Thesis Report

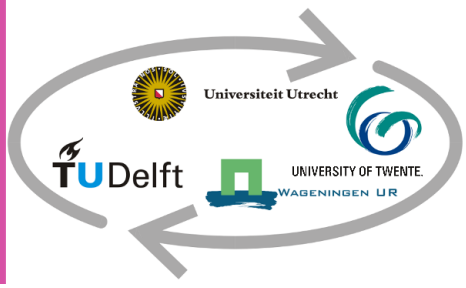
Journey reconstruction from social media posts

A study on data quality and privacy implications

Author
Betty Kronemeijer (4176081)
b.n.kronemeijer@students.uu.nl

Supervisor
Simon Scheider

Professor
Stan Geertman



Abstract

Studies of journeys can provide insights in the way in which movement influences the world socially and physically. However, journey data is difficult to obtain, as it entails private events over a longer period of time. Fortunately, smartphones with GPS functionality and social media usage have increasingly become a common part of daily life, for which reason geotagged personal posts from social media can serve as input journey data. Therefore, the objectives of this research are to assess whether or not social media personal posts are of sufficient quality for journey reconstruction, and to provide insight in the relation between quality of journey reconstruction and the associated privacy risk. This quality of personal posts for journey reconstruction is assessed by obtaining social media post histories from participants' Instagram accounts. From these post histories geotags and timestamps are extracted, that serve as input for individual journey reconstruction. This journey reconstruction is personally evaluated per participant based on six quality dimensions. The evaluation of journey reconstructions reveals that the better the quality of the journey reconstruction, the higher the risk of location privacy. Furthermore, temporal attributes are most influential on the quality of the journey reconstruction, because inconsistencies in the temporal attributes disarrange the sequence of stop places, that complicates the process of journey reconstruction.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Research problem

Due to the development of means of transport, people are nowadays enabled to travel larger distances in a shorter time than they used to in times before e.g. planes and trains were common modes of transport (Hall, 2005). There is a tight connection between this development of transport and the increase in individual mobility, which in its turn led to an increased scientific interest in travel behaviour (Høyer, 2000; Ellegård & Svedin, 2012; Büscher & Urry, 2009; Hall, 2005). Means of transport can be utilised for travel in either people's tied or in their untied time. These terms make a distinction between production and reproduction travel like a daily commute or a groceries trip on the one hand, and recreational travel on the other (Høyer, 2000). Within the domain of mobility in people's untied time, journeys can range from day trips that last under 24 hours, to cross-border travel in an unlimited spatial and temporal context (Hall, 2005). The common denominator of all journeys is that they start and end at some place, that there are possibly other stopping points in between, and that all these stops are connected by moves (Spaccapietra et al., 2008).

Studies of individual journeys can be applied to various fields in social science, and can uncover how movement can influence the world both socially and physically (Büscher & Urry, 2009). However, individual journey data is difficult to obtain, as it entails privacy sensitive data that should ideally be gathered over a longer period of time. Fortunately, volunteered geographic information (VGI) can offer a solution to this problem. Due to technological advancements, almost every modern smartphone is equipped with GPS functionality nowadays (Goodchild & Li, 2012;

Roick & Heuser, 2013; Spaccapietra et al., 2008). Furthermore, smartphones have increasingly become a common part of our daily lives, which stands in close relation to an increased use of social media (Roick & Heuser, 2013). Due to the opportunity to geolocate personal social media posts, social media posts can be used to gather journey data. Social media exists in large varieties on the web. OpenStreetMap can for example be seen as a social platform on which individuals try to gather data from an external source. On the other hand, there are social media platforms like Twitter or Instagram that enable personal posts. This distinction is important, as the former type of social media does not enable a reconstruction of personal events, whereas the latter does. Only personal posts are applicable to this research, because personal social media posts provide an interesting insight into individual journeys, as they originate from the private sphere of an individual. For this reason, data from personal social media posts reveal useful events from the perspective of research on personal journeys. However, the use of personal posts from social media could be dangerous, as it may threaten an individual's privacy when the quality of the reconstruction of events turns out to be high (Roick & Heuser, 2013; Scheider, 2019). The reason for this is that the better a reconstruction corresponds to actual events, the easier a person can be identified based on this reconstruction.

Due to the fact that data from personal social media posts is generated by non-professionals who are mostly unaware that their data might be used for research, quality assessment of this type of data is essential. In most cases of VGI, quality control can be done by comparing VGI datasets to a ground-truth dataset. For example by comparing OpenStreetMap to Google Maps, in which Google Maps serves as the ground-truth dataset (Goodchild & Li, 2012). However, this method is inapplicable for research that is based on data originating from personal social media posts. A ground-truth suggests that there is an objective external truth by which the quality of VGI can be measured. However, these social media posts are private for which reason there is no external truth, but only the experience of an individual by which the quality can be measured. Therefore, user studies are needed in order to assess the quality of data from social media geolocated personal posts. VGI data quality has yet been extensively researched by a variety of authors (Capineri et al., 2016; Jilani & Corcoran, 2013; Goodchild & Li, 2012; Mooney, Corcoran, & Winstanley, 2010; Cooper, Coetzee, Kaczmarek, Kourie, & Iwaniak, 2011; Flanagin & Metzger, 2008; Ali & Schmid, 2014), but it has only seldom been evaluated by qualitative methods (Goodchild & Li, 2012) or without reference dataset (Barron, Neis, & Zipf, 2014). Furthermore, Chen et al. (2016) and Kisilevich, Krstajic, Keim, Andrienko, and Andrienko (2010)

have researched the possibilities of the usage of social media geotags from personal posts for pattern discovery. However, both articles do not assess the geographic data quality in a qualitative and personal manner. Therefore, to the best of my knowledge and survey, there is not any work that has a comparable focus and corresponding method.

Therefore, in order to assess the feasibility of using social media based VGI to reconstruct personal events, this research is aimed at investigating whether social media geolocated posts contain enough information of sufficient quality to reconstruct a journey. This approach uses journey reconstruction as a proxy for the assessment of the quality of social media data from personal posts. Furthermore, as the research is based on personal data, geo-related privacy issues are relevant to both the process and the goal of the research, because when the quality is high, the privacy risk is also high. For that reason, an underlying goal of the investigation is to assess the security risks that are consequential of the use of personal geodata.

1.2 Research objectives

Within the scope of this research are multiple research objectives. The main objective is to assess the quality of personal geolocated posts as a source of VGI for the reconstruction of journeys. Another objective is to investigate privacy issues that are related to the processing of personal social media data.

1.2.1 Research questions

The main objective will be achieved by reconstructing journeys based on social media data from personal posts. The quality of the eventual journey reconstruction will be personally assessed by the producers of the social media data. In order to investigate this research problem, the following research question is addressed:

To what extent is the quality of geolocated social media posts sufficient for the reconstruction of journeys?

The research question will be answered based on the following subquestions:

- *What is the spatial data quality of personal posts in social media?*

  Whether or not social media geolocated posts are of sufficient quality, depends on how well they comply with the quality dimensions of geodata. These dimensions include among others properties like completeness and spatial accuracy (Jilani & Corcoran, 2013; Goodchild & Li, 2012).

- *How to reconstruct journeys from social media personal posts?*

  Each journey has a start and an end location, and one or several stops in between (Spaccapietra et al., 2008). This subquestion aims at defining the distinctive components that reconstruct a journey. It also addresses the question of how to handle spatial and temporal resolution for the representation of a journey.

- *What GIS methods are applicable to reconstruct a journey?*

  This subquestion aims at reviewing and finding the most suitable technologies that are able to process social media data in such a way that it can be used to reconstruct a journey. This includes the technical aspects of the automatic reconstruction of a journey, as well as the clustering methods that might be used for the categorisation of trajectory data into multiple journeys.

- *What privacy risk of social media data is incurred by the quality of it?*

  If social media data quality is of sufficient quality to reconstruct journeys, it might be a serious threat to an individual's privacy. Therefore, the aim of this subquestion is to assess to what extent social media data can be a threat to an individual's privacy. For example, social media data might have a resolution that is too coarse to extract a person's home location. In this case, a low quality incurs a low privacy risk.

1.2.2 Scope

The research is primarily focused on assessing data quality and providing a suitable method for journey reconstruction. Any collection or processing of data that does not contribute to this goal is therefore redundant. For example, it is possible to measure path similarities based on metric distance functions or cluster and pattern methods (Long & Nelson, 2013). Such methods calculate the extent of similarity between a number of movement patterns. This could for example

be a good method to implement when wanting to identify frequently used trajectories by certain groups of people. However, calculating path similarities does not fall within the scope of this research.

In short, individual movements and their properties fall within scope. This also means that individual movements against the background of larger group dynamics/characteristics fall out of scope. The reason for this is that the primary objective is to assess whether or not the quality of social media geolocated data is sufficient for the reconstruction of personal journeys. Looking into the social background of these journeys does not contribute to this objective.

Furthermore, out of multiple available methods for retrieving the location of a social media post, only those that provide a readily available geoposition fall within the scope of this research. This means that e.g. the geocoding of addresses that can be retrieved from social media posts by the use of natural language processing, are outside the scope.

1.3 Relevance

1.3.1 Societal relevance

The goal of this research is to investigate whether or not social media data is of sufficient quality to reconstruct journeys. The result of this main question could have benefits for multiple other fields. If social media data turns out to be of sufficient quality and the journey reconstruction method yields satisfying response, the method can be used for other applications too. For example, harvesting social media data could help the tourism sector in identifying popular sites and routes for different demographic groups. Another application that it could be used for is to trace migration movements, or to identify routes taken by refugees.

An additional motivation for this research subject, is the handling of privacy sensitive data. Due to the rise of social media and also because of the new European regulations on privacy (GDPR) (European Commission, 2018a), it is important to review and reveal the implications of posting a location online as many users do so without knowing the consequences (Stefanidis, Crooks, & Radzikowski, 2013). Assessing the privacy risk is difficult as the quality of journey reconstruction is not yet known, and quality and risk are highly related. A high quality of journey reconstruction will lead to a bigger risk of location privacy. Therefore, in this research journey

reconstruction serves as an attack strategy for the assessment of the privacy risk that is related to geolocating personal posts.

1.3.2 Scientific relevance

The data quality of Volunteered Geographic Information is a subject that has been elaborately reviewed by many authors (Jilani & Corcoran, 2013; Mooney et al., 2010; Cooper et al., 2011; Flanagin & Metzger, 2008; Goodchild & Li, 2012; Spielman, 2014; Haklay, Basiouka, Antoniou, & Ather, 2010; Hochmair & Zielstra, 2012). Some of these sources state that the more people contribute to the creation of the same spatial information, the less erroneous this information will become (Haklay et al., 2010; Long & Nelson, 2013). However, in the case of personal posts, there is only one source who can evaluate if the reconstruction of events is correct. This indicates that there is no ground-truth dataset, on which most methods of VGI quality control are based (Jilani & Corcoran, 2013; Mooney et al., 2010; Cooper et al., 2011). Because of the absence of a ground-truth dataset, the personal evaluation of the journey reconstruction, and the fact that there is only one person who can evaluate the quality of a reconstruction, this research requires a different approach to assess quality.

Chapter 2

Theoretical background

2.1 Volunteered Geographic Information

For many projects, research or applications, it is necessary to gather data. Depending on the objective of the research, the amount of data that needs to be gathered can vary from small amounts of easily retrievable data to large quantities of data that is difficult to collect. There are four general ways in which to obtain spatial data (Zheng, Zhang, Xie, & Ma, 2009):

- Use the currently available spatial data. The disadvantage to this method however, is that the data is not necessarily fit-for-purpose, and that payment might be needed in order to obtain the data.

- Retrieve the data from professional survey or mapping agencies. The advantage of this method is that the data will be fit-for-purpose. However, contracting professional corporations might turn out to be too costly for the project's resources.

- Use data from free map providers. This method is free of charge, which is beneficial for a project's budget. However, the data is not necessarily fit-for-purpose and of sufficient quality.

- Another option is to collect the data manually or by crowd-sourcing. This is a way in which to customise the data collection to be exactly fit-for-purpose. However, this method is usually more time-consuming than previously discussed methods.

There is no one solution that is the best option, as the method in which to obtain data depends on the use. The preferred method also depends on other factors like e.g. the available time

and budget. Projects that either do not have a sufficient budget, or that need too much data to be gathered or need a specific type of data that is hard to collect through the regular means of data collection as described in the first three methods, are generally prone to resort to the last option of manual or crowd-sourced data collection. Fortunately, due to the rise of the internet and the availability of smartphones, many people continuously generate data by the mere use of such devices. Such user-generated data is in general freely available, and can be filtered in such a way that it can become fit-for-purpose for numerous applications (Goodchild, 2007).

The method of data retrieval that is used and assessed in this research is the fourth option, as personal posts from social media can best be retrieved by a crowd-sourcing approach. User-generated data with geographical attributes is generally referred to as Volunteered Geographic Information (VGI) (Granell & Ostermann, 2016; Goodchild, 2007). The quantities in which this type of data is created keeps growing, as an increasing number of people continuously carry their smartphone with automated GPS traces with them, expanding the possibilities of harvesting VGI (Granell & Ostermann, 2016; Roick & Heuser, 2013; Spaccapietra et al., 2008; Van Exel, Dias, & Fruijtier, 2010). These possibilities are boosted even more by the intensification of social media use and the emergence of Location Based Social Networks (LBSN), in which people can geotag their social media posts (Roick & Heuser, 2013). For that reason, the crowd-sourcing approach is the most applicable method to obtain personal posts.

2.1.1 Social media as a source of VGI

Social media is a very potential source of geographic information, especially due to the increasing use of smartphones (Roick & Heuser, 2013). Before the emergence of spatially enabled social media platforms, it was quite difficult to obtain journey data (Roick & Heuser, 2013; Spaccapietra et al., 2008). Reason for this is that journey data is quite privacy sensitive, and that giving all respondents GPS devices is costly and slow. Obtaining journey data through social media is both quick and free of charge (Goodchild, 2007; Roick & Heuser, 2013), and as anonymisation techniques can be applied to mask participant's identities, VGI from social media platforms is a good solution to obtain journey data.

There are two ways to add geospatial components to social media data. First, there is the possibility to add a location to a digital medium like a photo. The addition of location to some

type of data is called geotagging (Roick & Heuser, 2013; Turner, 2006). This merely locates the digital post of a user, and not necessarily the user itself, and is therefore seen as a locative medium (Thielmann, 2010). Besides locative media, there is also the concept of mediated localities. This perspective on adding a geo-component to social media data embodies the process in which social media users can be located through time based on their social media use or check-ins (Roick & Heuser, 2013; Thielmann, 2010).

Where locative media is more about the locational annotation of digital posts, mediated localities seek to trace an individual's movement through the offline world based on digital personal posts (Thielmann, 2010). However, before being able to work with any of these two concepts, a location needs to be extracted from digital social media posts. There are three ways in which to obtain a location from social media (Roick & Heuser, 2013). First of all, a location of an individual can be gathered through GPS. This is a relatively reliable, accurate and easy method, as most smartphones are equipped with GPS abilities. However, this method is not always feasible, as there are monetary and privacy issues associated with individual GPS tracking (Roick & Heuser, 2013; Sui & Goodchild, 2011; Stefanidis et al., 2013). Another option to gather location from social media is to get the geotag that some social media networks have available on their website. However, sometimes a geotag is not available on the targeted social media platform, or a user does not geotag their posts. For such cases there is a final option, in which the researcher can geocode addresses or locations that are mentioned in the text of a social media post (Roick & Heuser, 2013). Unfortunately, this method is quite difficult, as there are multiple words that can be used to describe the same location and there is no actual coordinate tag. For example, when people always go to the same restaurant, in a social media post they might not use the exact name of the restaurant, but name it "the restaurant where we always celebrate my birthday".

2.2 Conceptualisation of journeys

Journeys consist of multiple common aspects. Each journey has a departure and arrival point, and possibly multiple stops and moves in between. These aspects will be elaborated on in a later section, but first it is important to substantiate the locational concepts on which journeys are based.

2.2.1 Space versus place

Location is an important aspect for the reconstruction of journeys. Social media users can tag their location that is subsequently transformed into a GPS track, based on which their journey can be reconstructed. The difficulty lies in the fact that location is not just the coordinate of a certain position on earth, also known as *space* (Roick & Heuser, 2013), because besides the notion of space, there is also the concept of *place*. A location is a place when it is referred to as a name or description, without a specific position being mentioned. In other words, space is a continuous concept and is the same for everyone (Taylor, 1999; Spaccapietra et al., 2008; Long & Nelson, 2013). On the other hand, two individuals can be positioned in the same space, but still experience another place, as their experience of the space is different.

Due to this different notion of space and place, quality of a reconstructed journey is not measurable in one of the three ways defined by Goodchild and Li (2012) in section 2.3.1. The quality of the reconstruction can only be assessed by the person who conceptualised the places they visited in a journey. In other words, the data of a journey reconstruction cannot be right or wrong in the objective sense, for which reason it is not possible to use either the crowd-sourced or social approach for improving the positional accuracy, as both of these approaches assume that the quality can be measured from an objective perspective (Goodchild & Li, 2012; Taylor, 1999). Therefore, a distinction can be made between spatial and placial accuracy. Spatial accuracy then refers to the accuracy of the coordinates of the geotag, whereas placial accuracy refers to the accuracy of the placename of the geotag.

2.2.2 Tracks versus journeys

Conceptually similar to the difference between space and place is the difference between tracks and journeys. Coordinates retrieved from i.e. smartphones can be stored in a coordinate log along with their respective time-stamp, describing an individual's spatial location at a certain time. The result of sequentially connecting and displaying these coordinates is called a track (Hu et al., 2013; Zheng et al., 2009; Spaccapietra et al., 2008). Tracks are coordinate sequences that can represent the movement of various entities like humans, animals, packages or vehicles. These entities can all be represented by a point feature (Spaccapietra et al., 2008).

Tracks do not yet conform to the full definition of a journey, as they are spatial point samples of journeys. Semantic information must be extracted from track data in order to be able to

reconstruct a journey (Guc, May, Saygin, & Körner, 2008; Dimond, Smith, & Goulding, 2013; Spaccapietra et al., 2008). For example, a track only displays the general path that an individual has taken and is therefore a sequence of spaces. This does not yet say anything about where and how long the individual has taken stops in between, or about how this individual experienced the place.

2.2.3 Journey aspects

Journeys consist of different types of events. The movement of an individual can be either stationary or progressive. Events where the individual is stationary are called stops. Stops are places, because the subject does not move outside of a certain defined threshold during the temporal interval of the stop (Spaccapietra et al., 2008). The threshold within which a subject is allowed to move within the boundaries of a stop depends on the spatial resolution of the application. Depending on the resolution of overall journey and the research context, a stop could be within the bounds of a city, a neighbourhood or even within the boundaries of a building. Events where the movement of an individual is not stationary are called moves.

Furthermore, each journey has a departure (start) and an arrival (end) point (Long & Nelson, 2013). Such start and end points are also stop places, but the conditions that decide which stops are marked as start and end points depend on the purpose of the application, and on the dataset itself (Spaccapietra et al., 2008). Take for example the daily commute to work. A possible departure point is the home location of the subject. The next question is where the trajectory should end, as there are multiple possibilities: for example, the trajectory either ends when the subject arrives at work, or it ends when the subject arrives back home. Such a choice depends on the issue that the research is applied to. As figure 2.1 shows, all components together are the basis of a journey reconstruction.

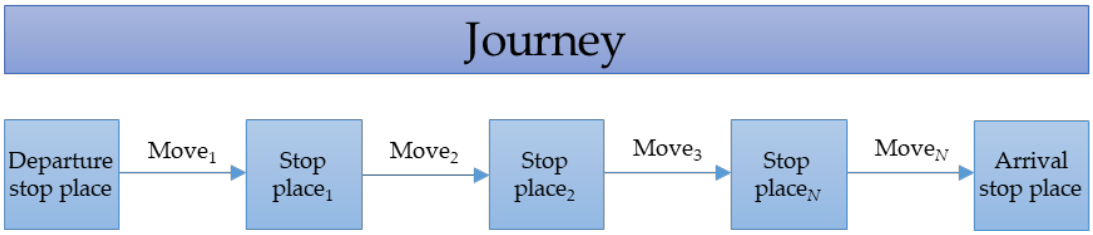


Figure 2.1: Components of a journey

2.3 Quality of VGI

As VGI is gathered by a large heterogeneous group of individuals, it is important that the quality of the resulting data is controlled in a reliable way by standardised quality measures (Spaccapietra et al., 2008; Goodchild & Li, 2012; Ali & Schmid, 2014; Xia, 2012). There are three quality components: intrinsic, extrinsic and pragmatic quality (Bordogna, Carrara, Criscuolo, Pepe, & Rampini, 2014; Criscuolo et al., 2016). Extrinsic quality measures the contextual factors to the generation of the data. This includes e.g. the trustworthiness of the volunteers and their (scientific) background. Extrinsic quality is irrelevant to this research, because it is only used for the assessment of external attributes, as is the case with e.g. OpenStreetMap. This research investigates personal posts, which means there is no external attribution for which the extrinsic quality should be measured. On the other hand, intrinsic quality measures the validity of the information itself based on quality dimensions of spatial data. This type of data quality is relevant to the research, as it measures the internal validity of the personal posts, which amounts to the ability to reconstruct journeys based on this data. Lastly, pragmatic quality describes the fitness for use of the dataset. This means that pragmatic quality is dependent on the application of the data (Bordogna et al., 2014; Criscuolo et al., 2016). Pragmatic quality is also irrelevant from the perspective of this research, as there are multiple purposes for which personal posts could be used. The discussion of possible applications of personal posts does not fall within the scope of this research. For the remainder of this chapter, only the intrinsic quality and quality dimensions of VGI will be discussed.

2.3.1 Quality assurance of crowd-sourced spatial data

The advantages of VGI are numerous, as VGI is almost always freely available in large quantities that cover a wide variety of scientific fields (Goodchild, 2007; Goodchild & Li, 2012). A problem however, is the quality control of VGI. The reason for this is that volunteered (geographic) information is generated by a largely untrained audience, that is mostly unaware of the fact that they are producing data (Granell & Ostermann, 2016; Goodchild & Li, 2012). In order to be able to produce reliable results based on VGI, it is essential to establish good quality assessment metrics (Senaratne, Mobasheri, Ali, Capineri, & Haklay, 2017).

The spatial component of VGI makes quality control more complex than the quality control of other types of open data. For the purpose of standardising quality control of geographical open

data, researchers have come up with five components of spatial data quality (Jilani & Corcoran, 2013; Goodchild & Li, 2012). These components for quality control state that the data must be positionally accurate, that the attributes to the data must be correct, that data entries must be logically consistent, that the data must be complete, and that the lineage is known (Jilani & Corcoran, 2013; Criscuolo et al., 2016; Zipf, Mobasheri, Rousell, & Hahmann, 2016). Besides these five components, spatial and temporal resolution are also considered to be factors that indicate the quality of geographical data (Jilani & Corcoran, 2013; Goodchild & Li, 2012).

Journey reconstruction is specific application of VGI, which influences quality control. Journeys are reconstructed based on *places* that are derived from personal posts, whereas the reconstruction shows *spaces*. The place that is tagged in some posts might not be the exact space that the person was in. There are multiple reasons for this, as people might for example geotag their posts in an unmatching spatial resolution. Another reason for uncertainty of place identification might be that there are multiple synonyms for the same place, which inflicts spatial uncertainty (Scheider & Janowicz, 2014). In general, the greatest difficulty lies in the matching of a place to a space (Scheider & Janowicz, 2014). However, in the case of social media geotags, this matching is usually already done by the platform itself.

The aforementioned quality dimensions assume a homogeneous and consistent quality throughout the dataset, which is not the case with VGI. The reason for this difference is that VGI is collected by different individuals with varying backgrounds, intelligence and reasons for generating data (Van Exel et al., 2010). Goodchild and Li (2012) and Ali and Schmid (2014) describe three approaches to quality assurance of VGI:

- *The crowd-sourcing approach*

  The crowd-sourcing approach to quality control of VGI assumes that the problem of VGI can be solved by its cause: the heterogeneity of the people who create it. This approach applies the concept that the more people evaluate and assess a dataset, the less faulty this dataset will be (Goodchild & Li, 2012; Flanagin & Metzger, 2008). However, this does not always seem to work in practice, as the people who voluntarily assess a dataset also have heterogeneous methods and opinions, because they are not bound by the methods of an organisation. Nevertheless, studies have shown that the positional accuracy of objects increases along with the increase of assessors (Goodchild & Li, 2012).

- *The social approach*

  Quality control by means of the social approach is based on the hierarchy of VGI contributors. Studies have pointed out that the majority of VGI is generated by only a small percentage of the overall amount of individuals that contribute to VGI (Goodchild & Li, 2012). The individuals that contribute the most to the dataset can function as so-called gate-keepers for all other contributors that commit to the dataset, and are therefore awarded the authority to maintain and control contributions on the site (Goodchild & Li, 2012; Spielman, 2014). Furthermore, the social approach can also be implemented in methods that measure quality based on e.g. the amount of contributors (Neis, Zielstra, & Zipf, 2011).

- *The geographical approach*

  The geographical approach to quality control relies on existing geographic rules. For example, if a digital medium that represents a cafeteria gets geotagged within the bounds of a lake, the chances are likely that this specific entry is incorrect (Goodchild & Li, 2012). Such a geographic approach to quality assurance can be applied through machine learning, where the algorithm trains on recognising patterns in geographic data (Jilani & Corcoran, 2013).

The most important difference between these three approaches to quality control, is that the first two are executed by people, while the geographic approach can be carried out by intelligent computing. The most appropriate method of quality control depends on the project that it is applied to. However, none of these approaches fully apply to the research problem. The crowd-sourcing and social approach presume there is an external truth that can be viewed and assessed by every individual, with some exceptions. Nevertheless, personally assessing the journey reconstruction might be seen as a very limited perspective on the social approach, as only one person is suitable to assess the quality of something that only they contributed to themselves. Furthermore, also the geographical approach might be applicable to this research to some extent, as machine learning algorithms can be applied to cluster the data from personal posts. This will be elaborated on in method section 3.2.2.

2.3.2 Quality dimensions for journey reconstruction

There are many quality dimensions in circulation for many different purposes. For the intrinsic quality of geographical data, positional accuracy, correctness of attributes, logical consistency, completeness and lineage are the most used (Goodchild & Li, 2012; Criscuolo et al., 2016; Jilani

& Corcoran, 2013; Van Exel et al., 2010; Barron et al., 2014). For the sake of research into journeys, spatial and temporal resolution are added as a complementary quality dimension (Hall, 2005; Spaccapietra et al., 2008).

These quality dimensions are designed for the measurement of regular spatial data, and apply differently to VGI than to regular spatial data. Therefore, in this section, quality dimensions that are used to measure the quality of personal posts for journey reconstruction, and their correspondence to general quality dimensions for regular spatial data is described.

Spatial accuracy of reconstructed stop places

This quality dimension corresponds with positional accuracy. Due to the notion of spatial uncertainty, only the spatial accuracy of stops is measured, whereas the placial accuracy is outside the scope of this quality dimension. Positional accuracy in general measures the correspondence of the tagged locations to the actual location (Goodchild & Li, 2012; Cooper, Coetzee, & Kourie, 2012). It is argued that the more volunteers contribute to the pinpointing of a location, the better the positional accuracy becomes (Neis et al., 2011), which coincides with the crowd-sourcing approach of VGI quality assurance. The concept that assumes a better quality due to a higher number of contributors is called Linus' Law, after the creator of Linux (Haklay et al., 2010; Goodchild & Li, 2012). However, Linus' Law is not entirely applicable here, as there is only one person that can assess the accuracy of the tagged position. From the perspective of journey reconstruction, spatial accuracy focuses on the accuracy of the tagged *space*.

Spatial resolution of reconstructed stop places

Depending on the chosen spatial resolution, an error of spatial accuracy is not necessarily fatal for the reconstruction of a journey. Especially when the difference between the actual position of the user and the position of the geotag is small enough to be covered by the granularity of the reconstruction. Therefore, inconsistent spatial resolution in geotags does not necessarily negatively influence the satisfaction of a journey reconstruction.

Spatio-temporal accuracy of stop place posts

Spatio-temporal accuracy of stop place posts is adapted from the quality dimension *correctness of attributes*, but is only focused on the spatio-temporal attributes. The original quality dimension *correctness of attributes* is focused on the part of data quality where the non-spatial information

of the dataset should be an accurate reflection of the event in the real world (Cooper et al., 2012). Especially in the case of VGI, correctness of attributes is a frequent point of discussion, as a heterogeneous group of people with heterogeneous perspectives on non-spatial events in the real world are responsible for the addition of these attributes (Van Exel et al., 2010; Neis et al., 2011). However, for a journey reconstruction, non-spatial attributes except for time attributes are irrelevant, for which reason the quality dimension is adapted to fit the estimation of the quality of journey reconstruction.

Spatio-temporal accuracy is is a different notion of accuracy than the spatial accuracy in the previous quality dimension, because the this quality dimension measures the influence of spatio-temporal accuracy on journey reconstruction, as opposed to only the spatial accuracy of geotags themselves. For example, footprint mismatch errors might occur when the location of an artefact is used for the geotag instead of the user's current position at the time of posting (Roick & Heuser, 2013; Hochmair & Zielstra, 2012). The severity of this footprint mismatch error increases along with the size of the discrepancy between the time of the event and the time of posting about the event. Another cause for discrepancies between the time of the original event and the time of posting about this event occur when users implement throwbacks in their respective post histories. A throwback is a post about a historical event (Kahle, Sharon, & Baram-Tsabari, 2016). Therefore, the use of throwbacks disturbs the sequence of posts, and simultaneously corrodes the spatio-temporal accuracy of stop place posts.

Logical consistency

Logical consistency is a quality dimension that represents the way in which all data entries follow the same structure and relationships (Zargar & Devillers, 2009; Cooper et al., 2012). This is especially difficult for VGI, as each individual adheres to different methods. For social media based reconstruction of journeys however, this is not an issue. The way in which individuals add their data to their social media account is structured in such a way that logical consistency is already secured to some extent, as every user has to enter attributes the same way. A possible threat to logical consistency is when people deviate from the indicated procedure when they e.g. do not geotag their entries, but instead mention their location in the caption (Roick & Heuser, 2013). However, this falls out of the scope of this research, and is therefore not used for this research.

Completeness and precision of reconstructed stop places

Completeness as a quality dimension is used to measure how well the data represents the totality of events. In the case of journey reconstruction, it measures the accuracy of correct identification of an event or place in a journey. This quality dimension has two components: precision and completeness (recall). Precision indicates the percentage of correctly identified stop places in a journey. The definition of the term precision is in this case taken from Information Retrieval. On the other hand, recall counts the errors of omission. Or in other words: how many events in the real world were failed to be taken up in the dataset (Devillers & Jeansoulin, 2010). A difficulty that especially occurs with VGI is that people have different understandings of what is and what is not supposed to be in a dataset (Cooper et al., 2012).

Precision measures the events that are considered to be correct in the reconstruction. This means that the event behind a place post corresponds to an actual event. On the other hand, recall measures the places where an individual has been without geotagging it.

Lineage

Lineage describes the development of a dataset, including the people that are responsible for each stage of the life cycle of the data; collection, compilation and end-product (Zargar & Devillers, 2009; Cooper et al., 2012). Describing the lineage of VGI is a difficult process, as each contributor has different methods of acquiring and assembling data. Furthermore, it is challenging to define the people responsible for the data, as the final product is a combination of different sources. This is complicated even more by the option of anonymous data entry in some cases. However, lineage for social media data is less complicated, as the source of the data is apparent because all data is linked to a personal account. Therefore, this quality dimension is not considered in the evaluation of journey reconstruction.

Quality of reconstructed journeys

There are different types of journeys, that cover varying spatial and temporal extents. The spatial and temporal attributes of journey data can therefore be used to define the type of trip that was conducted. In practice, there is not a single definition of what counts as which type of trip (Hall & Page, 2003; Hall, 2005). A good example of this is the definition of a day trip that was used by the Western Australian Tourism Commission (WATC): "a trip taken mainly for pleasure which lasts for at least four hours and involves a round trip distance of at least 50 km.

For trips to national parks, state forests, reserves, museums and other man-made attractions the distance limitation does not apply." (Hall, 2005). It is unknown where the WATC based these parameters on, and this definition of a day trip is not standardised. Therefore, the boundaries of journey types are subject to arbitrary parameters that a researcher deems appropriate for a project (Hall & Page, 2003; Hall, 2005). One of such parameters is generally accepted, as trips that last for longer than 24 hours are considered to fall under the denominator of journeys. Below the boundary of 24 hours, every movement is considered to be a day trip (Hall, 2005). Figure 2.2 displays what types of trip can be made along the spatial and temporal scale.

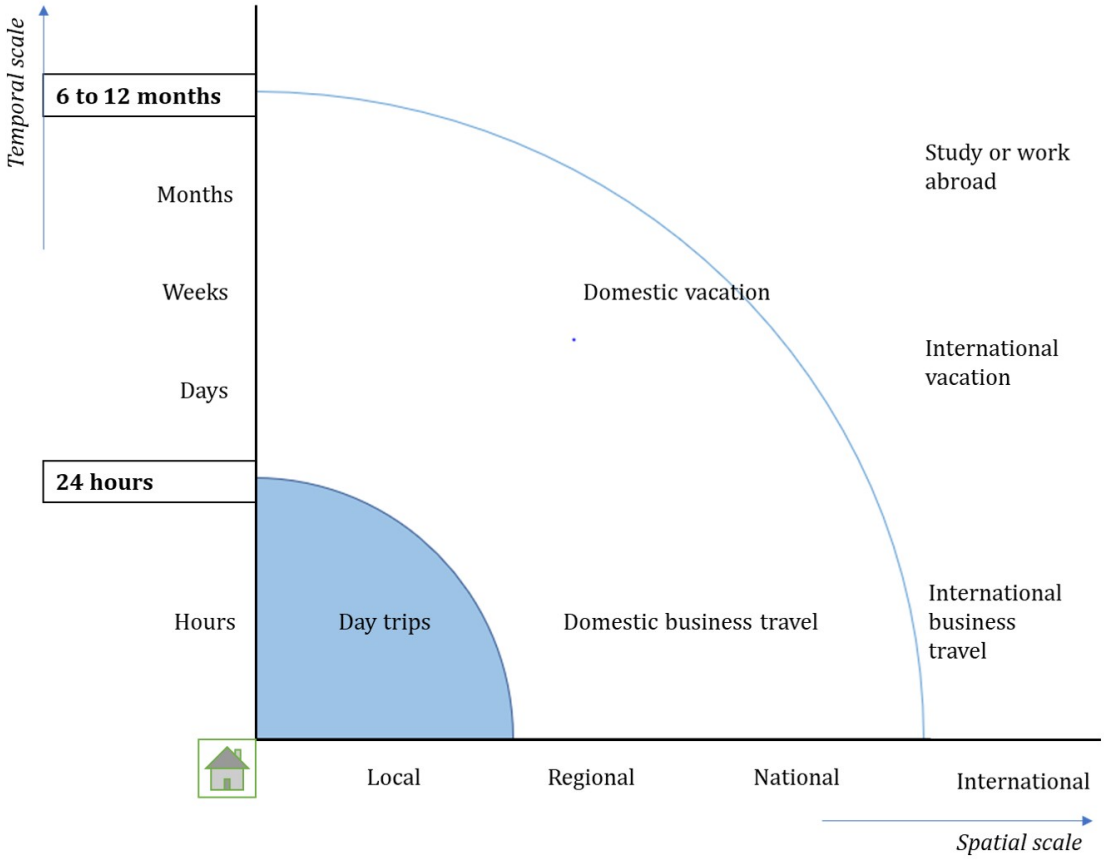


Figure 2.2: Approximation of types of trips on a spatio-temporal scale (Hall, 2005)

For the reconstruction of journeys, it is important to be aware of the different extents of journeys. Furthermore, track data is collected in an unorganised manner, as new points are only added when an individual posts something at their own convenience. Therefore, track data from social media personal posts is sparse and irregular by nature. These properties complicate the clustering of track data which will lead to reconstructed journeys, as there is no standard extent

of journeys and track data is irregular. The quality of reconstructed journeys therefore depends on how well a journey reconstruction process is able to correctly assign each place post to their corresponding journeys.

2.4 Privacy implications

Privacy is an important aspect related to the revelation of location data from personal posts. Concerns with respect to privacy issues have both changed and increased since the advancement of computers from approximately the 1960's onward (Curry, 1999). Before the use of computers, privacy was mostly the ability to be left alone. This changed when the computer came into existence, as privacy then became more focused around the ability to control what data is known (Curry, 1999). The increasing use of computers is connected to an increasing ability to geolocate individuals and households (Curry, 1999; Duckham & Kulik, 2006; Roick & Heuser, 2013). This development is related to yet bigger concerns with respect to privacy issues, as for example Duckham and Kulik (2006) state that an individual's exact location is an identifying property to such measure, that it is an even more unique identifier than the individual's genetic profile. Based on this statement, privacy can be seriously threatened by the use of geotags on location based social networks.

2.4.1 Types of spatial privacy

Roick and Heuser (2013) identify four different types of privacy threats that are related to the use of location extracted from social media personal posts. Two of which are directly involved with the location of a social media user itself. First, *location privacy* is a type of privacy issue where the location of one social media user at a certain time is revealed. On the other hand, a threat to *absence privacy* can reveal the absence of a social media user at a certain place and time. In this research, location privacy is the most important aspect of spatial privacy.

2.4.2 Risk of exposed location privacy

Many risks have been identified that are connected to the violation of an individual's location privacy. For example, an individual could become subject to stalking or other types of unwelcome visiting (Krumm, 2007). Regardless of these possible dangers of location exposure, people do not highly value the protection of their location privacy sensitive data (Danezis, Lewis, & Anderson, 2005; Krumm, 2007). Furthermore, when people feel like they are in control of whether or

not their location is shared along with their social media personal post, and they feel like they share it with people they are acquainted with, they generally do not take privacy concerns into account (Iachello et al., 2010). In general, the underlying danger seems to be that users share personal information without explicitly knowing that it could seriously harm their location privacy (Stefanidis et al., 2013; Krumm, 2007).

2.4.3 Attack strategies

Attack strategies entail malicious attempts to retrieve location data without having been given consent. There are multiple ways in which an attack strategy could be executed, which depends on the intention of the attacker. Attack strategies could for example be intended to identify an individual itself, or could be focused on finding their home location (Scheider, 2019; Li et al., 2014). Besides the use of attack strategies for criminal intent, attack strategies could also be used for research purposes, in order to identify possible risk of privacy breach. In the case of this research, reconstructing an individual's journey is an attack strategy, intended on registering the individual's history of whereabouts.

Chapter 3

Methodology

This research uses both quantitative and qualitative methods. The harvesting and processing of social media geolocated data is in its essence a quantitative procedure. However, as the reconstruction of each journey will be evaluated based on personal opinions, qualitative methods are used for the assessment of the quality of the journey reconstruction.

3.1 User study

In order to investigate whether or not social media geolocated posts are of sufficient quality to reconstruct journeys, the first essential step is to find participants that are willing to share personal data that is generated through the use of their respective social media accounts. Due to the new European General Data Protection Law (GDPR), all participants are directed to an informed consent form in order to ascertain the protection of all parties' personal data.

3.1.1 Participant sample

In total, 17 people were willing to participate in this research. Participants were selected based on their travel experience and their use of geotags. Other possible characteristics of participants are irrelevant for this research, and therefore only travel experience and use of geotags were considered as prerequisites for participation.

Out of the 17 initial participants, only 15 filled in the questionnaire about the journey reconstruction. Therefore, the results in this research are based on 15 participants. Of these 15

participants, 11 are male and 4 are female. 12 participants are from the Netherlands, whereas 3 participants are from elsewhere in the world. Furthermore, the participants' age ranges from 18 to 29 years old.

3.1.2 Securing personal data

The GDPR is a relatively new European regulation. It was accepted in 2016, after which it was enforced in May 2018 (European Commission, 2018a). The intentions of the GDPR are to harmonise the various privacy regulations member states of the European Union maintain, and to synchronously protect all EU citizens' personal data. In order to do so, the GDPR regulates the processing of any type of personal data that is related to an individual living in the EU (European Commission, 2018b). The collection, structuring, storage and alignment or combination of personal data falls within the definition of data processing (European Commission, 2018c), which means that the methodology of this research falls within the regulations of the GDPR. Therefore, the regulations of the GDPR are respected.

Before accessing any participant's personal data, it is important to establish informed consent between the researcher and the participant. In order to properly inform the participants, an informed consent form is written (see also: `https://refugeestorymaps.sites.uu.nl/informed-consent-form/`). Such an informed consent form has to contain various predefined pieces of information, that are designed to ascertain that the participants are aware of what their data is used for and that the researcher ensures that their personal data is protected (European Commission, 2018d). First of all, the respondent has to be aware of who the researcher is and for what purpose their data is requested. Another important point is the storage of the respondent's personal data, how long this data is in storage for and to inform the respondent on the possible risk of data breaches. Furthermore, it is important to inform the respondent on how they are referenced to in the final product.

In order to secure the safety of every respondent's personal data, all data from the respondent's social media pages and all information that is acquired through questionnaires has to be stored in a secure location. In the case of this research, this is done by storing all personal data on a secured storage drive that is provided by Utrecht University. All data is only held onto until the end of the research, by which time it will be completely deleted.

3.1.3 Data collection

After finding suitable participants and securing their personal data, the next step is to find social media platforms and corresponding APIs that are suitable for the retrieval of social media data from the participants.

API selection and use

There are many social media platforms that enable geotags in their posts, and would therefore be suitable for this research. Such social media platforms are e.g. Twitter, Flickr, Four-Square and Panoramio (Kisilevich, Krstajic, et al., 2010; Siła-Nowicka et al., 2016; Yin, Cao, Han, Luo, & Huang, 2013; Zheng, 2012). For this research, geotags from Instagram's API are used (Instagram, 2018). The reason for this is that Instagram appears to be a widely used social media platform among travellers, as people can only post photos on this platform. This is important, as the goal of this research is to assess the quality of social media geotags of people who travel. Furthermore, the general content of API response is quite similar between various social media platforms, such as for example Twitter and Facebook, for which reason the general method could also be applied to other social media platforms besides Instagram (Xie, Xia, Grinberg, Schwartz, & Naaman, 2014). A problem with the Facebook API for example, is that it's use is restricted. Facebook shows a user's complete history of posts, including the tags and timestamps, but only for the account of the person who calls the API. This means that access tokens to other users' accounts are not provided anymore if you do not have an authorised application. Because of such restrictions, the method in this research is centred around Instagram's geotags.

The first step in being able to retrieve users' posts from a social media platform, is to obtain access tokens that are linked to their accounts. A step-by-step description of how participants can generate these access tokens can be found on `https://refugeestorymaps.sites.uu.nl/participation/`. These access tokens can then be used to access the users' profiles, and to obtain their posts and the appurtenant attributes. Among these attributes are a unix timestamp and a geolocation. Both these attributes will be stored in a database. The APIs return both geographic coordinates and the name of the tagged location (e.g. "Eiffel Tower, Paris"), which will be added to the database as well (Instagram, 2018).

A problem with the Instagram API is that it only allows for the retrieval of the last twenty posts. This is not necessarily a problem for the course of the research, as a journey reconstruction

can still be made based on twenty geotag-timestamp sets. It will however be more difficult to extract the home location based on this restricted amount of geotags, which makes it more difficult to assess the privacy risk that is associated with geotags.

Data structure

The design of the database depends on the different factors that influence an object's track, and on the purpose of the measurement of the track. For e.g. bird migration, it is important to not only store the birds' spatio-temporal path, but also to monitor their height, weight, bodywarmth etc., depending on the research objective (Spaccapietra et al., 2008). However, all such properties are of no interest to this research. The reason for this is that the only necessary features of a track are in this case the location and timestamp, as it is the quality of the location data that is measured.

3.2 Journey reconstruction

The reconstruction of the participants' journeys is automated. The reason for this is that if the method by which the reconstruction is automated works, it can be applied to other or larger quantities of journey data. By means of Instagram's API, sets of temporal and spatial data points will be stored in a database. Together, these points from raw geocoded posts form a track (Hu et al., 2013). When these raw geocoded posts have been given semantic information, or in other words when geocoded posts are identified as stops or moves of a journey, they become a trajectory. Trajectories serve as input for the journey reconstruction. All python scripts that are used for data gathering and journey reconstruction can be found on https://github.com/bkronemeijer/GIMA_thesis_journey_reconstruction.

For the design of the journey reconstruction, it is first of all important to be aware of the stop places (Spaccapietra et al., 2008). Some of these stop places mark the start and end point of journeys. The post histories of participants likely contain multiple journeys, and therefore multiple start and end stop places. In short, each post history can contain multiple journeys, and each journey consists of a start and end point with a sequence of stops and moves in between. There is no convention on how to determine which stops are start and end stop places, as the definition of this depends on the application that it is used for. In the case of this research, each journey should theoretically begin and end at the same point: the home location. However,

due to the limitation of Instagram's API, it might well be possible that this home location is unknown based on the data retrieved from participants' social media content. As a solution, the research on this privacy related issue will be moved to the evaluation section, where participants will be asked how likely they regard the chances of their home location to be revealed based on their post histories. Furthermore, the start and end stop places of journeys will be identified by means of clustering algorithms.

3.2.1 Identification of stop places and moves

Stops are defined by Spaccapietra et al. (2008) as a location that is specifically marked as a stop by the tracked subject, and that has a certain time-span without a change of location. For data retrieved from social media however, it is quite difficult to measure whether or not a post was made and tagged during a stop or during a move. Messages and photos can for example be posted when one is staying at a certain place for a longer time, as well as when the participant is on their way to a next stop place.

On the subject of the identification of stops, literature is mostly concerned with tracks that have been measured at a regular interval that was set by the researcher (Moreno, Times, Renso, & Bogorny, 2010; Palma, Bogorny, Kuijpers, & Alvares, 2008). However, in the case of this research, location and timestamp pairs only become available at the convenience of the participant themselves, and are therefore significantly more sparse and irregular than general trajectories. Identification of stops in trajectories with regular and more frequent time intervals is usually done by clustering methods of the Stops and Moves of Trajectories (SMoT) family (Moreno et al., 2010; Rocha, Times, Oliveira, Alvares, & Bogorny, 2010; Palma et al., 2008; Alvares et al., 2007). Change in velocity (CB-SMoT) or direction (DB-SMoT) are also used as input for common methods in the identification of stops (Moreno et al., 2010; Rocha et al., 2010; Palma et al., 2008). The problem with these methods however, is that they only measure the stops based on objective measures, which require a denser and more consistent capture of time-location sets than social media usually provides. As a solution, each post will be categorised as a stop place, and the distance between the stops will be visualised as a move. Subsequently, in the evaluation the correctness of this method will be assessed.

In conclusion, there are multiple clustering algorithms that could be used for the clustering of trajectory data. However, due to the sparse and irregular nature of the input data for journey

reconstructions in this research, most of these are inapplicable. Therefore, a choice is made to implement the ST-DBSCAN algorithm to cluster journeys, so that each post gets assigned to an individual journey.

3.2.2 Identification of journeys

After identifying all stop places, the next step is to categorise which stop place posts were made during the same journey. In contrast to the research of Kisilevich, Krstajic, et al. (2010), who visually identify the journeys in the data they retrieved from social media geotags, journeys in participants' tracks in this research will be automatically identified by means of a clustering algorithm. Because of the heterogeneity of the input data, this method might be less accurate than manually and visually examining the data, but will enable the handling of a larger amount of data (Birant & Kut, 2007). The performance of the implemented cluster algorithm will be evaluated based on its outcomes, which will be assessed by the participants in the questionnaire. The clustering algorithm is used to group the points that belong to separate journeys. The first and the last place post of the resulting stop place clusters are then the start and end stop places.

Clustering algorithms

Many cluster algorithms for track data cluster data based on path similarity (Kisilevich, Mansmann, Nanni, & Rinzivillo, 2010; Palma et al., 2008), which is not applicable to this case, as each participant's journey will be reconstructed separately. Cluster methods in this research should cluster posts based on stop events. Out of the three most frequently used types of clustering methods, a density-based clustering method is preferred over a hierarchical or partitional method for the purpose of journey reconstruction (Madhulatha, 2013). The reason for this is that density based clustering algorithms are able to find clusters of varying amount, size and shape, and are able to filter outliers in the process (Ester, Kriegel, Sander, & Xu, 1996; Palma et al., 2008; Kisilevich, Krstajic, et al., 2010; Birant & Kut, 2007). One of the first density-based clustering algorithms is the Density-Based Spatial Clustering of Applications and Noise (DBSCAN). The main advantage of this specific clustering algorithm is that it is able to find clusters of different shapes (Palma et al., 2008; Birant & Kut, 2007). This is an important feature for the clustering of journey data, because the spatial distribution of journeys is not fixed.

The regular DBSCAN algorithm bases its clustering on two parameters. First of all, a threshold value is needed that depicts the radius around a core object. Furthermore, a minimum amount

of points that should fall within a cluster has to be set (Ester et al., 1996; Kisilevich, Mansmann, et al., 2010). This poses a problem to the reconstruction of journeys, as the temporal factor is essential in the clustering of this type of event-based movement data. Furthermore, the regular DBSCAN algorithm bases identification of noise on a general density among all clusters (Ester et al., 1996; Birant & Kut, 2007). This way of finding noise also poses a problem to the sparse and irregular trajectory data that social media provides. Because of the irregularity of the input data, densities may vary between the possible clusters. A solution for this is provided by Birant and Kut (2007), who designed the Spatial-Temporal DBSCAN (ST-DBSCAN). The ST-DBSCAN algorithm clusters the input data not only on the basis of spatial attributes, but also takes temporal aspects into account. Furthermore, each cluster gets labelled by a density factor, so that noise points can be detected even when clusters of different densities exist (Birant & Kut, 2007).

Clustering by means of the ST-DBSCAN algorithm is based on three parameters. Like the general DBSCAN algorithm, ST-DBSCAN takes a spatial threshold (Eps1) and a minimum amount of neighbourhoods (MinPts) parameter, but furthermore also takes a temporal threshold (Eps2) as a parameter (Birant & Kut, 2007). The MinPts parameter should be calculated as $\ln(n)$, where n is the number of entries in the database (Ester et al., 1996; Birant & Kut, 2007). Nevertheless, in case of the journey reconstruction, MinPts = 1 because of the sparse and irregular nature of the input data. The next parameter, Eps1, should be calculated based on the *k-nearest neighbourhoods* of the input points. In the calculations of the k-nearest neighbourhood, the parameter MinPts is equal to k (Birant & Kut, 2007). This means that for each point, the distance to the k-closest point is calculated. The maximum value of the resulting array of distances will be used as spatial threshold Eps1. Finally, the temporal threshold is calculated as the average amount of time between the posts of a participant. Once these inputs are set, the algorithm iterates over the entries of the input file, which is the track data per participant in the case of this research. By means of a function that retrieves the neighbourhoods of a point, it is assessed whether or not the point in question belongs to a cluster or is classified as *noise*.

The result of the implementation of this clustering algorithm is that each stop place post is assigned to a cluster, which represents a journey. A simplified visualisation of this process is shown in figure 3.1.

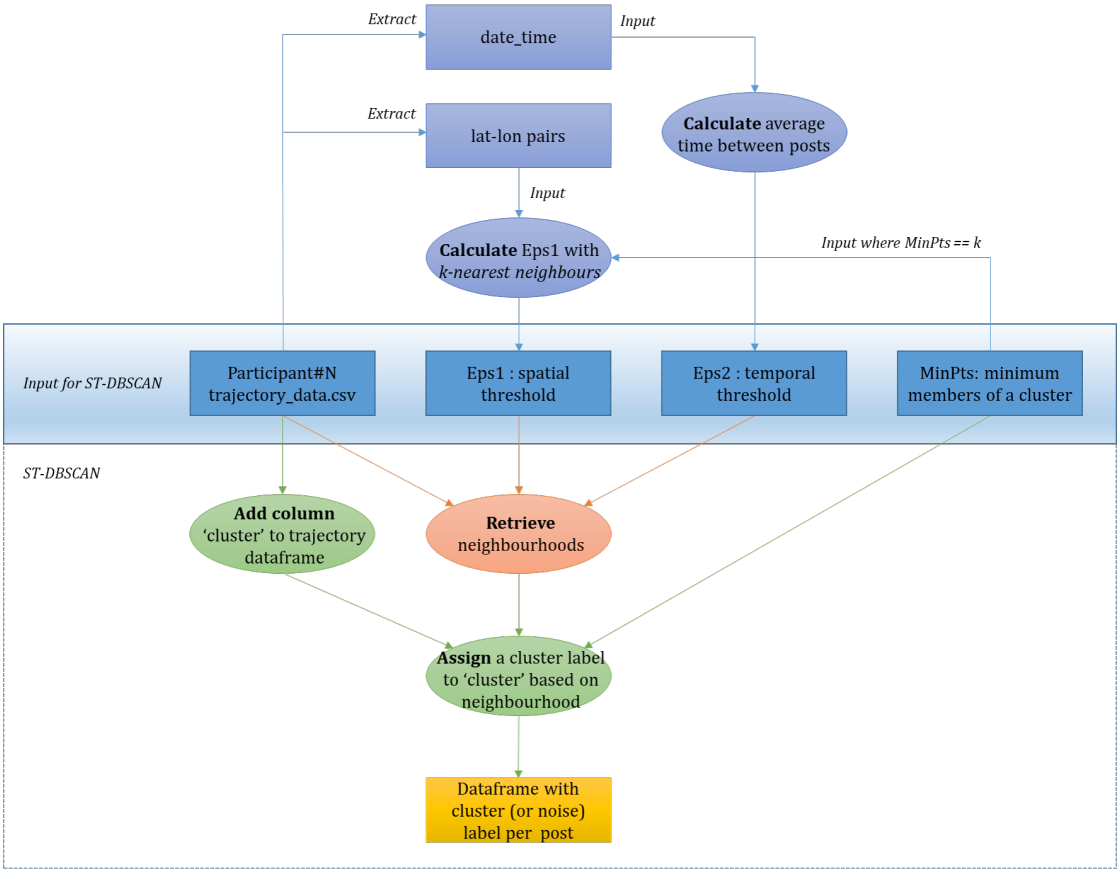


Figure 3.1: ST-DBSCAN process

3.2.3 Visualisation

For the visualisation of the journey reconstructions, the Folium python library is used. Folium is a python library that uses JavaScript's Leaflet to draw maps (Story, 2013). Because of the Leaflet interface, participants can zoom on their map, and click on the map markers that show the location name of their geotag. Stop places that together form a journey are represented by map markers in the same colour. Furthermore, stop places that are identified as noise points by the ST-DBSCAN algorithm are shown as black map markers. The coloured map markers and their meaning are displayed in a legend. The visualisation is handled in https://github.com/bkronemeijer/GIMA_thesis_journey_reconstruction/blob/master/visualisation.py.

3.3 Evaluation

The user experience will be measured by means of a structured questionnaire that is based on the defined quality criteria for journey reconstruction. The reason that this qualitative method is the preferred way of assessing the journey reconstruction, is that the assumed accuracy of the reconstruction is based on personal experiences. However, the purpose of the evaluation is not to investigate how respondents *feel* about the reconstruction, but rather to quantify accuracy of the reconstruction as assessed by participants. Therefore, questionnaires are the most appropriate way to get respondents' quality assessments. This way, respondents can both answer some quantifiable questions about the reconstruction, as well as elaborate on their answers in a personal manner.

An issue to address is the anonymisation of the respondents. For this reason, participants will not be numbered when mentioned in the results, but will merely be addressed by means of the word participant. For example: "a participant mentions that...", is a possible reference to a random participant. In the elaborate textual answers, all information that might possibly lead to the identification of a participant will either be generalised or deleted, so that the respondents remain anonymous. Furthermore, all data will be erased directly after the research is finished.

3.3.1 Questionnaire construction

In this research, the quality estimation is focused on the intrinsic quality of social media data. The questions of the questionnaire can be divided into six categories, that serve as the quality dimensions based on which the intrinsic quality is measured. The used quality properties are:

1. *Location privacy*

   This category of quality measures the extent to which a person's home location can be estimated based on their post history. Furthermore, location and absence privacy are assessed. These types of privacy regard the probability of discovering where a person is, or knowing where a person is not based on their social media behaviour. For example, when people do not post about an event at the same time the event is happening, it will be less easy to threaten their location and absence privacy, as their information is not entirely up to date.

2. *Spatial accuracy of reconstructed stop places*

   This category measures the spatial accuracy of posts. People tag places they have visited, but these places are not necessarily accurately geotagged, as people tag a place name instead of a space location.

3. *Spatial resolution of reconstructed stop places*

   This category assesses the spatial resolution of posts. The resolution in which people geotag places varies greatly, which could influence the quality of the reconstruction.

4. *Spatio-temporal accuracy of reconstructed stop places*

   This category examines the behaviour of a social media user concerning geotags. The footprint mismatch error e.g. measures whether people are generally more prone to tagging the place they took a picture of, or to tagging the place they were in at the moment of posting.

5. *Completeness and precision of reconstructed stop places*

   This category measures the completeness and precision of the set of stop place posts. Precision entails the amount of stop places that are correct, whereas recall measures the amount of events that did occur in reality, but that were not recorded in the dataset. Furthermore, it is possible that people do not geotag their location, but instead describe or mention the location in e.g. the caption.

6. *Quality of reconstructed journeys*

   This category measures the precision and recall on a journey level. The ST-DBSCAN algorithm clusters entries based on their spatial and temporal properties, but might falsely aggregate separate journeys into one, or might on the other hand falsely separate one journey into multiple smaller journeys. Furthermore, because of throwbacks, the sequence of the journeys might be disordered.

The second, third, fourth and fifth property cover the individual posts that are marked as a stop, whereas the sixth covers the evaluation of the clustering of the posts as journeys. The second, third, fourth and fifth property are considered to be most important for the measurement of the quality, because they measure the data that comes directly from the platform itself. On the other hand, the sixth property measures a version of the data that is edited by a clustering algorithm, which makes the outcome of the sixth property dependent on not only the data itself,

but also on the performance of the clustering algorithm. This section continues with describing how the quality of each of the quality dimensions is measured.

Location privacy

Table 3.1: Overview of testing methods for privacy

Topic	**Question**	**Operationalisation**
Extent of awareness of data production	To what extent are you aware that you were producing the information shown in the reconstruction?	5 point Likert scale
Confidence of finding home location in reconstruction	To what extent are you confident that your home location can be seen in the reconstruction?	5 point Likert scale
Confidence of finding home location in entire post history	To what extent are you confident that your home location can be revealed based on your full post history?	5 point Likert scale
Location and absence privacy	How long do you generally wait between visiting a place and posting about it?	Ordinal time-scale

Spatial accuracy of reconstructed stop places

Table 3.2: Overview of testing methods for spatial accuracy

Topic	**Question**	**Operationalisation**
Amount of inaccurate geotags	In which cases is a post about a different place than the one you geotagged?	Descriptive statistics of questionnaire results
Spatial accuracy of each post	For each post, how far away are they from the true place you visited?	Descriptive statistics of questionnaire results

Spatial resolution of reconstructed stop places

Table 3.3: Overview of testing methods for spatial resolution

Topic	**Question**	**Operationalisation**
Influence of spatial resolution on satisfaction reconstruction	To what extent do you feel the reconstruction is less accurate because the geotag was set in a broader spatial resolution?	5 point Likert scale

Spatio-temporal accuracy of reconstructed stop places

Table 3.4: Overview of testing methods for spatio-temporal accuracy

Topic	**Question**	**Operationalisation**
Footprint mismatch error	Do you usually tag the location in which you are in at the moment? Or do you tag the place that you took a photo of?	Percentage of total
Time between event and post	How long do you generally wait between visiting a place and placing a post about it?	Descriptive statistics of questionnaire results
Throwbacks	How many of the shown stops can you identify as so-called throwbacks?	Descriptive statistics of questionnaire results

Completeness and precision of reconstructed stop places

Table 3.5: Overview of testing methods for completeness

Topic	**Question**	**Operationalisation**
Precision	Are the places shown on the reconstruction all of the places you have stopped?	Percentage of total
	In which cases is a post about a different place than the one you geotagged?	Descriptive statistics of questionnaire results
Recall	How many are lacking?	Textual analysis of open question results
Geotag behaviour; why geotag?	Do you in general always geotag your posts? and What are reasons for you to geotag your post?	Textual analysis of open question results
Implicit identification of location	Do you describe the place you visited in the caption?	Textual analysis of open question results

Quality of reconstructed journeys

Table 3.6: Overview of testing methods for reconstructed journeys

Topic	**Question**	**Operationalisation**
Precision of journeys	How many of these journeys correspond to journeys you actually made?	Percentage of total
Recall of journeys	How many are lacking?	Percentage of total
Identification of cluster errors	For the journeys that are incorrect, can you identify the error?	Textual analysis of open question results
Temporal resolution; journey sequence	How many of the stops are shown in the correct sequence?	Percentage of total & textual analysis of open question results

3.4 Analysis scheme

The analysis scheme summarises how the research is carried out. The first phase consists of a literature study on the existing state of the art in the field of VGI and mobility. The information that is gathered in the literature study will serve as input for the user study, in which participants will be gathered and where their information will be extracted from the web by the use of Instagram's API. For the security of the personal data of the participants, it is important that privacy measures are taken. The data that is extracted from the participants' social media accounts serves as input for the tracks that are the first step in phase 3. An elaboration on the analysis methods can be found in figure 3.3.

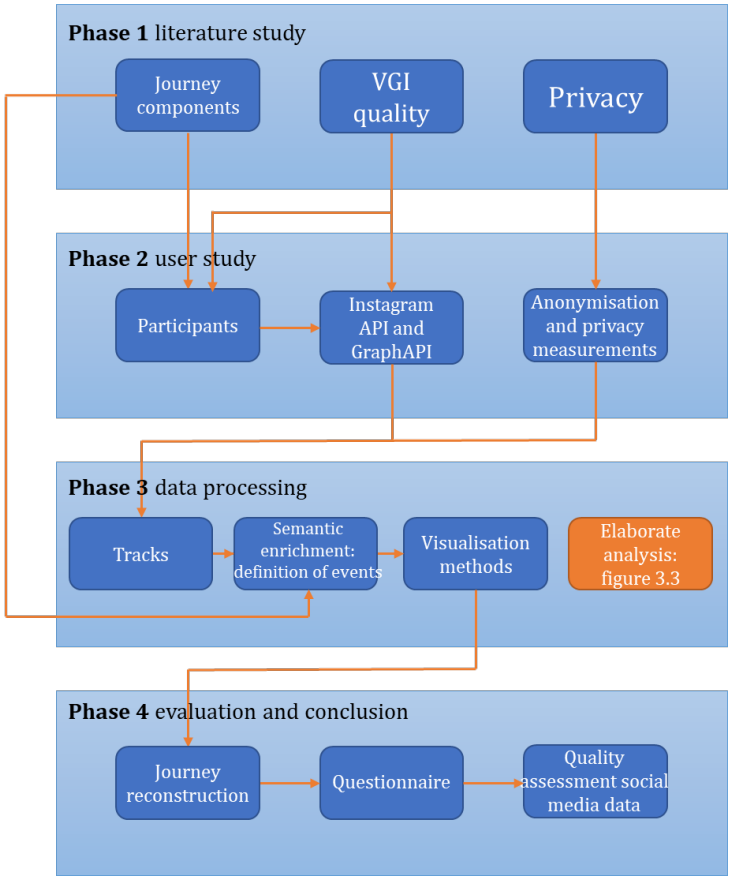


Figure 3.2: Analysis scheme

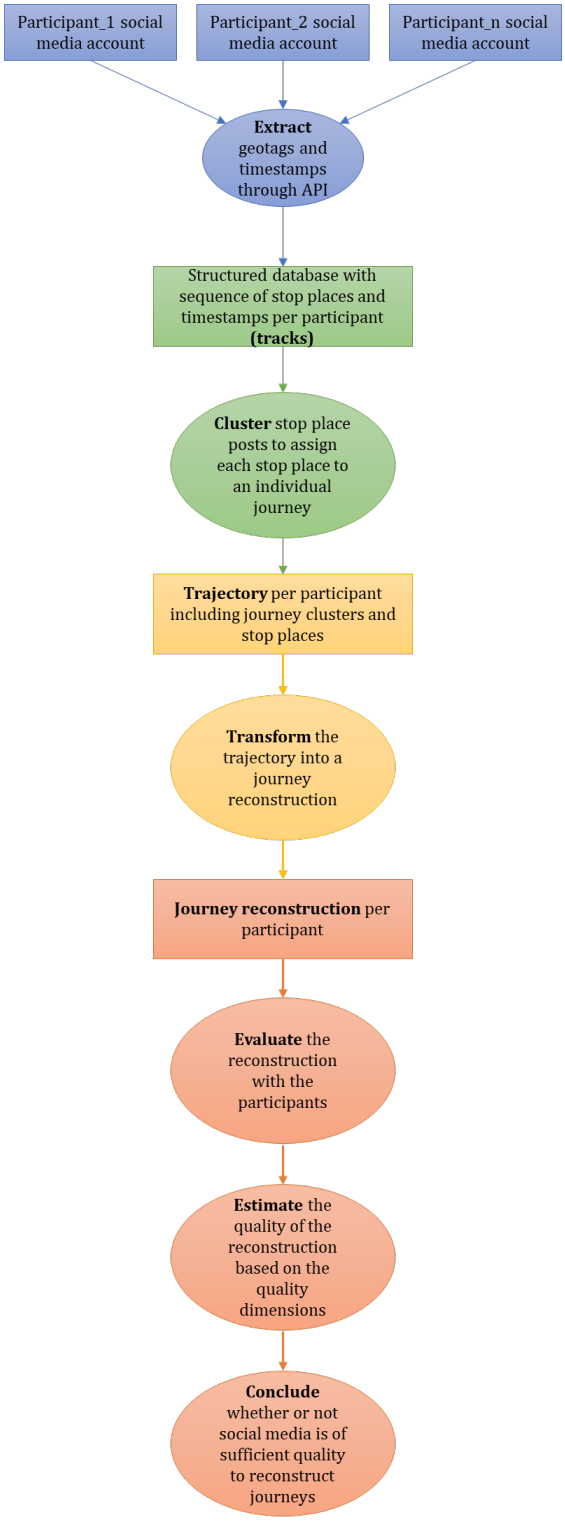


Figure 3.3: Elaborate analysis description

Chapter 4

Results and discussion

This chapter describes the results of the methodology as substantiated in chapter 3. First, results of the journey reconstruction are shown. Subsequently, the results of the evaluation are presented and discussed per quality dimension. Finally, the implications of these results are discussed.

4.1 Journey reconstruction

This section shows some examples of journey reconstructions. The journey reconstruction results in an HTML document, for which reason the examples are shown by means of screenshots of the corresponding HTML page.

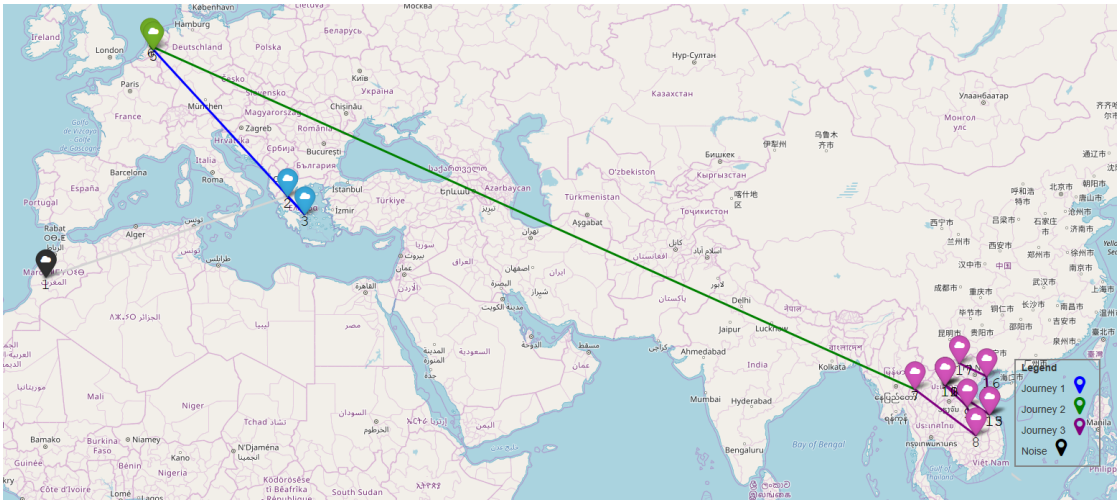


Figure 4.1: Example of a journey reconstruction

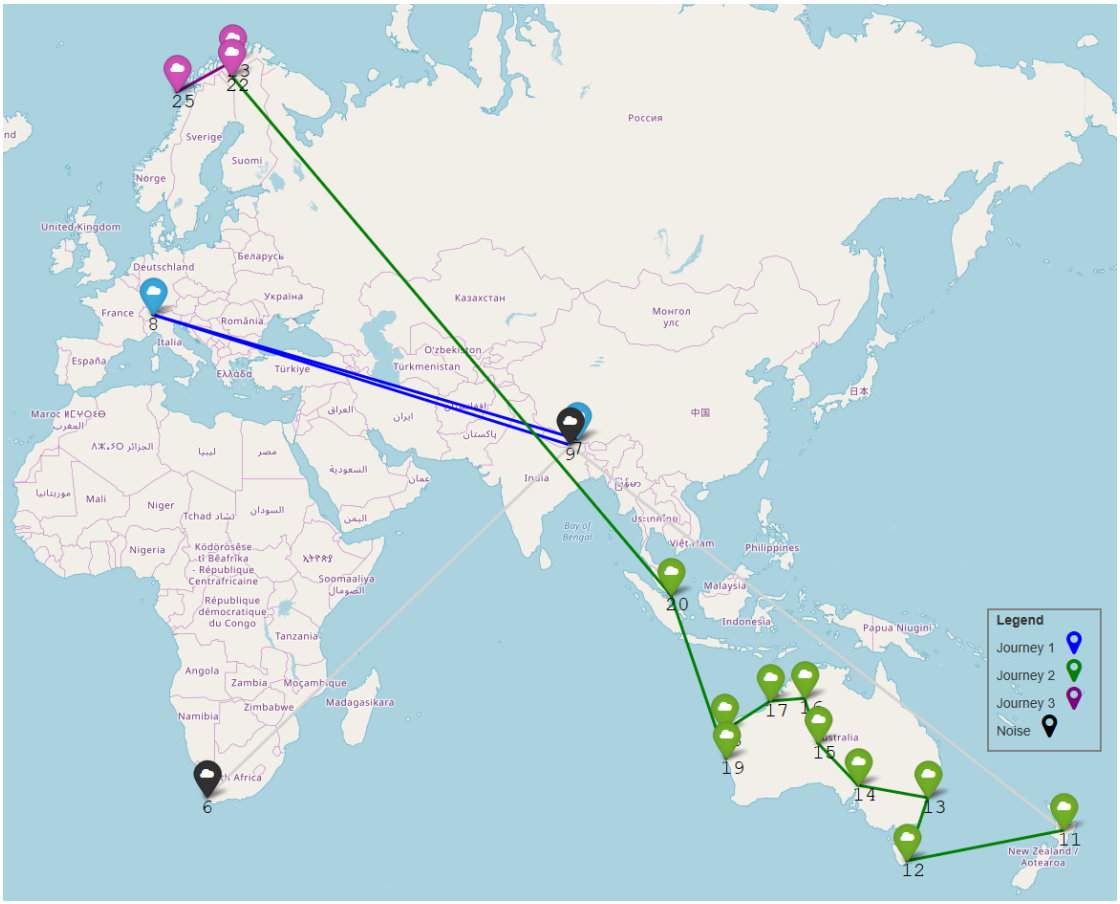


Figure 4.2: Example of a journey reconstruction

Figure 4.1 shows a clear example of a successful journey reconstruction. This participant clearly made one journey in southeast Asia, one in the Netherlands, one in Greece and one in Morocco. Figure 4.2 shows another example of a journey reconstruction that was mostly correct. The reconstruction shows the journey through Australia and New-Zealand, and a journey in the north of Norway. However, the blue points and the noise points (in black) seem to make less sense. The underlying data for the reconstruction in figure 4.2 can be viewed in table 4.1. In the data, stop place posts that are categorised as ST-DBSCAN noise are marked as *-1*, and it means that the participant only posted once during a journey. This makes sense for the entries in Cape Town (OID 4) and Kathmandu (OID 7). However, the ST-DBSCAN algorithm linked San Gottardo in Switzerland (OID 6) to Tingri in China (OID 5), even though these two stop place posts do not seem to belong to the same journey at a first glance.

Table 4.1: Input table for the journey reconstruction in figure 4.2

OID	date_time	latitude	longitude	location_name	cluster
4	13/03/2019 08:10	-33.9253	18.4239	Cape Town, Western Cape	-1
5	04/10/2018 10:55	28.5667	86.6333	Tingri, Xizang, China	1
6	20/09/2018 10:01	46.55608	8.565637	San Gottardo, Uri, Switzerland	1
7	27/08/2018 06:39	27.71378	85.31024	Kathmandu, Nepal	-1
9	04/05/2018 06:49	-38.6868	176.0696	Taupo, New Zealand	2
10	22/04/2018 12:00	-42.8802	147.3284	Hobart, Tasmania	2
11	17/04/2018 23:24	-33.8611	151.2065	Shangri-La Hotel, Sydney	2
12	16/04/2018 11:30	-30.5311	139.3038	Vulkathuna-Gammon Ranges	2
13	13/04/2018 13:12	-25.2432	130.9842	Uluru-Kata Tjuta National Park	2
14	10/04/2018 13:40	-17.4489	128.5456	Purnululu National Park	2
15	08/04/2018 15:54	-17.9667	122.233	Broome, Western Australia	2
16	06/04/2018 12:59	-23.1416	113.7733	Ningaloo Reef, Coral Bay	2
17	04/04/2018 12:08	-27.7065	114.1679	Kalbarri, Western Australia	2
18	28/03/2018 18:19	1.28259	103.8644	Gardens by the Bay Singapore	2
20	20/03/2018 09:14	69.01707	23.04527	Kautokeino	3
21	19/03/2018 15:41	69.96841	23.27076	Alta, Norway	3
23	16/03/2018 10:51	67.93337	13.08964	Reine	3

Despite the successful examples shown in figure 4.1 and 4.2, some journey reconstructions turned out to be chaotic, as for example shown in figure 4.3. The corresponding table 4.2 shows that over half of the input stop places are categorised as noise points. Another remarkable occurrence in these input tables, is the difference in decimal precision between geotags. For example, in table 4.2, the post place with OID 7 has one decimal for latitude, while other post places have up to 6 decimals. This case is further shown in another participant's input table, in table 4.3. This table shows that the location named *Sweden* (OID 15) and *Norway* (OID 2) have coordinates with zero decimals, while the locations with inputs *Jachthaven Vlieland* and *Friesland, Netherlands* have a precision of eight decimals. This makes sense in the case of *Jachthaven Vlieland*, as this is a small harbour on one of the Dutch islands. However, *Friesland, Netherlands* is a province, and therefore covers a bigger area than the harbour on the island Vlieland. This indicates that there is no clear correlation between the amount of decimals in the geotag and the spatial resolution of the place of the geotag.

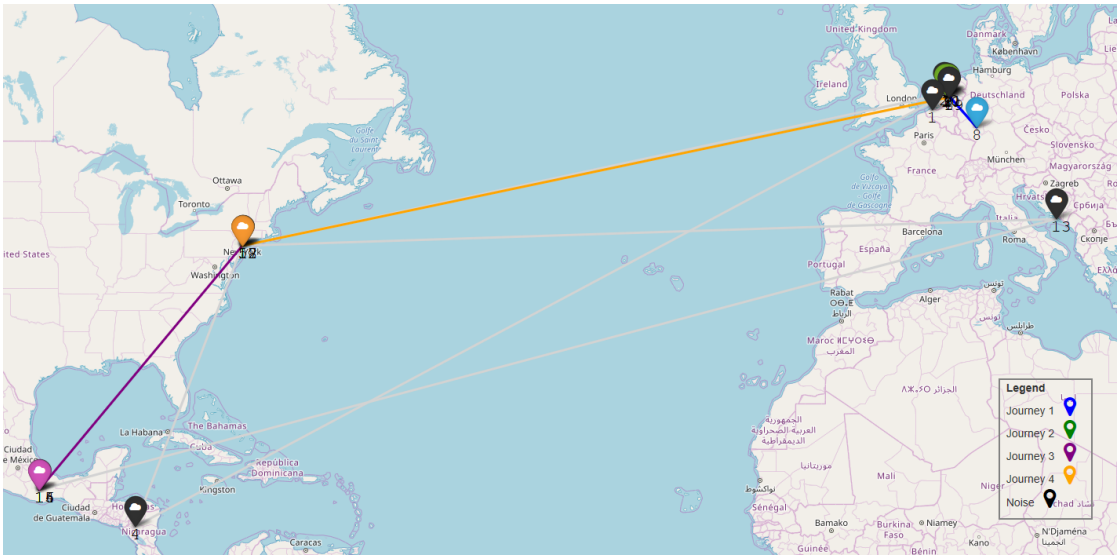


Figure 4.3: Example of a journey reconstruction

Table 4.2: Input table for the journey reconstruction in figure 4.3

OID	date_time	latitude	longitude	location_name	cluster
1	05/03/2019 19:09	51.2085	3.2249	Brugge, Belgium	-1
2	20/02/2019 12:17	52.31056	4.973333	Amsterdam-Zuidoost	-1
3	07/02/2019 09:09	52.33967	4.874213	Metro 50 Amsterdam	-1
4	26/01/2019 17:52	13.09138	-86.001	Nicaragua	-1
5	09/01/2019 20:57	40.7142	-74.0064	New York, New York	-1
6	01/01/2019 01:02	52.0874	5.1068	Utrecht	-1
7	13/12/2018 15:47	52.36667	4.9	A'dam	1
8	08/12/2018 20:17	50.003	8.2602	Mainz, Germany	1
9	17/11/2018 18:19	52.31056	4.973333	Amsterdam-Zuidoost	2
10	10/11/2018 18:59	52.30931	4.761543	Capitale des Pays-Bas	2
11	28/10/2018 21:32	52.3777	4.9001	Amsterdam, Netherlands	-1
12	24/10/2018 14:28	40.7142	-74.0064	New York, New York	-1
13	19/10/2018 09:08	42.9622	17.1369	Korcula	-1
14	11/10/2018 05:02	17.0603	-96.7255	Oaxaca, Mexico	-1
15	24/09/2018 13:57	17.0693	-96.726	Monte Albán, Mexico	3
16	21/09/2018 20:46	17.0606	-96.7254	Oaxaca City	3
17	19/09/2018 12:09	40.7142	-74.0064	New York, New York	4
18	16/09/2018 21:05	40.71327	-73.971	Williamsburg Bridge	4
19	09/09/2018 20:41	52.0874	5.1068	Utrecht	-1
20	06/08/2018 08:36	52.4054	4.545002	Woodstock Bloemendaal	-1

Table 4.3: Input table for a journey reconstruction

OID	date_time	latitude	longitude	location_name	cluster
1	06/03/2019 13:45	53.29574968	5.076145408	Veerboot Vlieland - Harlingen	-1
2	24/02/2019 19:29	61	8	Norway	-1
3	26/01/2019 17:27	53.36033226	5.214452523	Brandaris	-1
4	13/09/2018 16:53	53.36029	5.21409	West-Terschelling	1
6	19/08/2018 14:58	53.29670908	5.089419014	Jachthaven Vlieland	1
7	29/07/2018 18:29	53.29670908	5.089419014	Jachthaven Vlieland	1
8	23/07/2018 20:40	53.29627415	5.075306554	Vlieland Island	1
10	29/04/2018 16:53	53.36029	5.21409	West-Terschelling	2
12	31/03/2018 21:03	53.36029	5.21409	West-Terschelling	2
13	16/03/2018 22:34	53.36029	5.21409	West-Terschelling	2
14	28/12/2017 18:51	53.2	5.78333	Leeuwarden, friesland	-1
15	24/10/2017 15:20	61	15	Sweden	-1
16	09/09/2017 17:01	53.29670908	5.089419014	Jachthaven Vlieland	-1
17	14/06/2017 17:33	53.08805747	5.820695934	Friesland, Netherlands	3
18	22/05/2017 21:06	53.0919996	6.0860253	KV Drachten	3
19	21/02/2017 19:54	46.19570271	7.343201717	Veysonnaz, 4 Vallées	-1

4.2 Evaluation

In this section, the results of the evaluation are presented and discussed per quality dimension. Each quality dimension is first discussed separately, beginning with a table that summarises the questionnaire results. The table displays the explicit and, if available, the implicit results. The explicit results are results based on questions in the questionnaire that correspond to the questions as defined in the methodology section 3.4. The implicit results are based on answers to other questions than defined in methodology section 3.4, but that can be used to derive answers indirectly. For example, participants are asked to count the number of inaccurately placed geotags. In order to indirectly check for the validity of their answers, these answers are compared to the answers of another question, in which participants have to say how far away the geotag is from the true place they visited per post. Subsequently, the results are interpreted. Quality depends on the application, which is a journey reconstruction in this case. If another application is maintained, the quality might be estimated differently. Furthermore,

all interpretations are only applicable to the participant sample, as it cannot be generalised towards the population of all social media users. In order to do that, further research with a larger sample and a wider variety is needed. This has not yet been done in this research, as this research entails an exploratory investigation in the ability to reconstruct and assess personal posts, and is therefore mainly focused on the method rather than on the extent of the sample.

4.2.1 Location privacy

Table 4.4: Summarised results for location privacy

Question	**Explicit result**	**Implicit result**
To what extent are you aware that you were producing the information shown in the reconstruction?	$Mean = 4.40$ $Median = 5$ $Mode = 5$	N.a.
To what extent are you confident that your home location can be seen in the reconstruction?	$Mean = 2.33$ $Median = 3$ $Mode = 1$	N.a.
To what extent are you confident that your home location can be revealed based on your full post history?	$Mean = 3.13$ $Median = 4$ $Mode = 4$	N.a.
How long do you generally wait between visiting a place and posting about it?	Within approx. 1 day: 40% More than a day: 60%	N.a.

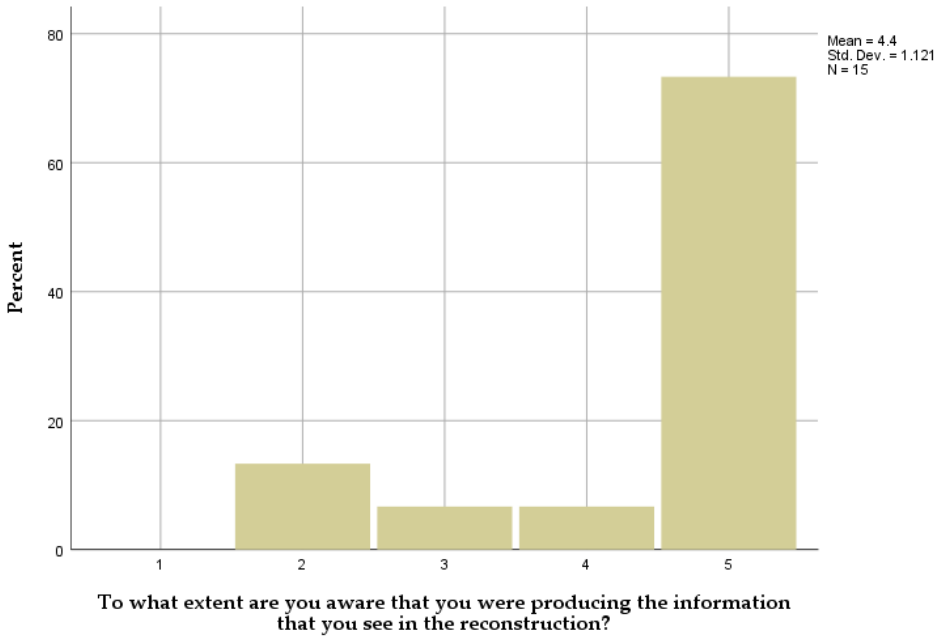
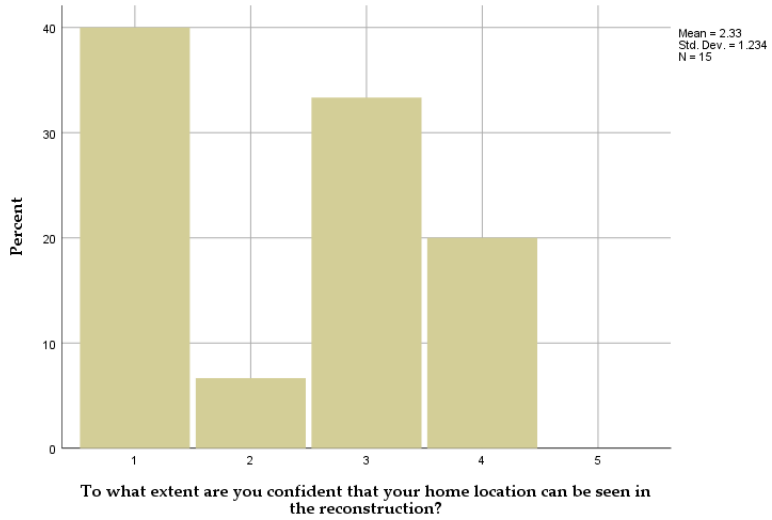
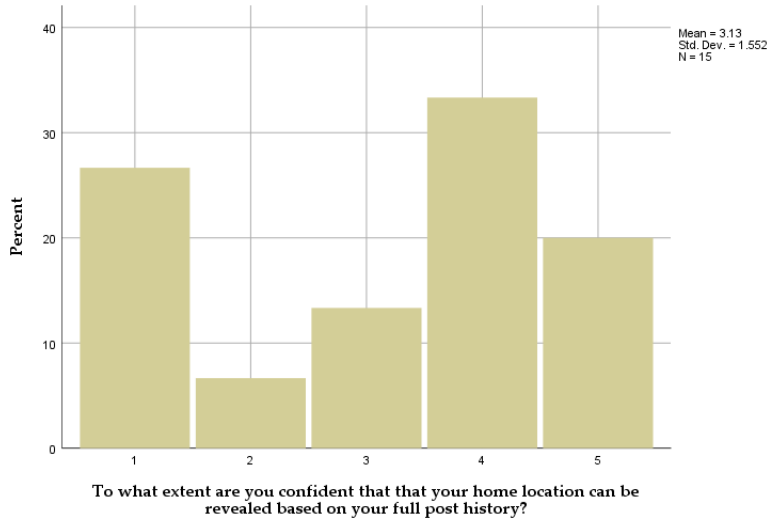


Figure 4.4: Histogram of extent of awareness of data production

As table 4.4 and figure 4.4 reveal, participants are generally aware that they produce data based on which their journey can be reconstructed. This has multiple implications for location privacy. First of all, it is positive that people are aware that they produce this kind of location data, because awareness is the first step towards change, if change is deemed necessary. Furthermore, the fact that the participants are generally aware they produce this information could implicate that they feel that their location privacy is not significantly threatened by placing geotags in their posts.



(a) In reconstruction



(b) In full post history

Figure 4.5: Histograms of possibility to reveal home location

Figure 4.5 reveals that participants are in general quite convinced that their home location can be found in their post histories, which implies that participants do not only geotag their travel events, but also geotag events that happen within their own surroundings. However, one needs to take into account that participants generally answered that they mostly tag places when they feel like that place is special and therefore deserves to be shown.

A comparison of these figures shows that the possibility of finding participants' home location has diminished due to Instagram's policy of limiting the API response for draft clients to only the last 20 posts. This means that Instagram's policy has a positive effect on their participants' privacy. Nevertheless, a large part of the home locations of participants can still be revealed even based on only the last 20 posts.

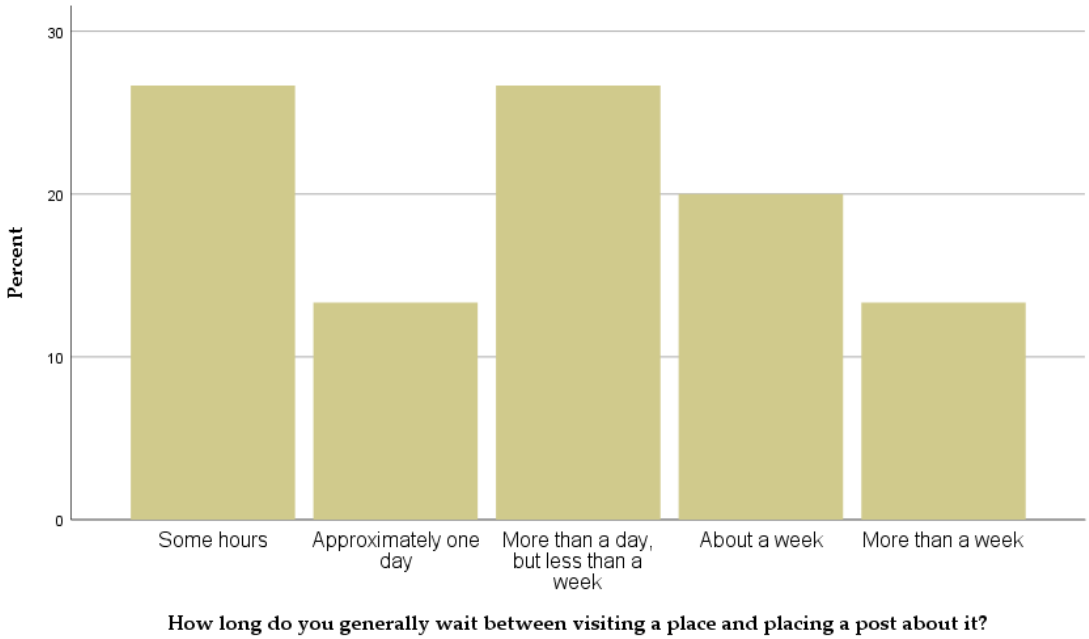


Figure 4.6: Location and absence privacy

40% of the participants posts approximately within one day of taking a picture. The rest of the participants takes longer than a day to post and geotag their whereabouts. This means that the majority of the participants cannot be correctly located at the appropriate time, as they might have moved between taking the picture and posting the picture with a geotag. The longer participants wait between taking a picture and posting it, the lesser the threat on their location privacy. Furthermore, these results imply that the quality of geotags is not high from the temporal perspective, because the longer participants wait with posting their pictures, the less accurate the timestamp becomes.

Summary

Most of the participants were aware that their personal posts could serve as input data for a journey reconstruction. This is in contradiction with the statement of Stefanidis et al. (2013),

who say that users are not aware of the implications of posting their personal location. Furthermore, participants are generally convinced that their home location can be revealed in a journey reconstruction. This could mean that they do not estimate the danger of geotags as high, because they post geotags even though they are aware of the implications. This is in accordance with the study of Danezis et al. (2005) and Krumm (2007), who discovered that individual's do not have a high regard of the protection of their privacy sensitive location data. This tendency may be explained by the temporal discrepancy between actual events and the posts participants place about it, as 60% of the participants generally waits longer than a day before they post about an event. Due to this discrepancy, participants cannot directly be placed in a location at the correct time, which conceals their true whereabouts. For that reason, individuals might feel that their identity is relatively safe, which could cause them to not have a high regard of the protection of their privacy sensitive location data.

4.2.2 Spatial accuracy and spatial resolution of reconstructed stop places

Table 4.5: Summarised results for spatial accuracy

Question	**Explicit result**	**Implicit result**
In which cases is a post about a different place than the one you geotagged?	Inaccurate: 23.5% Accurate: 76.5%	Inaccurate: 51.9% Accurate: 48.1%
For each post, how far away are they from the true place you visited?	*Mean of all posts* Exactly correct : 48.1% $<$ 1 km : 16.2% 1 - 2 km : 14.4% 2 - 5 km : 13.9% 5 - 10 km : 2.8% $>$ 10 km : 4.6%	*Mean per participant* Exactly correct : 50.3% $<$ 1 km : 15.4% 1 - 2 km : 13.6% 2 - 5 km : 13.5% 5 - 10 km : 2.6% $>$ 10 km : 4.7%

The results of the question "In which cases is a post about a different place than the actual place you visited?" reveal that 23.5% of the place posts are tagged in the wrong place. However, results of the question about the spatial accuracy of each post returned an incorrectness of 51.8%. This percentage was generated as follows. In the questionnaire, participants could indicate per post

number how far away the geotag is from the real place the event happened. For this they had six options, by which participants can indicate that the post corresponding to the post number is:

- Exactly correct
- $<$ 1 km off
- 1 - 2 km off
- 2 - 5 km off
- 5 - 10 km off
- $>$ 10 km off

In order to extract the amount of incorrectly placed geotags, every post that was indicated to not be *exactly correct* was accumulated per participant. This procedure resulted in 51.8% incorrectly placed geotags among participants.

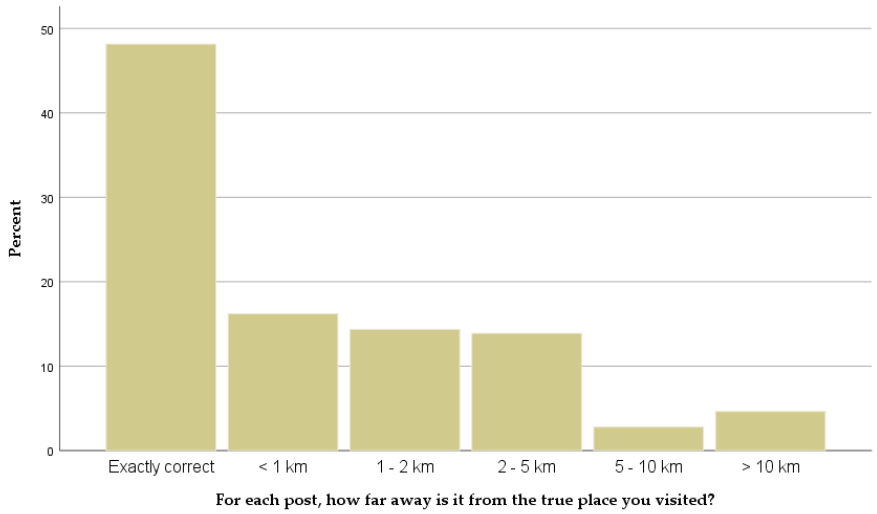


Figure 4.7: Distribution of average spatial accuracy of all posts in percent

Figure 4.7 shows the average spatial accuracy of all posts in this research. This means that all participants together on average answered *exactly correct* for 48.1% of their posts. Another measure for spatial accuracy shows the average of spatial accuracy per participant, and this reveals that participants on average answered *exactly correct* for 50.3% for their own posts. Both measures show the same tendency that about half of the posts are exactly correct, and that the less accurate the spatial accuracy, the less participants choose this spatial inaccuracy for their

post. This means that participants generally correctly geotag their posts. Furthermore, in cases where the geotag is not correct, the geotag is usually within 5 kilometres of the original place.

Summary

In general, about half of the place posts are accurately geotagged. The posts that are inaccurately geotagged are generally close to the original location of the event behind the geotagged post. For the purpose of journey reconstruction, this level of spatial accuracy is sufficient. Furthermore, participants tend to grossly overestimate the accuracy of their geotags. Results show that participants estimated 23.5% to be inaccurate, whilst the actual percentage of inaccurate geotags is 51.9%. This discrepancy could originate from the possibility that participants do not consider some inaccurate geotags as inaccuracies, but rather as a correct placial representation of the event they post about. In other words, it is likely that place matters more to the participants than the spatial accuracy of the coordinates in their geotags. This might be the reason that they consider less geotags to be inaccurate, than the actual amounts of geotags that are not exactly correct.

4.2.3 Spatial resolution of reconstructed stop places

Table 4.6: Summarised results for spatial resolution

Question	**Explicit result**	**Implicit result**
To what extent do you feel the reconstruction is less accurate because the geotag was set in a broader spatial resolution?	$Mean = 2.80$ $Median = 3$ $Mode = 3$	N.a.

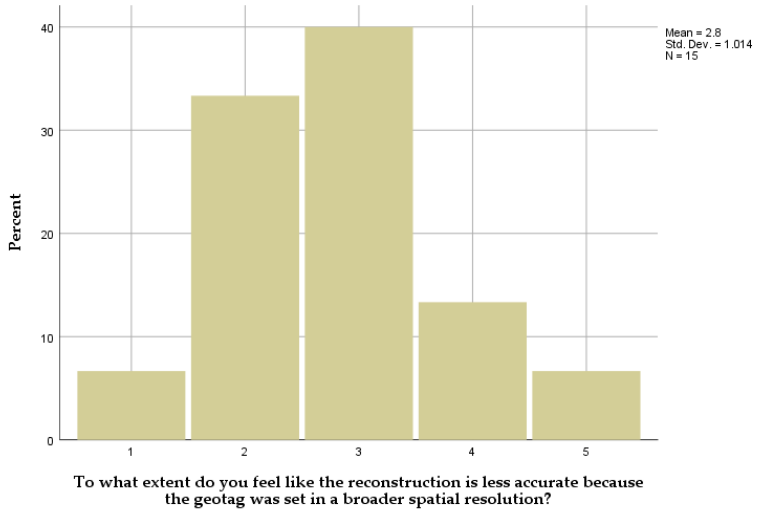


Figure 4.8: Histogram of satisfaction of reconstruction

The median of the results shown in figure 4.8 is 3, which implies that the spatial resolution of geotags has some effect on the satisfaction of the reconstruction, but is not generally disturbing. However, the results of this question might not be the best possible representation about how respondents experience the spatial resolution of geotags in the reconstruction. For example, the respondent who entered *5* as an answer to this question, also entered that the geotag of each post was *exactly correct*. This means that the participant indicated that every post was geotagged in the exact location where the participant had been, for which reason the reconstruction should also be exactly correct. If this were the case, the participant logically cannot answer *5*, which means that the participant strongly feels like the reconstruction is less accurate due to incorrectly geotagged places. Furthermore, out of two respondents who answered that they geotag the place they are in at the moment instead of tagging the place shown in their post, one answered *2*, and the other answered *4*. Therefore, it is difficult to draw a conclusion based on this question, as apparently the satisfaction of the reconstruction does not seem to depend only on the distance between the geotag and the actual event.

4.2.4 Spatio-temporal accuracy of reconstructed stop places

Table 4.7: Summarised results for spatio-temporal accuracy

Question	**Explicit result**	**Implicit result**
Do you usually tag the location in which you are in at the moment? Or do you tag the place that you took a photo of?	Geotag corresponds to event location: 86.7% Geotag corresponds to current location: 13.3%	Geotag corresponds to event location: 93.3% Geotag corresponds to current location: 6.7%
How long do you generally wait between visiting a place and placing a post about it?	Some hours: 26.7% A day: 13.3% Less than a week: 26.7% A week: 20.0% Over a week: 13.3%	N.a.
How many of the shown stops can you identify as so-called throwbacks?	$Mean = 31.3\%$ $Median = 10.0\%$ $Std.Dev. = 38.5\%$	$Mean = 58.5\%$ $Median = 78.6\%$ $Std.Dev. = 38.5\%$

86.7% of the participants indicate that when they geotag their posts, they tag the location in which the event took place. The other 13.3% indicates that they tag the location in which they are at the moment of posting about an event. The extent to which this difference results in a footprint mismatch error depends on how long the participant generally waits between taking a picture of an event, and placing this picture on Instagram. The longer the wait, the larger the footprint mismatch error, because the longer the period of time between the event and the post, the further the participant could have travelled in the meantime.

Only one of the participants who indicated that they tag the location in which they currently are when posting, also indicated to almost immediately post about events, for which reason the geotag likely is similar to the event location. Therefore, in the implicit results the percentage of participants whose geotags correspond to the location of the events in the posts is 93.3%, whereas the percentage of participants whose geotags do not correspond to the same location in which the posted event took place is 6.7%. This means that the spatial attribute in general corresponds to the location of the event in the post. In other words, the correctness of the spatial attribute is

good. On the other hand, as also revealed in figure 4.6, the correctness of the temporal attribute is flawed, because only 26.7% of the participants posts about an event within some hours. This means that 73.3% of the participants have incorrect temporal attributes, as the timestamp of the post does not match the time of the event that is posted about.

The correctness of temporal attributes is not only endangered because the temporal lag between an event and the post, but also because participants might post so-called throwbacks. This is a specific type of temporal lag, where a participant posts about an event after other events have happened in between. Therefore, not only the temporal correctness is off, but also the sequence of visited places. For example, a participant tagged a location in Bolivia, even though the participant had been home for several months and had also posted about other events in the meantime. This might lead to a reconstruction as shown in figure 4.9. Most of the last 20 posts of this participants are centred around the participant's alleged home location, indicating that the participant is not currently travelling. However, the post with post number 14 is situated in Bolivia. The clustering algorithm identified this stop place as a noise point, for which reason it is displayed as a black marker.

Throwbacks are of significant influence to the correctness of journeys. First of all, because of throwbacks, the general sequence of stop places is inherently incorrect when a participant uses throwbacks. Furthermore, the ST-DBSCAN algorithm cannot function properly due to such throwbacks, because the less true the data, the less true the outcome of the algorithm. For this reason, the larger the amount of throwbacks, the smaller the chance that the journeys that are clustered by ST-DBSCAN correspond to journeys that are actually made by the participant. When asking participants in the questionnaire how many of their posts are throwbacks, they answered 31.3% on average. However, when asking how many stop places in the reconstruction are shown in the correct sequence, the percentage is higher and it shows that 58.5% of the posts is shown in an incorrect sequence. This implies that the amount of throwbacks should be higher than 31.3%, as throwbacks are a reason that stops are reconstructed in the wrong sequence.

Summary

The spatial attribute of Instagram data from personal posts is quite correct, in the sense that participants generally try to geotag the place of the event in the post, instead of tagging the location in which they are at the moment of posting. The interpretation that this is indeed correct

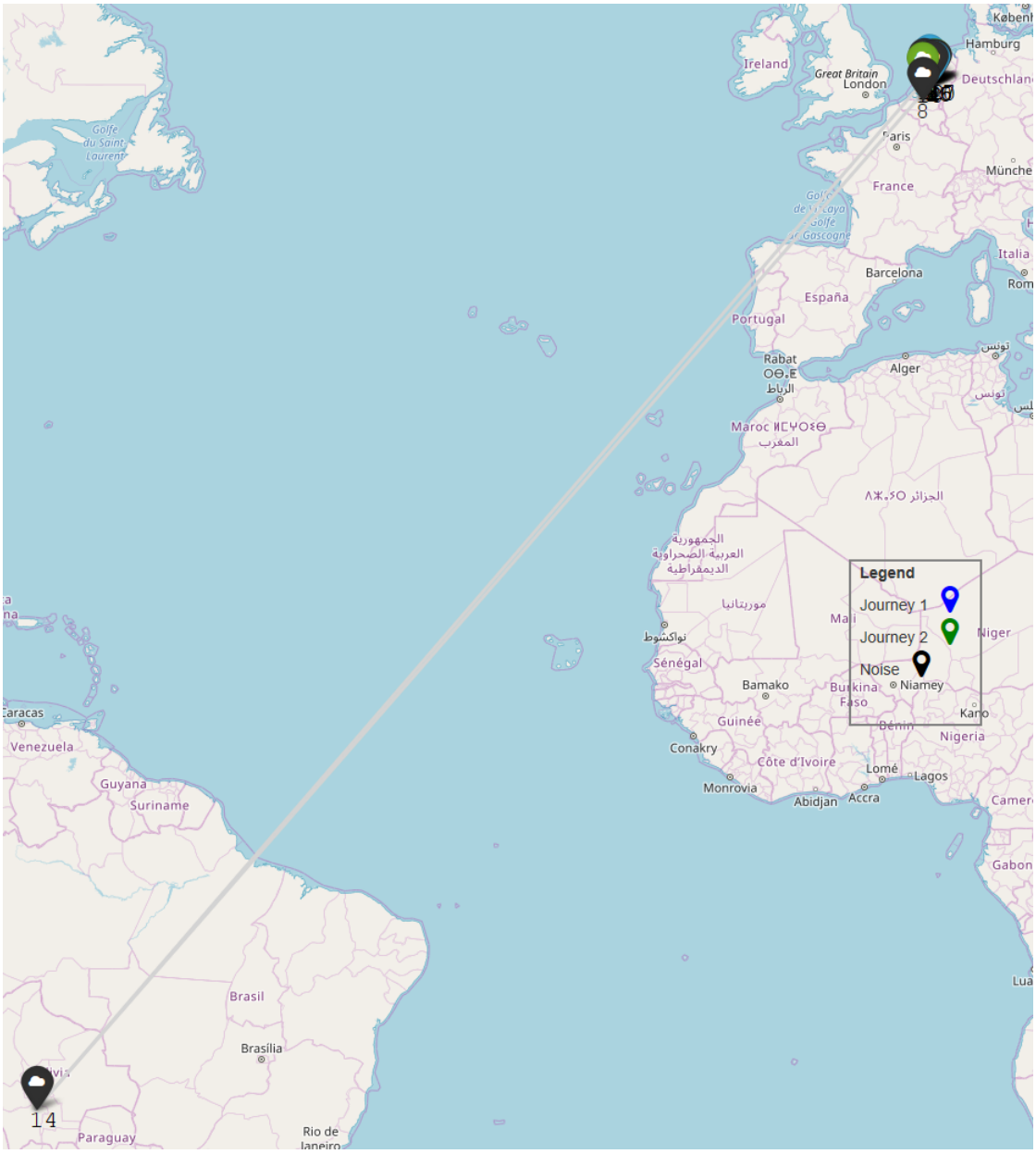


Figure 4.9: Example of a throwback in a journey reconstruction

is debatable, because it depends on the application. In this case, it is deemed correct because the picture displayed in a post corresponds to the geotag. For other applications, it might be more beneficial if participants tag the location in which they are in at the moment of posting, because then the temporal attribute would exactly match the participant's location at that moment. The correctness of this spatial attribute is a different discussion than the discussion about spatial accuracy as discussed in subsection 4.2.2.

Furthermore, the results reveal that the correctness of the temporal attribute is untrustworthy, because most participants wait before they post about an event. Moreover, the use of throwbacks causes even more errors in the correctness of spatial and temporal attributes.

4.2.5 Completeness and precision of reconstructed stop places

Table 4.8: Summarised results for completeness and precision

Question	**Explicit result**	**Implicit result**
Are the places shown on the reconstruction all of the places you have stopped?	84.7% correspondence	82.2% correspondence
How many are lacking?	"A lot"	N.a.
Do you in general always geotag your posts? and What are reasons for you to geotag your post?	To show special locations	N.a.
Do you describe the place you visited in the caption?	Sometimes available in caption or hashtags	N.a.

84.7% of the reconstructed stop places corresponds to places participants had visited in reality. In order to check this percentage, an implicit result was generated from the question "In which cases is a post about a different place than the one you geotagged?". Participants are here supposed to fill in the post numbers that do not correspond to places they have visited in reality, for which reason the implicit result should match the explicit result. As shown in table 4.7, the differences between the explicit and implicit result is minimal. It is interpreted that the precision of data

from personal posts is high, because the percentage of reconstructed stop places that correspond with actual events is high.

There are no explicit numerical results available for recall. Nevertheless, participants indicated that a lot of stop places are missing. The reason for this according to one of the participants is "I generally post 1 photo of a trip but in reality I have been to a lot of places during that trip.", and another reason is "I don't post about every place that I go to. I post pictures based on the fact that I like them and if they would look good on my Instagram". This is related to the reasons why participants geotag their posts. A common reason is that participants want to show the special locations they have been to, or as one participants expresses it: "To brag about being in places". Theoretically, if participants do not qualify a place as special enough to geotag the post, they might also not post about it at all, which makes the set of personal posts less complete.

If people do not qualify a place as special enough to geotag the post, but still place a post about it on their Instagram page, the location might be revealed by implicit information in the description of the post. This is shown by the following quotes of participants:

> "I usually geotag the location, when I don't mention my location in the text under the post."

> "When I do this [add a description to a post], I usually mention the town where the picture was taken. Sometimes in text, sometimes in hashtags. There is no particular reason for this I guess."

Summary

The precision of the data from personal posts is high. This is an expected result, because participants post about their personal events of which they are the expert, for which reason there should be a high correspondence between reconstructed stop places and actual events. The recall (completeness) on the other hand is low. Completeness could be better if implicit information were extracted from personal posts, because some participants indicate that they sometimes mention the location in the description of a post when they do not explicitly geotag the event they post about.

4.2.6 Quality of reconstructed journeys

Table 4.9: Summarised results for quality of reconstructed journeys

Question	**Explicit result**	**Implicit result**
How many of these journeys correspond to journeys you actually made?	$Mean = 58.5\%$ $Median = 60.0\%$ $Std.Dev = 40.5\%$	N.a.
How many journeys are lacking?	$Mean = 40.3\%$ $Median = 33.3\%$ $Std.Dev = 33.8\%$	N.a.
For the journeys that are incorrect, can you identify the error?	Errors in stop place sequence	N.a.
How many of the stops are shown in the correct sequence?	$Mean = 41.5\%$ $Median = 21.4\%$ $Std.Dev = 38.5\%$	N.a.

On average, participants indicate that 58.5% of the clustered stop places correspond to journeys they actually made. 33.3% of the participants indicate that 100% of the clustered stop places correspond to journeys they actually made, and 20% of the participants indicate that 0% of the clustered stop places correspond to journeys they actually made.

On average, a 40.3% recall is reported by participants. This means that 40.3% of the journeys participants conducted in total is not accounted for in their geotagged personal posts. The recall of clustered journeys is probably less than the recall of overall stop places. The chances that a participant posts once during an entire journeys are higher than the chances of a participant posting once in every stop place. For example, a participant mentions that "I generally post 1 photo of a trip but in reality I have been to a lot of places during that trip. [...] I haven't posted about the same trip twice". However, to be able to be able to make a statement about this, further research about the recall of stop places is needed.

Throwbacks are mentioned as a possible explanation for the error in journey clustering: "One journey that lasted three months has been split into separate journeys, but this also has to do with the throwback element. Some places are connected that were not part of the same journey"

and "I posted a picture half a year after I visited the location (the so-called throwback)". This is also revealed by the amount of stop places that are shown in the correct sequence, because on average only 41.5% of the stop places is shown in the correct sequence. Another error that is mentioned is that journeys are falsely split into separate parts. When examining the results, a relationship might be found between the percentage of stop places in the correct sequence, and errors in journey clustering. Further research with more personal posts per participant might also yield better results from the ST-DBSCAN algorithm.

Summary

Less than half of the reconstructed journeys are in the correct sequence, and as participants indicate, throwbacks are of influence on the journeys shown in the reconstruction. Therefore, it can be assumed that disruptions in the temporal aspect of data from personal posts are of major influence on the results yielded by the ST-DBSCAN algorithm. The less people use throwbacks, the more accurate the temporal dimension of the data is, the more accurately the ST-DBSCAN algorithm computes clusters in the track data. Furthermore, trajectories can be conceptualised in different ways among participants. This means that they would split the track data into different journeys than either the ST-DBSCAN or other participants would, because of their individual perspective on the concept of journeys. This results in a larger standard deviation, but does not necessarily inflict erroneous results.

4.3 Evaluation summary

This section discusses the quality dimensions that are most important to the overall quality of social media data from personal posts for the purpose of journey reconstruction.

One of the most important quality aspects that influence the overall quality of social media data from personal posts, is embodied by the temporal attribute. This aspect influences the quality of all quality dimensions, except for *spatial accuracy of reconstructed stop places*, *spatial resolution of reconstructed stop places* and *completeness and precision of reconstructed stop places*. First of all, the temporal aspect influences *location privacy*, because the less accurate the timestamp due to the fact that participants tend to wait between an event and posting about this event, the less participant's location privacy is endangered. On a less positive note, the inconsistencies in the temporal attribute cause incorrectness of attributes and disarrange the

sequence of stop places. Due to this, the ST-DBSCAN algorithm cannot properly cluster stop place posts into correct journeys. Therefore, the more inaccuracies in the temporal aspect - for example due to the use of throwbacks - the lesser the quality of *spatio-temporal accuracy of reconstructed stop places* is. Furthermore, the lesser the quality of this quality dimension, the less the ST-DBSCAN algorithm is able to properly cluster journeys, for which reason the *quality of clustered journeys* will be low. However, a low quality of clustered journeys will obscure the participants whereabouts, which in its turn increases their location privacy. The *spatio-temporal accuracy of reconstructed stop places* is therefore an important quality dimension, as it influences both the *quality of reconstructed journeys* and the *location privacy*.

Opposite to the temporal aspect, the quality of the spatial aspect scores quite well. Participants generally tend to correctly geotag their posts, with respect to the goal of journey reconstruction, whereas other applications might demand stricter requirements regarding the spatial accuracy of geotags. In case participants happen to incorrectly geotag their posts, the error is not so large that it negatively influences the journey reconstruction. The *spatial accuracy of reconstructed stop places* is therefore not of as much influence in the overall quality of journey reconstructions based on social media personal posts. However, in order to control the quality, it is necessary to test the spatial accuracy and also the spatial resolution, as incorrectly placed geotags do have some effect on the personal satisfaction of the journey reconstruction.

In short, the quality dimension that is most influential on the quality of journey reconstructions from personal posts is *spatio-temporal accuracy of reconstructed stop places*. Furthermore, there is a negative relationship between *location privacy* and all other quality dimensions. This means that the lower the quality of the other quality dimensions, the better the location privacy.

Chapter 5

Conclusion

This research is based on the following main question:

> **To what extent is the quality of geolocated social media posts sufficient for the reconstruction of journeys?**

In conclusion, this chapter first covers the answers to the defined subquestions, after which the answer to this main question is substantiated. Thereafter, the strengths and limitations of this thesis research are covered. Lastly, recommendations for future research are presented.

5.1 Answers

First, a short description of the answer to each of the subquestions is given. Thereafter, the answer to the main question is substantiated.

5.1.1 Subquestions

What is the spatial data quality of personal posts in social media?
The spatial data quality of personal posts in social media is not necessarily trustworthy, as it depends on multiple factors that are embodied as quality dimensions in this research. Out of these factors, the correctness of the time attribute is most unstable. The incorrectness of the attribute influences other quality dimensions such as the *quality of clustered journeys* and

location privacy. The spatial accuracy is in most cases sufficient for journey reconstruction, as half of the geotags are positioned on the correct location, and the other half is generally close to the original location.

How to reconstruct journeys from social media personal posts?
The API delivers the content of participants' social media personal posts in a JSON format. After retrieving this JSON data, the necessary attributes need to be extracted. These attributes are the timestamp, the latitude and longitude, and the location name. Together, these attributes form a track, based on which the whereabouts of a participant can be visualised. The start and end points of journeys can be extracted by the output of the ST-DBSCAN algorithm, which is described in the next paragraph. It is not yet possible to extract home locations of participants, due to the API limit of 20 personal posts.

What GIS methods are applicable to reconstruct a journey?
Initially, GIS methods were thought to be needed for the distinction between stops and moves in the personal post data. However, due to the sparse collection of personal post data, there are no suitable methods to classify whether or not a personal post is made during a stop or a move. For this reason, all personal posts are categorised as stops. Furthermore, the ST-DBSCAN algorithm is the most applicable clustering algorithm to be used to cluster the stop places in order to recreate journeys. The ST-DBSCAN algorithm is a spatio-temporal adaptation of the DBSCAN algorithm, and clusters points based on both spatial and temporal thresholds, whilst allowing the occurrence of noise points (outliers).

What privacy risk of social media data is incurred by the quality of it?
The better the overall quality of the data from personal posts, the bigger the privacy risk. Privacy risks can entail the ability to locate participants at the correct time and place, to find the home location of the participant, or to be able to know whether or not the participant is currently at home. The likeliness of these situations to occur diminishes along with decreasing quality. In this case, the privacy risk is not high. Due to the large differences between participants concerning the time between events and posts, it is difficult to estimate the exact time of a stop place. Therefore, it is difficult for possible attackers to find out exactly where an individual is at a given time. Nevertheless, the reconstruction of journeys does give insight into where an individual has been exactly, which might inflict privacy risk from other perspectives.

5.1.2 Main question

The extent to which the quality of geolocated social media personal posts is sufficient for the reconstruction of journeys in this research largely depends on the correctness of the time attributes. The spatial accuracy is a rather stable factor, as most geotags are accurate, and the geotags that are inaccurate are not too far from the original place that was visited for the reconstruction of journeys. The factor that varied most among participants, and that mostly caused the differences in quality of the reconstructions is the time attribute. The participants that often used throwbacks in their post history tend to have a reconstruction of lower quality than the participants who mainly post in the right sequence and that post shortly after the event they post about happened. Furthermore, the precision of the data from personal posts is high, as almost all posts are about events in places that actually occurred. On the other hand, the data from personal posts is not at all complete, as participants do not post and geotag every place they visit. Moreover, recall errors could occur to participants during the course of filling in the questionnaire. The journey reconstruction is about events that happened long ago, and participants might not correctly recall the exact details. This could inflict a bias in the results of the evaluation, and therefore a bias in the journey statistics.

In conclusion, the quality of social media personal posts is sufficient to some extent, depending on the purpose it is used for. The quality of journey reconstructions could be increased by the ability to handle cases of throwbacks, and by having more complete post histories.

5.2 Limitations

A limitation of this research was the limit of the Instagram API, that only allowed for the last 20 posts to be accessed. Due to this API limit, many personal posts could not been accessed, even though having access to full post histories would be beneficial for the research. For example, the questionnaire revealed that home locations were less easy to find based on the last 20 posts, compared to the full post histories of participants. Furthermore, the more geotagged personal posts per participant, the better the ST-DBSCAN algorithm is able to cluster stop places, and the higher the quality of journey reconstructions will be.

Another limitation is the conceptual difference between space and place. For participants, the notion of place is more important when assessing the positional accuracy in the questionnaire.

The reason for this is that this is their way to know where they have been, as participants rarely think of a location as a set of coordinates, but rather think of a location as an experience. This poses a difficulty for the assessment of data, because spatial accuracy should be assessed in terms of space.

Furthermore, limitations are posed by the technical aspect of participating in this research. Due to the way in which the API operates, participants are supposed to create a developer account and accept multiple invites in order to be able to participate. The process of participation is not simple, for which reason participation is not encouraged.

5.3 Recommendations for future research

First of all, a recommendation for future research is to test the quality of social media personal posts based on a larger participant sample, with better access to their post histories. More data will likely enable a better quantification of both the quality of journey reconstructions and questionnaire results.

Another recommendation is to not only take explicit geotags into account in the localisation of personal posts, but to also find a way to use the implicit location information of a personal posts that can for example be found in hashtags and captions.

The privacy risk of geotagging could be further explored by using alternate attack strategies to assess the vulnerability of location privacy inflicted by the use of geotags. For example, the possibility of revealing an individual's home location based on geotags could be researched more thoroughly.

Furthermore, after the quality of journey reconstructions from personal posts has been established, interesting results could be yielded by comparing journey reconstructions against the background of participant characteristics.

References

Ali, A. L., & Schmid, F. (2014). Data Quality Assurance for Volunteered Geographic Information.. doi: 10.1007/978-3-319-11593-1{_}9

Alvares, L. O., Bogorny, V., Kuijpers, B., Moelans, B., Antonio, J., Macedo, F. D., & Palma, A. T. (2007). Towards Semantic Trajectory Knowledge Discovery. *Data Mining and Knowledge Discovery*.

Barron, C., Neis, P., & Zipf, A. (2014). A Comprehensive Framework for Intrinsic OpenStreetMap Quality Analysis. *Transactions in GIS*. doi: 10.1111/tgis.12073

Birant, D., & Kut, A. (2007). ST-DBSCAN: An algorithm for clustering spatial-temporal data. *Data and Knowledge Engineering*. doi: 10.1016/j.datak.2006.01.013

Bordogna, G., Carrara, P., Criscuolo, L., Pepe, M., & Rampini, A. (2014). A linguistic decision making approach to assess the quality of volunteer geographic information for citizen science. *Information Sciences*. doi: 10.1016/j.ins.2013.07.013

Büscher, M., & Urry, J. (2009). Mobile methods and the empirical. *European Journal of Social Theory*. doi: 10.1177/1368431008099642

Capineri, C., Haklay, M., Huang, H., Antoniou, V., Kettunen, J., Ostermann, F., & Purves, R. (2016). *European Handbook of Crowdsourced Geographic Information*. doi: 10.5334/bax

Chen, S., Yuan, X., Wang, Z., Guo, C., Liang, J., Wang, Z., ... Zhang, J. (2016). Interactive Visual Discovering of Movement Patterns from Sparsely Sampled Geo-tagged Social Media Data. *IEEE Transactions on Visualization and Computer Graphics*. doi: 10.1109/TVCG.2015.2467619

Cooper, A. K., Coetzee, S., Kaczmarek, I., Kourie, D. G., & Iwaniak, A. (2011). Challenges for quality in volunteered geographical information. *AfricaGEO 2011*. doi: 10.1080/14498596.2014.927337

Cooper, A. K., Coetzee, S., & Kourie, D. G. (2012). Assessing the quality of repositories of volunteered geographical information. *GISSA Ukubuzana 2012*.

Criscuolo, L., Carrara, P., Bordogna, G., Pepe, M., Zucca, F., Seppi, R., ... Rampini, A. (2016). Handling quality in crowdsourced geographic information. *European Handbook of Crowdsourced Geographic Information*. doi: http://dx.doi.org/10.5334/bax.e

Curry, M. (1999). Rethinking Privacy in a Geocoded World. In *Geographical information systems: principles, techniques, management and applications.*

Danezis, G., Lewis, S., & Anderson, R. (2005). How Much Is Location Privacy Worth. *Fourth Workshop on the Economics of Information Security*. doi: 10.1111/j.1747-0285.2011.01230 .x

Devillers, R., & Jeansoulin, R. (2010). *Fundamentals of Spatial Data Quality.* doi: 10.1002/ 9780470612156

Dimond, M., Smith, G., & Goulding, J. (2013). Improving route prediction through user journey detection. In *Proceedings of the 21st acm sigspatial international conference on advances in geographic information systems - sigspatial'13.* doi: 10.1145/2525314.2525464

Duckham, M., & Kulik, L. (2006). Location Privacy and Location-Aware Computing. In R. Billen, E. Joao, & D. Forrest (Eds.), *Dynamic & mobile gis: Investigating change in space and time* (chap. 3). Boca Raton: CRC Press.

Ellegård, K., & Svedin, U. (2012). Torsten Hägerstrand's time-geography as the cradle of the activity approach in transport geography. *Journal of Transport Geography*. doi: 10.1016/ j.jtrangeo.2012.03.023

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). Density-Based Algorithm for Discovering Clusters. *KDD-96*. doi: 10.1016/B978-044452701-1.00067-3

European Commission. (2018a). *2018 reform of EU data protection rules.* Retrieved from https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/ data-protection/2018-reform-eu-data-protection-rules_en

European Commission. (2018b). *What constitutes data processing?* Retrieved from https://ec.europa.eu/info/law/law-topic/data-protection/reform/ what-constitutes-data-processing_en

European Commission. (2018c). *What does the General Data Protection Regulation (GDPR) govern?* Retrieved from https://ec.europa.eu/info/law/law-topic/data-protection/ reform/what-does-general-data-protection-regulation-gdpr-govern_en

European Commission. (2018d). *What your company must do.* Retrieved from https://ec .europa.eu/justice/smedataprotect/index_en.htm

Flanagin, A. J., & Metzger, M. J. (2008). *The credibility of volunteered geographic information.*

doi: 10.1007/s10708-008-9188-y

Goodchild, M. F. (2007). *Citizens as sensors: The world of volunteered geography.* doi: 10.1007/s10708-007-9111-y

Goodchild, M. F., & Li, L. (2012). Assuring the quality of volunteered geographic information. *Spatial Statistics*, *1*, 110–120. doi: 10.1016/j.spasta.2012.03.002

Granell, C., & Ostermann, F. O. (2016). Beyond data collection: Objectives and methods of research using VGI and geo-social media for disaster management. *Computers, Environment and Urban Systems*. doi: 10.1016/j.compenvurbsys.2016.01.006

Guc, B., May, M., Saygin, Y., & Körner, C. (2008). Semantic Annotation of GPS Trajectories. In *11th agile international conference on geographic information science.* doi: 10.1016/j.compbiomed.2008.02.004

Haklay, M. M., Basiouka, S., Antoniou, V., & Ather, A. (2010). How Many Volunteers Does it Take to Map an Area Well? The Validity of Linus' Law to Volunteered Geographic Information. *The Cartographic Journal*. doi: 10.1179/000870410X12911304958827

Hall, C. M. (2005). *Reconsidering the geography of tourism and contemporary mobility.* doi: 10.1111/j.1745-5871.2005.00308.x

Hall, C. M., & Page, S. J. (2003). *The geography of tourism and recreation: Environment, place and space: Second edition.* doi: 10.4324/9780203246276

Hochmair, H. H., & Zielstra, D. (2012). Positional Accuracy of Flickr and Panoramio Images in Europe. In *Gi_forum 2012: Geovizualisation, society and learning.*

Høyer, K. G. (2000). Sustainable tourism or sustainable mobility? The norwegian case. *Journal of Sustainable Tourism*. doi: 10.1080/09669580008667354

Hu, Y., Janowicz, K., Carral, D., Scheider, S., Kuhn, W., Berg-Cross, G., ... Kolas, D. (2013). A geo-ontology design pattern for semantic trajectories. In *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics).* doi: 10.1007/978-3-319-01790-7-24

Iachello, G., Smith, I., Consolvo, S., Abowd, G. D., Hughes, J., Howard, J., ... LaMarca, A. (2010). Control, Deception, and Communication: Evaluating the Deployment of a Location-Enhanced Messaging Service.. doi: 10.1007/11551201{_}13

Instagram. (2018). *Location Endpoints.* Retrieved from https://www.instagram.com/developer/endpoints/locations/

Jilani, & Corcoran, P. (2013). Automated Quality Improvement of Road Network in OpenStreetMap. *Agile Workshop (Action and Interaction in Volunteered Geographic Informa-*

tion), 3–6.

Kahle, K., Sharon, A. J., & Baram-Tsabari, A. (2016). Footprints of fascination: Digital traces of public engagement with particle physics on CERN's social media platforms. *PLoS ONE*. doi: 10.1371/journal.pone.0156409

Kisilevich, S., Krstajic, M., Keim, D., Andrienko, N., & Andrienko, G. (2010). Event-based analysis of people's activities and behavior using Flickr and Panoramio geotagged photo collections. In *Proceedings of the international conference on information visualisation.* doi: 10.1109/IV.2010.94

Kisilevich, S., Mansmann, F., Nanni, M., & Rinzivillo, S. (2010). Spatio-temporal clustering. In *Data mining and knowledge discovery handbook.* doi: 10.1007/978-0-387-09823-4{_}44

Krumm, J. (2007). Inference Attacks on Location Tracks. In *Pervasive computing.* doi: 10.1007/978-3-540-72037-9{_}8

Li, M., Zhu, H., Gao, Z., Chen, S., Ren, K., Yu, L., & Hu, S. (2014). All Your Location are Belong to Us: Breaking Mobile Social Networks for Automated User Location Tracking. In *15th acm international symposium on mobile ad hoc networking and computing* (p. 43 - 52). Philadelphia.

Long, J. A., & Nelson, T. A. (2013). *A review of quantitative methods for movement data.* doi: 10.1080/13658816.2012.682578

Madhulatha, T. S. (2013). AN OVERVIEW ON CLUSTERING METHODS. *IOSR Journal of Engineering*. doi: 10.9790/3021-0204719725

Mooney, P., Corcoran, P., & Winstanley, A. C. (2010). Towards quality metrics for OpenStreetMap. In *Proceedings of the 18th sigspatial international conference on advances in geographic information systems - gis '10.* doi: 10.1145/1869790.1869875

Moreno, B. N., Times, V. C., Renso, C., & Bogorny, V. (2010). Looking inside the stops of trajectories of moving objects. In *Proceedings of the brazilian symposium on geoinformatics.*

Neis, P., Zielstra, D., & Zipf, A. (2011). The Street Network Evolution of Crowdsourced Maps: OpenStreetMap in Germany 2007–2011. *Future Internet*, *4*(4), 1–21. Retrieved from http://www.mdpi.com/1999-5903/4/1/1/ doi: 10.3390/fi4010001

Palma, A. T., Bogorny, V., Kuijpers, B., & Alvares, L. O. (2008). A clustering-based approach for discovering interesting places in trajectories.. doi: 10.1145/1363686.1363886

Rocha, J. A. M., Times, V. C., Oliveira, G., Alvares, L. O., & Bogorny, V. (2010). DB-SMoT: A direction-based spatio-temporal clustering method. In *2010 ieee international conference on intelligent systems, is 2010 - proceedings.* doi: 10.1109/IS.2010.5548396

Roick, O., & Heuser, S. (2013). *Location based social networks - definition, current state of the art and research agenda.* doi: 10.1111/tgis.12032

Scheider, S. (2019). Obfuscating spatial point tracks with simulated crowding. *Manuscript submitted for publication*.

Scheider, S., & Janowicz, K. (2014). Place reference systems:: A constructive activity model of reference to places. *Applied Ontology*. doi: 10.3233/AO-140134

Senaratne, H., Mobasheri, A., Ali, A. L., Capineri, C., & Haklay, M. M. (2017). *A review of volunteered geographic information quality assessment methods.* doi: 10.1080/13658816 .2016.1189556

Siła-Nowicka, K., Vandrol, J., Oshan, T., Long, J. A., Demšar, U., & Fotheringham, A. S. (2016). Analysis of human mobility patterns from GPS trajectories and contextual information. *International Journal of Geographical Information Science*. doi: 10.1080/13658816.2015 .1100731

Spaccapietra, S., Parent, C., Damiani, M. L., de Macedo, J. A., Porto, F., & Vangenot, C. (2008). A conceptual view on trajectories. *Data and Knowledge Engineering*. doi: 10.1016/ j.datak.2007.10.008

Spielman, S. E. (2014). Spatial collective intelligence? credibility, accuracy, and volunteered geographic information. *Cartography and Geographic Information Science*. doi: 10.1080/ 15230406.2013.874200

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

Story, R. (2013). *Folium.* Retrieved from https://python-visualization.github.io/ folium/

Sui, D., & Goodchild, M. (2011). *The convergence of GIS and social media: Challenges for GIScience.* doi: 10.1080/13658816.2011.604636

Taylor, P. J. (1999). Places, spaces and Macy's: Place - space tensions in the political geography of modernities. *Progress in Human Geography*. doi: 10.1191/030913299674657991

Thielmann, T. (2010). Locative Media and Mediated Localities: An Introduction to Media Geography. *Aether. the Journal of Media Geography*.

Turner, A. J. (2006). *Introduction to Neogeography.*

Van Exel, M., Dias, E., & Fruijtier, S. (2010). The impact of crowdsourcing on spatial data quality indicators. In *Proceedings of the 6th giscience international conference on geographic information science.*

Xia, J. (2012). Metrics to measure open geospatial data quality. *Issues in Science and Technology Librarianship*. doi: 10.5062/F4B85627

Xie, K., Xia, C., Grinberg, N., Schwartz, R., & Naaman, M. (2014). Robust detection of hyper-local events from geotagged social media data.. doi: 10.1145/2501217.2501219

Yin, Z., Cao, L., Han, J., Luo, J., & Huang, T. (2013). Diversified Trajectory Pattern Ranking in Geo-Tagged Social Media.. doi: 10.1137/1.9781611972818.84

Zargar, A., & Devillers, R. (2009). An operation-based communication of spatial data quality. In *Proceedings of the international conference on advanced geographic information systems and web services, geows 2009.* doi: 10.1109/GEOWS.2009.8

Zheng, Y. (2012). Tutorial on Location-Based Social Networks. In *proceeding of the 21st international conference on world wide web.*

Zheng, Y., Zhang, L., Xie, X., & Ma, W.-Y. (2009). Mining interesting locations and travel sequences from GPS trajectories. In *Proceedings of the 18th international conference on world wide web - www '09.* doi: 10.1145/1526709.1526816

Zipf, A., Mobasheri, A., Rousell, A., & Hahmann, S. (2016). Crowdsourcing for individual needs – the case of routing and navigation for mobility-impaired persons. In *European handbook of crowdsourced geographic information* (pp. 325–337). Retrieved from http://www.ubiquitypress.com/site/chapters/10.5334/bax.x/ doi: 10.5334/bax.x