of false re-identification - predominantly in comparison to other commonly applied geomasks - and has reached the first part of the main objective of this study to create the mechanism AVM through which the individual's privacy is protected and the false re-identification risk is decreased.

Research question 3: *To what degree can the proposed geomasking technique provide confidentiality for the individuals while maintaining data utility?*

The displacement process of AVM includes four steps: At first, the considered polygons are dissolved based on a selected DV. Afterward, all data points which are situated within a merged polygon with at least one more data point are constructing a Voronoi polygon to which closest boarders the points are relocated to. If the edge of the Voronoi polygon is located outside the SKA polygon, the data point will be transferred to the outer boundary of the joined polygon. Points that have not been displaced by the Voronoi mechanism, are randomly dispersed within their SKA polygons. Finally, all data points are snapped to the closest intersection. Due to this complex displacement procedure, it is rather challenging to trace back the original location and, hence, identify an individual. Besides, as discussed before, in case of several data points sharing the same intersection, their K-anonymity is decreased. Further, points situated on an intersection decrease the chance of appointing the right house of all buildings surrounding the intersection - not knowing *if* the right building is located at this intersection.

Providing the individual's confidentiality is one of the tasks a geomask has to offer next to maintaining a similar point pattern to that of the ODP and the capability of distinguishing clusters and spatial relations (as elaborated in chapter 2.3). Therefore, the preservation of the individuals' confidentiality is absolutely useless in terms of spatial analysis when the masked data set does not sustain data utility. For instance, previous authors have criticized masking techniques that aggregate data points to one point or within one area since the spatial resolution is extremely reduced and impede the ability to examine spatial relations or clusters (Armstrong et al., 1999, Zandbergen, 2014, Kwan et al., 2004, Hampton et al., 2010). Other methods such as the affine transformation rotating method sustain the relative point density but may move data points to another residence, increasing the risk of false re-identification (see chapter 2.3). Consequently, this study applied ESDA to assess whether AVM sustains data utility and whether information loss can be detected. For this, the examination of the mean and median center for the overall location of the data points, Ripley's K-function to explore the spatial distribution of the ODP and MDP, as well as the NNHCA to investigate the MDP's clusters' density, size, and orientation and compare them with the ellipsoids of the ODP, were implemented. Further, a visualization of the data points is conducted to scrutinize the outer shape and the risk of false re-identification of the MDP. The selection of these methods is based on other researches that evaluate these techniques as beneficial for perusing a geomask's performance (Seidl et al., 2015, Kwan et al., 2004). Additionally, the methods were not only applied on the ODP and data sets obfuscated by AVM, but they were also implemented on data sets masked by AAE, VM, and DM. As a reminder, VM and AAE were picked since the proposed method had been established on their concepts whereas DM was chosen due to its small effect on the spatial characteristics of the ODP as discussed by Hampton et al. (2010), Zandbergen (2014), Allshouse et al. (2010), Lu et al. (2012). To the first statistical approach, the visualization of the MDP, AVM achieved auspicious results: Only two data sets represent three outliers (Leipzig 200 and Zwickau 200, refer to figures 33c and 35c in chapter 4.1) while another sample (Zwickau 2000) exhibits only one AVM data point outside of the original extent (see figure 36c). Here, the visualized outcome also proved that no AVM data point was neither situated on an illogical location such as a forest or lake nor to a residence. Every AVM data point was moved to an intersection decreasing the risk of false re-identification.

The analysis of the mean and median centers revealed that AVM has only a small displacement distance for high population density data. For example, the mean centers of Leipzig 2000 and Zwickau 2000 had displacement distances of less than 11 m to the mean center of the ODP. Concerning the median centers, AVM performed less strikingly. Here, it demonstrated the best displacement distance at 1,61 m for Zwickau 2000 whereas, for Saxony 200, it fared the worst with 1528 m.

The Ripley's K-function divulged that AVM achieved significantly similar obsK-values to those of the ODP throughout all distances. Moreover, AVM even attained the closest obsK-values for the data sets Leipzig 2000 and Zwickau 2000. This places stable confidence in the new method since the MDP maintain a spatial dependency on the ODP and are not randomly dispersed as observed by other methods (see chapter 4.3) .

The third analysis proved that AVM scores a similar amount of clusters as the ODP as well as for the mean cluster density. Nevertheless, the method developed ellipsoids which do not always correlate with the location of the original cluster ellipsoids. Adding to that, the AVM clusters do not always align with those of the ODP regarding their size and orientation. This is particularly

observable in figure 49.

Based on the results of the statistical methods and the visualization of the MDP, it can be indicated that AVM provides an efficient data utility due to its strong point linkage to the ODP, the maintenance of the original spatial extent, and the overall location of its MDP. Only with regards to the derived clusters, AVM did not succeed very well, however, AVM does not transfer data points to other residencies decreasing the risk of false re-identification and providing confidentiality protection. Thus, it can be concluded that AVM has found a balance between the protection of individuals while retaining data utility as suggested by Zandbergen and contested in chapter 1.3.2.

Research question 4: *How does the proposed method comply with the rules and regulations by the General Data Protection Regulation (GDPR)?*

With respect to this research question, art. 4, paragraphs 1 and 5, were further considered since they identify locational and health data as personal data, as well as art. 9, paragraph 1, which describes the processing of personal data is prohibited in case of revealing health data. Hence, studies using original data points without obscuring or aggregating them or studies applying a geomasking technique with a high risk of re-identification do not comply with the GDPR because the data points resemble coordinates that reveal an individual's location. And location is considered a personal identifier, violating the right of the protection of personal data. A geomask obscuring health data is only complying with the GDPR in case of providing full confidentiality, meaning a risk of re-identification is non-existing. This is due to the fact, that personal data cannot be processed in case of revealing health data, but in case of full confidentiality, the point data would not disclose individuals anymore. Since AVM dislocates the ODP from their original location and transfers them to a street intersection - which does not have an address that could be associated with an individual -, the opportunity of correctly identifying the original location and, therefore, an individual is not given and, hence, ensures confidentiality. Thus, AVM does comply with the rules and regulations of the GDPR and can be applied for masking health or other personal data, fulfilling the second study objective of this research (see chapter 1.3.2).

Research question 5: *To what level is the GDPR protecting the individual's privacy with respect to information derived from spatial data?*

Spatial data is part of personal data because it can be implemented directly or indirectly as an identifier. This was elaborated in chapter 2.1 where spatial data can reveal an individual by his or her address, and, therefore, has to be treated as personal data. It is evident that the GDPR is generally vaguely worded and difficult to comprehend, which can result in various implementations and interpretations. For instance, the statement that a data subject *should* give permission for specific areas of scientific research proves the vagueness and the room for interpretation plus that decisions base on personal interests. As elaborated, the term *certain areas of scientific research* was not defined by the GDPR creating room for interpretability for the scientist as well. As highlighted in chapter 1.1, several researches in the field of health geography or reproductive health have published individual data. However, no consent needs to be given for this since it can be argued that this research was conducted in the interest of "provision of health or social care" as stated in paragraph 57. Therefore, this is can be interpreted as a sign of an emerging gap between ethical and legal requirements which was also discussed by Staunton et al. (2019). The definition of transparent data processing is also more or less broadly held, leaving again enough scope for interpretation. Generally, transparency practices on processing personal data are very low, questioning whether the GDPR protects the data subject's personal data in this regard. To conclude, it can be said, that the GDPR surely aims to preserve personal data, it's processing, and storage, and, therefore, the individual's privacy. Still, there is room for further protection, especially regarding interpretations and implementations or when considering that every EU member state has room for their own rules and regulations, too. Therefore, it can be said that the GDPR more or less illustrates another "disclosure limitation strategy".

6 Conclusion

This study sought to develop the new geographical masking method AVM based on the concepts of AAE and VM to decrease the risk of false re-identification and to protect the individual's privacy while maintaining data utility. Hereby, AVM considers the underlying population density by defining a level of K-anonymity as AAE does and displaces a part of the ODP based on the concept of VM. Additionally, AVM also follows an alternative approach than other obfuscating techniques by considering the underlying geography and transfers data points to the closest street intersection. Thereby, it decreases the risk of false re-identification immensely and does not relocate data points to illogical positions.

To evaluate whether the proposed method maintains data utility and preserves the point pattern of the original data set, the new mask as well as AAE, VM, and DM were applied on six data sets resembling either 200 or 2.000 address data points in the city of Leipzig, district of Zwickau, and the free state of Saxony in Eastern Germany. Subsequently, the MDP were examined by conducting a thorough visualization of the data points, the interpretation of the mean and median centers, the Ripley's K-function as well as the NNHCA. Apart from the NNHCA, AVM scored promising results for all methods: The visualization of the MDP including a base map demonstrated that only three out of six data sets presented one to three MDP outside of the original point extent and that all data points were situated on an intersection decreasing the risk of false re-identification. However, in terms of the analysis of the preservation of the original outer shape, VM outperformed all methods with only one data point situated beyond the original extent. Contrary to that, AAE performed the worst by transferring up to 21 points outside the extent. Regarding the mean and median centers, DM fares the best at the mean centers whereas VM succeeds the most auspicious for the median centers. Nevertheless, AVM also receives short displacement distances - predominantly in areas with a high population density - which do not distinguish so much of the displacement distances of DM and VM. Again, AAE demonstrates the worst performance by moving the data points at the farthest distances. About the Ripley's K-function, AVM retained similar obsK-values as the ODP's throughout all distances describing a strong point linkage to the ODP, putting strong confidence in the new method. Still, DM exceeds AVM and the other methods due to the number of most similar obsK-values whilst AAE scored the lowest results again as it scatters its points more than the ODP and abandons the point linkage to them. VM has a stronger point accumulation than the ODP (except for Zwickau 2000, bands one to ten) and represents the most dissimilar values for the data sets with a low point density which can be justified by the mask's disadvantage of changing the point pattern in scattered areas. Finally, the NNHCA revealed that DM eclipses all other methods regarding all parameters (size, orientation, amount, mean points) except for the cluster density. Apart from the other statistical analyses, AAE overcomes the other three masks regarding this parameter followed by AVM. AVM scores similar values in regard to the number of clusters and mean points as well as cluster density to that of the ODP. nevertheless, some ellipsoids do not align with the ODP clusters regarding orientation and size. Additionally, AVM produces some outliers.

The statistical analyses evidenced that AVM is an auspicious technique to protect confidential discrete spatial data while maintaining data utility. It did not outperform the other masking techniques DM or VM regarding the mean and median centers nor for the NNHCA but received very promising results for the Ripley's K-function preserving the original point pattern. Adding to that, apart from AAE it is the only method that preserves the SKA accurately (DM does this only partly) **and** minimizes the risk of false re-identification. The great efficiency regarding this parameter is originating from the fact that the proposed mask considers the underlying topography and moves the data points to an intersection. Since some of the data points are snapped to the same intersection, their K-anonymity is increased, too. However, it can be argued that a map viewer will count fewer data points influencing the spatial analysis of the map (for instance, illustrating fewer data points representing a lethal disease as Ebola could influence disaster management or the spatial analysis of the spreading of the disease). Contrary to that, DM and VM as well as AAE can transfer data points to other residences or parcels increasing the risk of false re-identification. Based on these three factors, it can be concluded that AVM is the most encouraging method in terms of the preservation of data utility and decreasing the risk of false re-identification to protect the individual's privacy and has met the first objective of this research. Finally, AVM is both applicable for web-services or for maps in scientific research creating opportunities for unique users.

Another focus of this research was the investigation of geoprivacy and its growing significance in our web-enabled world and whether rules and regulations exist to protect the individual and his or her personal data. Here, it was concluded that the GDPR is vaguely worded and leaves a lot of room for interpretation to the benefit of companies. Moreover, each European country has additional laws giving companies different opportunities or even gaps to process personal data. Due to that, it must be emphasized that an individual should view her or his responsibilities regarding data usage and production as well. Privacy settings should be read and in case of not understanding them, it should be considered whether consent will be given. Regarding the issue of publishing an individual's data in scientific research, the authors need to be aware of the consequences for an individual being exposed - no matter whether the map represents a stigmatizing disease or the purchase of a specific vehicle. In addition to that, the author should not talk their way out by saying they did nothing illegal, yet, they did something beneficial for the community. This can only increase the gap between the ethical considerations and the legal regulations which can also be linked to the fact that our technological world is constantly growing while our legal system cannot keep up with it.

Finally, the second objective of this study was to investigate whether AVM complies with the rules and regulations by the GDPR. AVM does comply with it since the new method moves the ODP to a street intersection that does not contain an address that can disclose an individual. Ergo, AVM also fulfills the second research objective.

6.1 Limitations and Future Recommendations

To not exceed the frame of this thesis, only a few statistical measurements were applied to evaluate the efficiency of AVM. Still, many other measurements exist such as the global Moran's I for spatial autocorrelation or distance to K-nearest neighbor. It is of valuable research to implement these and other statistical methods in future research to evaluate AVM. In addition, it is of great interest to apply AVM on larger data sets as originally planned to analyze its performance on enormous point data sets.

These limitations suggest new prospects for future research: First, it is recommended to juxtapose AVM with other obfuscating techniques that had not been applied in this research to gather more knowledge about the new approach. Second and in addition to that, it would be worthwhile to compare the applied geomasks - but also others - with alternative approaches protecting the individual's confidentiality (such as anonymization methods) to investigate their weaknesses and strengths. Third, it would be interesting to alter the AVM-algorithm and merge it with the very efficient DM-algorithm. For example, data points that are not affected by VM might be masked by DM. Fourth, when developing new geomasks, the underlying geography should be automatically considered to decrease the risk of false re-identification and, hence, to circumvent stigmatization and harassment. It has become evident that almost none of the obfuscating methods do consider the surrounding topography. Fifth, the power of mapping has to become more popular. Here, an increase in the comprehension of this issue was detected already which is a step forward. However, privacy guidelines as established by Kounadi and Resch (2018) as well as the existence of geomasks have to become more well-known to researchers, institutions, companies, or the public sector. A first step to reach this goal is to make geomasks more accessible. During this research, it was discovered that only the geomask DM is retrievable online for free. This is confounding considering the fact that many researchers stress to obfuscate confidential discrete spatial data (Zandbergen, 2014; Ajayakumar et al., 2019; see chapter 1.1). A second step is to employ geomasks for open source software such as QGIS as well since ESRI software is rather expensive. Through that, companies, researchers, and institutions can share their findings with the public without jeopardizing an individual too much.

References

AbdelMalik, P., Boulos, M. N. K., & Jones, R. (2008). The perceived impact of location privacy: A web-based survey of public health perspectives and requirements in the uk and canada. *BMC Public Health*, *8*(1), 156.

Ajayakumar, J., Curtis, A. J., & Curtis, J. (2019). Addressing the data guardian and geospatial scientist collaborator dilemma: How to share health records for spatial analysis while maintaining patient confidentiality. *International Journal of Health Geographics*, *18*(1), 1–12.

Allshouse, W. B., Fitch, M. K., Hampton, K. H., Gesink, D. C., Doherty, I. A., Leone, P. A., Serre, M. L., & Miller, W. C. (2010). Geomasking sensitive health data and privacy protection: An evaluation using an e911 database. *Geocarto international*, *25*(6), 443–452.

Alrayes, F., & Abdelmoty, A. (2014). No place to hide: A study of privacy concerns due to location sharing on geo-social networks. *International Journal On Advances in Security*, *7*(3/4), 62–75.

ArcGIS PRO. (2020). How Multi-Distance Spatial Cluster Analysis (Ripley's K-function) works [Last accessed July 21st, 2020]. https://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/h-how-multi-distance-spatial-cluster-analysis-ripl.htm

ArcGIS Pro. (2020). Dissolve (data management) [Last accessed July 23rd, 2020.]. https://pro.arcgis.com/de/pro-app/tool-reference/data-management/dissolve.htm

Ardagna, C. A., Cremonini, M., di Vimercati, S. D. C., & Samarati, P. (2008). Privacy-enhanced location-based access control, In *Handbook of database security*. Springer.

Armstrong, M. P. (2002). Geographic information technologies and their potentially erosive effects on personal privacy. *Studies in the Social Sciences*, *27*(1), 19–28.

Armstrong, M. P., Rushton, G., & Zimmerman, D. L. (1999). Geographically masking health data to preserve confidentiality. *Statistics in medicine*, *18*(5), 497–525.

Aurenhammer, F., & Klein, R. (2000). Voronoi diagrams. *Handbook of computational geometry*, *5*(10), 201–290.

Bailey, T., & Gatrell, A. (1995). Interactive spatial data analysis. 1995. *Harlow: Longman*.

Balser, M., & Hurtz, S. (2020, March 18). Warum die Telekom Bewegungsdaten von Handynutzern weitergibt. *Süddeutsche Zeitung* [Last accessed 6 April 2020]. https://www.sueddeutsche.de/digital/coronavirus-telekom-smartphone-tracking-datenschutz-1.4850094

Bamberger, K. A., & Mulligan, D. K. (2015). *Privacy on the ground: Driving corporate behavior in the united states and europe*. MIT Press.

Beresford, A. R., & Stajano, F. (2003). Location privacy in pervasive computing. *IEEE Pervasive computing*, *2*(1), 46–55.

Bettini, C., Wang, X. S., & Jajodia, S. (2005). Protecting privacy against location-based personal identification, In *Workshop on secure data management*. Springer.

Blumberg, A. J., & Eckersley, P. (2009). On locational privacy, and how to avoid losing it forever. *Electronic frontier foundation*, *10*(11).

Bob, Y., & Hoffman, G. (2020, March 17). Shin Bet confirms it is currently using surveillance tools. *The Jerusalem post* [Last accessed April 6th, 2020]. https://www.jpost.com/breaking-news/use-of-digital-means-to-track-coronavirus-patients-approved-621237

Botha, J., Grobler, M., Hahn, J., & Eloff, M. (2017). A high-level comparison between the south african protection of personal information act and international data protection laws, In *International conference on cyber warfare and security conference proceedings*.

Boulos, M. N. K., Cai, Q., Padget, J. A., & Rushton, G. (2006). Using software agents to preserve individual health data confidentiality in micro-scale geographical analyses. *Journal of Biomedical Informatics*, *39*(2), 160–170.

Boulos, M. N. K., Curtis, A. J., & AbdelMalik, P. (2009). Musings on privacy issues in health research involving disaggregate geographic data about individuals. BioMed Central.

Bourke, P. (1988). Calculating the area and centroid of a polygon. *Swinburne Univ. of Technology*, *7*.

Bridwell, S. A. (2007). The dimensions of locational privacy, In *Societies and cities in the age of instant access*. Springer.

Brownstein, J. S., Cassa, C. A., & Mandl, K. D. (2006). No place to hide—reverse identification of patients from published maps. *New England Journal of Medicine*, *355*(16), 1741–1742.

Bundesamt für Kartographie und Geodäsie. (2019). Verwaltungsgebiete 1:250 000 (Ebenen), Stand 01.01. (VG250 01.01.) [Last accessed April 16th, 2020]. https://gdz.bkg.bund.de/index.php/default/open-data/verwaltungsgebiete-1-250-000-ebenen-stand-01-01-vg250-ebenen-01-01.html

Burt, J. E., Barber, G. M., & Rigby, D. L. (2009). *Elementary statistics for geographers*. Guilford Press.

Cassa, C. A., Grannis, S. J., Overhage, J. M., & Mandl, K. D. (2006). A context-sensitive approach to anonymizing spatial surveillance data: Impact on outbreak detection. *Journal of the American Medical Informatics Association*, *13*(2), 160–165.

Cassa, C. A., Wieland, S. C., & Mandl, K. D. (2008). Re-identification of home addresses from spatial locations anonymized by gaussian skew. *International journal of health geographics*, *7*(1), 45.

Chainey, S., Tompson, L., & Uhlig, S. (2008). The utility of hotspot mapping for predicting spatial patterns of crime. *Security journal*, *21*(1-2), 4–28.

Christin, D., Reinhardt, A., Kanhere, S. S., & Hollick, M. (2011). A survey on privacy in mobile participatory sensing applications. *Journal of systems and software*, *84*(11), 1928–1946.

Cole, D., & Fabbrini, F. (2016). Bridging the transatlantic divide? the united states, the european union, and the protection of privacy across borders. *International Journal of Constitutional Law*, *14*(1), 220–237.

Council, N. R. Et al. (2007). *Putting people on the map: Protecting confidentiality with linked social-spatial data*. National Academies Press.

Curtis, A. J., Mills, J. W., & Leitner, M. (2006). Spatial confidentiality and gis: Re-engineering mortality locations from published maps about hurricane katrina. *International Journal of Health Geographics*, *5*(1), 44.

Curtis, A., Mills, J. W., Agustin, L., & Cockburn, M. (2011). Confidentiality risks in fine scale aggregations of health data. *Computers, Environment and Urban Systems*, *35*(1), 57–64.

Custers, B., Dechesne, F., Sears, A. M., Tani, T., & van der Hof, S. (2018). A comparison of data protection legislation and policies across the eu. *Computer Law & Security Review*, *34*(2), 234–243.

Dekkers, M., Polman, F., Te Velde, R., & De Vries, M. (2006). Measuring european public sector information resources. *Final Report of Study on Exploitation of public sector information–benchmarking of EU framework conditions*.

Dixon, P. M. (2014). R ipley's k function. *Wiley StatsRef: Statistics Reference Online*.

Dresden, L. (2020). Bevölkerungsbestand [Last accessed April 14th, 2020]. https://www.dresden.de/de/leben/stadtportrait/statistik/bevoelkerung-gebiet/Bevoelkerungsbestand.php

Duckham, M., & Kulik, L. (2006). Location privacy and location-aware computing, In *Dynamic and mobile gis*. CRC press.

Duncan, G. T., Pearson, R. W. Et al. (1991). Enhancing access to microdata while protecting confidentiality: Prospects for the future. *Statistical Science*, *6*(3), 219–232.

Ebdon, D. (1988). Statistics in geography.

El Emam, K. (2013). *Guide to the de-identification of personal health information*. Auerbach Publications.

ESRI. (2020a). ArcGIS Pro - 2D and 3D mapping software [Last accessed on February 19th, 2020]. https:%20//www.esri.com/de-de/arcgis/products/arcgis-pro/overview

ESRI. (2020b). How mean center works [Last accessed March 13th, 2020]. https://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/h-how-mean-center-spatial-statistics-works.htm

ESRI. (2020c). How median center works [Last accessed March 13th, 2020]. https://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/h-how-median-center-spatial-statistics-works.htm

ESRI. (2020d). Multi-distance spatial cluster analysis (ripley's k function) [Last accessed March 12th, 2020]. https://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/multi-distance-spatial-cluster-analysis.htm

ESRI. (2020e). Reverse geocode. [Last accessed January 31st, 2020]. https://desktop.arcgis.com/de/arcmap/10.3/tools/geocoding-toolbox/reverse-geocode.htm

ESRI. (2020f). Robert Koch-Institut: COVID-19-Dashboard [Last accessed April 7th, 2020]. https://www.dresden.de/de/rathaus/aktuelles/pressemitteilungen/2020/01/pm_036.php

ESRI Deutschland. (2020). Adressen - Sachsen [Last accessed April 14th, 2020]. https://opendata-esri-de.opendata.arcgis.com/datasets/esri-de-content::adressen-sachsen?geometry=6.577%5C%2C49.700%5C%2C20.452%5C%2C52.124&selectedAttribute=REGBEZ

European Data Protection Supervisor. (2020). Health [Last accessed March 31st, 2020]. https://edps.europa.eu/data-protection/our-work/subjects/health_en

European Parliament. (2016). Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation). *Official Journal ofthe European Union*, (59: L 119).

Everitt, B. (1974). Cluster analysis heinemann. London.

Föderl-Schmidt, A. (2020, March 21). Coronavirus in Israel: Geheimdienst soll Infizierte aufspüren. *Süddeutsche Zeitung* [Last accessed April 6th, 2020]. https://www.sueddeutsche.de/politik/israel-coronavirus-geheimdienst-1.4852398

Fronterrè, C. (2018). Spatial analysis of geomasked and aggregated data.

Fu, W. J., Jiang, P. K., Zhou, G. M., & Zhao, K. L. (2014). Using moran's i and gis to study the spatial pattern of forest litter carbon density in a subtropical region of southeastern china. *Biogeosciences*, *11*(8), 2401.

Furini, M., & Tamanini, V. (2015). Location privacy and public metadata in social media platforms: Attitudes, behaviors and opinions. *Multimedia Tools and Applications*, *74*(21), 9795–9825.

Gambs, S., Killijian, M.-O., & del Prado Cortez, M. N. (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*.

Ghinita, G., Zhao, K., Papadias, D., & Kalnis, P. (2010). A reciprocal framework for spatial k-anonymity. *Information Systems*, *35*(3), 299–314.

Graham, C. (2012). Anonymisation: Managing data protection risk code of practice. *Information Commissioner's Office*.

Grubesic, T. H., & Murray, A. T. (2001). Detecting hot spots using cluster analysis and gis, In *Proceedings from the fifth annual international crime mapping research conference*.

Gruteser, M., & Grunwald, D. (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*.

Gupta, R., & Rao, U. P. (2020). Preserving location privacy using three layer rdv masking in geocoded published discrete point data. *World Wide Web*, *23*(1), 175–206.

Haley, D. F., Matthews, S. A., Cooper, H. L., Haardörfer, R., Adimora, A. A., Wingood, G. M., & Kramer, M. R. (2016). Confidentiality considerations for use of social-spatial data on the social determinants of health: Sexual and reproductive health case study. *Social Science & Medicine*, *166*, 49–56.

Hampton, K. H., Fitch, M. K., Allshouse, W. B., Doherty, I. A., Gesink, D. C., Leone, P. A., Serre, M. L., & Miller, W. C. (2010). Mapping health data: Improved privacy protection with donut method geomasking. *American journal of epidemiology*, *172*(9), 1062–1069.

Hardwick, D., & Patychuk, D. (1999). Geographic mapping demonstrates the association between social inequality, teen births and stds among youth. *The Canadian Journal of Human Sexuality*, *8*(2), 77.

International Atomic Energy Agency. (2015). Incident and Trafficking Database (ITDB) [Last accessed May 11th, 2020]. http://www-ns.iaea.org/security/itdb.asp

Kaminski, M. E. (2015). When the default is no penalty: Negotiating privacy at the ntia. *Denv. L. Rev.*, *93*, 925.

Kar, B., Crowsey, R. C., & Zale, J. J. (2013). The myth of location privacy in the united states: Surveyed attitude versus current practices. *The Professional Geographer*, *65*(1), 47–64.

Keßler, C., & McKenzie, G. (2018). A geoprivacy manifesto. *Transactions in GIS*, *22*(1), 3–19.

Kounadi, O. (2015). *Geospatial privacy framework for confidential discrete data with emphasis on spatial crime analysis & visualization* (Doctoral dissertation). Department of Criminal Justice, Temple University, Philadelphia, USA.

Kounadi, O., Bowers, K., & Leitner, M. (2015). Crime mapping on-line: Public perception of privacy issues. *European journal on criminal policy and research*, *21*(1), 167–190.

Kounadi, O., & Leitner, M. (2014). Why does geoprivacy matter? the scientific publication of confidential data presented on maps. *Journal of Empirical Research on Human Research Ethics*, *9*(4), 34–45.

Kounadi, O., & Leitner, M. (2015a). Defining a threshold value for maximum spatial information loss of masked geo-data. *ISPRS international journal of geo-information*, *4*(2), 572–590.

Kounadi, O., & Leitner, M. (2015b). Spatial information divergence: Using global and local indices to compare geographical masks applied to crime data. *Transactions in GIS*, *19*(5), 737–757.

Kounadi, O., & Leitner, M. (2016). Adaptive areal elimination (aae): A transparent way of disclosing protected spatial datasets. *Computers, Environment and Urban Systems*, *57*, 59–67.

Kounadi, O., & Resch, B. (2018). A geoprivacy by design guideline for research campaigns that use participatory sensing data. *Journal of Empirical Research on Human Research Ethics*, *13*(3), 203–222.

Kounadi, O., Resch, B., & Petutschnig, A. (2018). Privacy threats and protection recommendations for the use of geosocial network data in research. *Social Sciences*, *7*(10), 191.

Kramer, A. D., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, *111*(24), 8788–8790.

Krieger, N. (2003). Place, space, and health: Gis and epidemiology. *Epidemiology*, *14*(4), 384–385.

Krieger, N., Waterman, P., Chen, J. T., Soobader, M.-J., Subramanian, S. V., & Carson, R. (2002). Zip code caveat: Bias due to spatiotemporal mismatches between zip codes and us census–defined geographic areas—the public health disparities geocoding project. *American journal of public health*, *92*(7), 1100–1102.

Krumm, J. (2007). Inference attacks on location tracks, In *International conference on pervasive computing*. Springer.

Kulk, S., & Van Loenen, B. (2012). Brave new open data world? *International Journal of Spatial Data Infrastructures Research*, *7*, 196–206.

Kwan, M.-P., Casas, I., & Schmitz, B. (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.

Leitner, M., & Curtis, A. (2004). Cartographic guidelines for geographically masking the locations of confidential point data. *Cartographic Perspectives*, (49), 22–39.

Leitner, M., & Curtis, A. (2006). A first step towards a framework for presenting the location of confidential point data on maps—results of an empirical perceptual study. *International Journal of Geographical Information Science*, *20*(7), 813–822.

Levine, N. (2004). Hot spot analysis i [Last accessed July 26$^{\text{rd}}$, 2020.]. https://www.icpsr.umich.edu/CrimeStat/files/CrimeStatChapter.6.pdf

Levine, N. (2007). Crimestat: A spatial statistics program for the analysis of crime incident locations (v 3.1). ned levine associates, houston, tx, and the national institute of justice, washington, dc. march [Last accessed July 26$^{\text{rd}}$, 2020.]. https://www.icpsr.umich.edu/CrimeStat/contact.html

Levine, N. (2008). The "hottest" part of a hotspot: Comments on "the utility of hotspot mapping for predicting spatial patterns of crime". *Security journal*, *21*(4), 295–302.

Lu, Y., Yorke, C., & Zhan, F. B. (2012). Considering risk locations when defining perturbation zones for geomasking. *Cartographica: The International Journal for Geographic Information and Geovisualization*, *47*(3), 168–178.

Lynskey, O. (2015). *The foundations of eu data protection law*. Oxford University Press.

McDonald, A. M., & Cranor, L. F. (2008). The cost of reading privacy policies. *Isjlp*, *4*, 543.

Miller, J. (2013). City of london calls halt to smartphone tracking bins [Last accessed February 27$^{\text{th}}$, 2020]. https://www.bbc.com/news/technology-23665490

Monmonier, M. (2004). *Spying with maps: Surveillance technologies and the future of privacy*. University of Chicago Press.

Murad, A., Hilton, B., Horan, T., & Tangenberg, J. (2014). Protecting patient geo-privacy via a triangular displacement geo-masking method, In *Proceedings of the 1st acm sigspatial international workshop on privacy in geographic information collection and analysis*.

Oakes, J. M., & Johnson, P. J. (2006). Propensity score matching for social epidemiology. *Methods in social epidemiology*, *1*, 370–393.

ODS. (2020). Deutschland landkreise [Last accessed April 7$^{\text{th}}$, 2020]. https://public.opendatasoft.com/explore/dataset/landkreise-in-germany/table/

Olson, K. L., Grannis, S. J., & Mandl, K. D. (2006). Privacy protection versus cluster detection in spatial epidemiology. *American Journal of Public Health*, *96*(11), 2002–2008.

Onsrud, H. J., Johnson, J. P., & Lopez, X. (1994). Protecting personal privacy in using geographic information systems. *Photogrammetric Engineering and Remote Sensing*, *60*(9), 1083–1095.

Pasotti, A. (2020). Qgis python plugins repository - geocoding [Last accessed February 19th, 2020]. https://datausa.io/profile/geo/minneapolis-st.-paul-bloomington-mn-wi-metro-area#about

Ricker, B., Schuurman, N., & Kessler, F. (2015). Implications of smartphone usage on privacy and spatial cognition: Academic literature and public perceptions. *GeoJournal*, *80*(5), 637–652.

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

Sachsen, S. L. (2018). Bevölkerungsbestand [Last accessed April 16th, 2020]. https://www.statistik.sachsen.de/html/426.htm

Schilit, B., Hong, J., & Gruteser, M. (2003). Wireless location privacy protection. *Computer*, *36*(12), 135–137.

Schuppe, J. (2020). Google tracked his bike ride past a burglarized home. that made him a suspect.n [Last accessed March 13th, 2020]. https://www.nbcnews.com/news/us-news/google-tracked-his-bike-ride-past-burglarized-home-made-him-n1151761

Schwab, K., Marcus, A., Oyola, J., Hoffman, W., & Luzi, M. (2011). Personal data: The emergence of a new asset class, In *An initiative of the world economic forum*.

Seidl, D. E., Jankowski, P., & Clarke, K. C. (2018). Privacy and false identification risk in geo-masking techniques. *Geographical Analysis*, *50*(3), 280–297.

Seidl, D. E., Jankowski, P., & Nara, A. (2019). An empirical test of household identification risk in geomasked maps. *Cartography and Geographic Information Science*, *46*(6), 475–488.

Seidl, D. E., Paulus, G., Jankowski, P., & Regenfelder, M. (2015). Spatial obfuscation methods for privacy protection of household-level data. *Applied Geography*, *63*, 253–263.

Shu, J., Jia, X., Yang, K., & Wang, H. (2018). Privacy-preserving task recommendation services for crowdsourcing. *IEEE Transactions on Services Computing*.

Stadt Leipzig. (2020). Stadtwald und Auenwald [Last accessed July 19th, 2020]. https://www.leipzig.de/freizeit-kultur-und-tourismus/parks-waelder-und-friedhoefe/stadtwald-und-auenwald/

Statistisches Landesamt Sachsen. (2020). Bevölkerung [Last accessed April 16th, 2020]. https://www.statistik.sachsen.de/html/369.htm

Staunton, C., Slokenberga, S., & Mascalzoni, D. (2019). The gdpr and the research exemption: Considerations on the necessary safeguards for research biobanks. *European Journal of Human Genetics*, *27*(8), 1159–1167.

Sweeney, L. (2002). K-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(05), 557–570.

Thatcher, J. (2013). From volunteered geographic information to volunteered geographic services, In *Crowdsourcing geographic knowledge*. Springer.

Tor, B. (2020). Qgis - geocoding addresses tutorial [Last accessed February 13th, 2020]. https://guides.library.ucsc.edu/DSCguides/QGIS_GeocodingAddresses

Toubiana, V., & Nissenbaum, H. (2011). An analysis of google log retention policies.

Van den Berg, B., & Van der Hof, S. (2012). What happens to my data? a novel approach to informing users of data processing practices. *First Monday*, *17*(7).

Voronoi, G. (1908). Nouvelles applications des paramètres continus à la théorie des formes quadratiques. deuxième mémoire. recherches sur les parallélloèdres primitifs. *Journal für die reine und angewandte Mathematik*, *134*, 198–287.

Waller, L. A., & Gotway, C. A. (2004). *Applied spatial statistics for public health data* (Vol. 368). John Wiley & Sons.

Wang, H., Zhang, Z., & Taleb, T. (2018). Special issue on security and privacy of iot. *World Wide Web*, *21*(1), 1–6.

Wartell, J., & McEwen, J. T. (2001). Privacy in the information age: A guide for sharing crime maps and spatial data series. *Research Report, NCJ*, *188739*.

Weiser, P., & Scheider, S. (2014). A civilized cyberspace for geoprivacy, In *Proceedings of the 1st acm sigspatial international workshop on privacy in geographic information collection and analysis*.

Wieland, S. C., Cassa, C. A., Mandl, K. D., & Berger, B. (2008). Revealing the spatial distribution of a disease while preserving privacy. *Proceedings of the National Academy of Sciences*, *105*(46), 17608–17613.

Wong, D. W. (1999). Several fundamentals in implementing spatial statistics in gis: Using centrographic measures as examples. *Geographic Information Sciences*, *5*(2), 163–174.

Xu, R., & Wunsch, D. (2008). *Clustering* (Vol. 10). John Wiley & Sons.

Zandbergen, P. A. (2007). Influence of geocoding quality on environmental exposure assessment of children living near high traffic roads. *BMC public health*, *7*(1), 37.

Zandbergen, P. A. (2014). Ensuring confidentiality of geocoded health data: Assessing geographic masking strategies for individual-level data. *Advances in medicine*, *2014*.

Zhang, S., Freundschuh, S. M., Lenzer, K., & Zandbergen, P. A. (2017). The location swapping method for geomasking. *Cartography and Geographic Information Science*, *44*(1), 22–34.

Zimmerman, D. L., & Pavlik, C. (2008). Quantifying the effects of mask metadata disclosure and multiple releases on the confidentiality of geographically masked health data. *Geographical Analysis*, *40*(1), 52–76.

Zimmerman, D. L., Rushton, G., Armstrong, M. P., Gittler, J., Greene, B. R., Pavlik, C. E., & West, M. M. (2007). *Geocoding health data: The use of geographic codes in cancer prevention and control, research and practice*. CRC Press.

Appendices

A Appendix: Code of Voronoi Masking

```
# Voronoi Masking
# Concept from Seidl, Paulus, Jankowski and Regenfelder (2015). Spatial obfuscation methods
# for privacy protection of household-level data
# Implementation by Fiona Polzin
# f.s.polzin@students.uu.nl
# June 2020
#---------------------------------------------------------------------------

# The external variables in the lines 20, 22, and 24 must be set by the user.
# Code can be executed in ArcMAP 10.x and ArcGIS Pro

#---------------------------------------------------------------------------
# Import system modules
import arcpy

inputflag = 0
# Set user input or receive the input as parameters? (1=receive as parameters, 0= set in program)
if inputflag==0:
   # EXTERNAL variables to be inputed by user
    workspace = r""
    # User determines the local variables of the point case file
    PointFile = ""
    # user determines the search radius of the research area (i.e., 1000 meters)
    SearchRadius = "100000"
if inputflag==1:
    workspace = arcpy.GetParameterAsText(0)
    PointFile = arcpy.GetParameterAsText(1)
    AreaFile = arcpy.GetParameterAsText(2)
    SearchRadius = arcpy.GetParameterAsText(3)
    Copy_PointFile = arcpy.GetParametersAsText(4)
#---------------------------------------------------------------------------
# Set workspace
arcpy.env.workspace = workspace
arcpy.env.overwriteOutput = True #controls whether the algorithm will replace any present output automatically

#---------------------------------------------------------------------------
# Step1 = Create Thiessen Polygons
AreaFile = "VoronoiPolygon"
# Step 2 = Copy PointFile to compare the original with the masked positions at the end
Copy_PointFile = "VM_Points"
# Step 3 = Snap the Copy_Pointfile to closest segment of its corresponding Voronoi polygon

#---------------------------------------------------------------------------
# Step 1 - Create Thiessen Polygon (FeatureToPoint)
inFeatures = PointFile
outFeatureClass = AreaFile
outFields = "ALL"

#Execute Create Thiessen Polygon
arcpy.CreateThiessenPolygons_analysis(inFeatures, outFeatureClass, outFields)

#---------------------------------------------------------------------------
# Step 2 - Copy PointFile
in_features = PointFile
out_feature_class = Copy_PointFile

#Execute copy features
arcpy.CopyFeatures_management(in_features, out_feature_class)
#---------------------------------------------------------------------------
# Step 3 - Snap CaseFile to closest Voronoi polygon
inFeature = Copy_PointFile
snap_feature = AreaFile
```

```
snap_Type = "EDGE"
snap_Distance = SearchRadius

#Execute Snap
arcpy.Snap_edit(inFeature, [[snap_feature, snap_Type, snap_Distance]])
```

B Appendix: Code for Spatial K-Anonymity Polygons

```
# ----------ADAPTED FOR "SPACE FOR ETHICS" COURSE, LECTURER: OURANIA KOUNADI ----------
# Name: Adaptive Aerial Elimination (AAE) with dissolve
# Description: Dissolving polygons until all polygons have attribute values that are equal or greater than a minimum value.
# Author: O.Kounadi, 1/03/2019
# Purpose: To be used for educational purpose only
# Reference: This script is part of a code written for the adaptive geographical masking method that is published in https://doi.org/10.1016/j.compenvurbsys.2016.01.004

# Copy confidential points and polygon file to a new personal geodatabase
# All working output will be saved in the geodatabase
# -------------------------------------------

# -----------------1st BLOCK------------------
# Set workspace variable
workspace = r""

# Set disclosure value
Ka = 5

# Set variables for paths (the file name, e.g "\areas",  should not be replaced)
path = r"\areas"
path2 = r"\areas_new"
path3 = r"\new"
path4 = r"\new1"
path5= r"\o_points1"
path6= r"\m_points1"

# Import system modules
import arcpy, os, sys
import arcpy.mapping
from arcpy import env

# Set workspace
arcpy.env.workspace = workspace
arcpy.env.overwriteOutput = True

#  Exclude unnecessary features from the dissolving process
arcpy.CopyFeatures_management(path, "areas_new")
arcpy.SelectLayerByAttribute_management ("areas_new", "NEW_SELECTION",'[RORI]>= 5')
arcpy.CalculateField_management("areas_new","ID",'NULL')
arcpy.SelectLayerByAttribute_management("areas_new","CLEAR_SELECTION")


# Define variable to check for disclosure value
arcpy.SelectLayerByAttribute_management("areas_new", "NEW_SELECTION", '[RORI] in ( SELECT min( [RORI] ) FROM areas_new)')
check1 = arcpy.SearchCursor("areas_new")
for row in check1:
    dv = row.getValue("RORI")
del check1, row

# -----------------2nd BLOCK------------------
#  Dissolving process
while dv < Ka:
    arcpy.SelectLayerByAttribute_management("areas_new", "NEW_SELECTION", '[ID] in (SELECT min( [ID] ) FROM areas_new)')
    arcpy.SelectLayerByLocation_management("areas_new", "SHARE_A_LINE_SEGMENT_WITH", "", "", "NEW_SELECTION")
    # Select neighbors until min DV is reached
    list = []
    check2 = arcpy.SearchCursor("areas_new")
    for row in check2:
        dv1 = row.getValue("RORI")
        list.append(dv1)
        dv2 = sum(list)
    del check2, list, row, dv1
    while dv2 < Ka:
        arcpy.SelectLayerByLocation_management("areas_new", "
```

```
    SHARE_A_LINE_SEGMENT_WITH", "", "", "ADD_TO_SELECTION")
        list1 = []
        check3 = arcpy.SearchCursor("areas_new")
        for row in check3:
            dv1 = row.getValue("RORI")
            list1.append(dv1)
            dv2 = sum(list1)
        del check3, list1, row, dv1
    del dv2
    # Dissolve selected features
    arcpy.Dissolve_management("areas_new", "new", "#", "RORI SUM;ID MIN")
    arcpy.AddField_management("new", "RORI", "LONG")
    arcpy.AddField_management("new", "ID", "LONG")
    arcpy.CalculateField_management('new', 'RORI', '[SUM_RORI]', 'VB', '#')
    arcpy.CalculateField_management('new', 'ID', '[MIN_ID]', 'VB', '#')
    arcpy.DeleteField_management("new", "SUM_RORI")
    arcpy.DeleteField_management("new", "MIN_ID")
    arcpy.SelectLayerByAttribute_management("areas_new", "CLEAR_SELECTION")
    arcpy.Update_analysis("areas_new", "new", "new1")
    arcpy.CopyFeatures_management("new1", path)
    # Delete unnecessary files
    arcpy.Delete_management(path2)
    arcpy.Delete_management(path4)
    arcpy.Delete_management(path3)
    # Exlude previous feature from next iteration
    arcpy.CopyFeatures_management(path, "areas_new")
    arcpy.SelectLayerByAttribute_management("areas_new", "NEW_SELECTION", '[ID] in
    (SELECT min( [ID] ) FROM areas_new)')
    arcpy.CalculateField_management("areas_new", "ID", 'NULL')
    # Check for the next iteration
    arcpy.SelectLayerByAttribute_management("areas_new", "NEW_SELECTION",
                                            '[RORI] in (SELECT min( [RORI] ) FROM
    areas_new)')
    check4 = arcpy.SearchCursor("areas_new")
    for row in check4:
        dv = row.getValue("RORI")
    del check4, row

else:
    del dv
    print
    "K-anonymized areas are created."
    arcpy.CopyFeatures_management("areas_new", path)
```

C Appendix: Code of Adaptive Voronoi Masking

```
# Adaptive Voronoi Masking
# Implementation by Fiona Polzin
# f.s.polzin@students.uu.nl
# July 2020
#---------------------------------------------------------------------
# This code does not include the creation of spatial K-anonymity polygons. This can be implemented via the first part of AAE.

# The external and internal variables in the lines 16, 18, 19, 20, and 21 must be set by the user.
# Code can be executed in ArcMAP 10.x and ArcGIS Pro.
#---------------------------------------------------------------------
# Import system modules
import arcpy

# ----------------------------------------
# EXTERNAL variables to be inputed by user
workspace = r""
# User determines the local variables of the point case file
PointFile = ""
BlockFile = ""
StreetFile = ""
SearchRadius = "100000 Meters"

# ----------------------------------------------
# Step 1: Find points for VM
pointsJoin = "points_join"
pointsForVM = "Points_VM"
pointsNoVM = "Points_No_VM"
# Step 2: Create Thiessen and split(Union) them based on blocks
VoronoiPolygon = "VoronoiPolygon"
VoronoiUnion = "VoronoiPoly_Union"
# Step 3: Now, snap the pointsForVM to the closest edge of their corresponding Voronoi_clip polygon
# Step 4: Find the polygons with only data point
blocks= "Blocks"
# Step 5: Create one random points (points not for VM)
RandomPoints = "RandomPoints"
# Step 6: Merge StreetFile
MergedStreets = "UnsplitStreets"
# Step 7: Find street intersections
Intersection = "Intersection"
# Step 8: Move pointsForVM to closest intersection
# Step 9: Move pointsNoVM to closest intersection
# Step 10: merge pointsForVM and pointsNoVM to a new layer
AVM = "AVM_points"

# ----------------------------------------------
# Set workspace
arcpy.env.workspace = workspace
arcpy.env.overwriteOutput = True  # controls whether the algorithm will replace any present output automatically
arcpy.env.addOutputsToMap = True  # add output to map automatically

# ----------------------------------------------
# Step 1 - Find points for VM and not for VM
arcpy.analysis.SpatialJoin(PointFile, BlockFile, "points_join", "JOIN_ONE_TO_ONE", "KEEP_ALL", )

arcpy.analysis.Statistics("points_join", "stat", "id COUNT", "id")
arcpy.management.CopyFeatures(BlockFile, "Blocks")
arcpy.JoinField_management("Blocks", "id", "stat", "ID", "COUNT_ID")

arcpy.SelectLayerByAttribute_management("Blocks", selection_type="NEW_SELECTION", where_clause='"COUNT_ID" = 1')
arcpy.SelectLayerByLocation_management(in_layer=PointFile, overlap_type="INTERSECT", select_features="Blocks", search_distance="", selection_type="NEW_SELECTION", invert_spatial_relationship="NOT_INVERT")
arcpy.SelectLayerByLocation_management(in_layer=PointFile, overlap_type="INTERSECT", select_features="Blocks", search_distance="", selection_type="NEW_SELECTION", invert_spatial_relationship="NOT_INVERT")
```

```
arcpy.CopyFeatures_management(PointFile, "Points_No_VM")


arcpy.SelectLayerByLocation_management(in_layer=PointFile, overlap_type="INTERSECT", select_features="Blocks", search_distance="", selection_type="SWITCH_SELECTION", invert_spatial_relationship="NOT_INVERT")
arcpy.CopyFeatures_management(PointFile, "Points_VM")

arcpy.SelectLayerByAttribute_management("Blocks", selection_type="CLEAR_SELECTION")
arcpy.SelectLayerByAttribute_management(PointFile, selection_type="CLEAR_SELECTION")

# ----------------------------------------------
# Step 2 - Create Thiessen and split(Union) them based on blocks
arcpy.env.extent = BlockFile
arcpy.CreateThiessenPolygons_analysis(pointsForVM, VoronoiPolygon, "ALL")
arcpy.Union_analysis([VoronoiPolygon,BlockFile], VoronoiUnion)


# ----------------------------------------------
# Step 3 - Loop to snap data points to closest segment
inFeature = pointsForVM
snap_feature = VoronoiUnion
snap_Type = "EDGE"
snap_Distance = SearchRadius

arcpy.Snap_edit(inFeature, [[snap_feature, snap_Type, snap_Distance]])

#----------------------------------------------
#Step 4 - Find all polygons which contain data points not transferred by VM
arcpy.SelectLayerByAttribute_management("Blocks", selection_type="NEW_SELECTION", where_clause='"COUNT_ID" = 1')

#----------------------------------------------
#Step 5 - Create random points (for points no VM)
arcpy.CreateRandomPoints_management(out_path=workspace, out_name=RandomPoints, constraining_feature_class="Blocks", number_of_points_or_field="COUNT_ID")
arcpy.SelectLayerByAttribute_management("Blocks", selection_type="CLEAR_SELECTION")


# ------------------------------ Move all data points to closest intersection ----------------------------------------------
# Step 6 - To assure that the input streetfile is not divided into multiple substreets causing various streets to intersect but do not produce "cross" street intersections, streets will be "unsplitted"
in_features = StreetFile
out_feature_class = MergedStreets

# Run UnsplitLine
arcpy.UnsplitLine_management(in_features, out_feature_class)

# ----------------------------------------------
# Step 7 - Use "Intersect"-tool to find street intersection and receive output as points
in_features = MergedStreets
out_feature_class = Intersection

arcpy.Intersect_analysis(in_features, out_feature_class, "", "", "POINT")

# ----------------------------------------------
# Step 8 - Move pointsForVM to closest intersection
infc = pointsForVM
snapClass = Intersection
snapType = "VERTEX"
snapDistance = SearchRadius
arcpy.Snap_edit(infc, [[snapClass, snapType, snapDistance]])

# ----------------------------------------------
# Step 9 - Move "Points_Not_For_VM" to closest intersection
in_fc = RandomPoints
snapFeature = Intersection
snapChar = "VERTEX"
snapWidth = SearchRadius
```

```
arcpy.Snap_edit(in_fc, [[snapFeature, snapChar, snapWidth]])

#-----------------------------------------------
#Step 10 - Merge all masked data points into a new layer
arcpy.Merge_management('RandomPoints;Points_VM', AVM, "")
print ("Program successfully finished: AVM points are created")
```

D Contents additional ZIP-file