GIMA
Geographical Information Management and Applications

Final Thesis Report

Crawling the world wide web to find hidden patterns of people

A proof-of-concept of using UGC data to investigate spatial-temporal patterns in Rotterdam & Veere

Author

Bram Roozen

6001297

b.roozen@uu.nl

Supervisor

Dr. Egbert van der Zee

Professor

Prof. Dr. Stan Geertman

26th of February 2021

Universiteit Utrecht
TU Delft
University of Twente.
Wageningen UR

Final Thesis Report

Crawling the world wide web to find hidden patterns of people

A proof-of-concept of using UGC data to investigate spatial-temporal patterns in Rotterdam & Veere

Author

Bram Roozen

6001297

b.roozen@uu.nl

Supervisor

Dr. Egbert van der Zee

Professor

Prof. Dr. Stan Geertman

26th of February 2021

Preface

In this document I present my Master thesis entitled “Crawling the world wide web to find hidden patterns of people: A proof-of-concept of using UGC data to investigate spatial-temporal patterns in Rotterdam & Veere”. It has been written for the M.Sc. program Geographical Information Management & Applications (GIMA), which is a cooperation between Utrecht University, Technical University of Delft, Technical University of Twente, and Wageningen University & Research.

First and foremost, I would like to express my gratitude to my supervisor Egbert van der Zee for sharing his feedback and infectious enthusiasm about every detail of the subject. Second, I would like to thank Christian and Daan, with whom I have spent many hours of time on the code for obtaining the required data. I also want to thank Stan Geertman for sharing his wisdom and critical feedback on the project, Remco Bron for his expert opinion, and Manon for proofreading the report.

Last, but certainly not least, I would also like to thank my fellow students for the regular online coffee breaks to still provide support regardless of the ongoing pandemic. I hope you will enjoy the read.

Bram Roozen, 26th of February 2021

Abstract

As cities get bigger and more crowded, questions arise on how to deal with the issues and challenges that come paired with it. Knowledge of the exact situation may help to get an insight into the situation and potentially adjust policies. For this, it is required to gain understanding in the spatial and temporal behaviour of individuals. Simply put, an answer is needed to the question “Where are people, and when are they there?” Over the last years, advancements in data collection, storage and analysis have been made. People are posting pictures, sharing information, and reviewing meals when they have been at a certain place. All this data is stored in the social media applications where they were posted to. When the posts are provided with a geotag and a timestamp, this geosocial data could be used to trace the digital footsteps of individuals.

Therefore, the aim of this study is to investigate how different sources of geosocial data can be used to visualise both spatial and temporal patterns of individuals. As this study will be a proof-of-concept of using geosocial data, the study area will consist of an urban area (Rotterdam) and a more rural area (Veere). By doing a thorough investigation, the choice is made for using Instagram and TripAdvisor as geosocial data sources. After an extensive data collection and conversion process, the data can be used for analysis. Displaying the descriptive statistics for both the temporal and spatial results reveal many of the behavioural patterns of people. To add to that, the Moran’s I has been calculated to check for spatial autocorrelation and a hot spot analysis has been performed. This came to numerous interesting results, among which an interesting visualisation on an invisible boundary around the city centre of Rotterdam. There are however some ifs and buts when working with geosocial data. Concluding, it can be stated that where temporal patterns are better visible in rural Veere, the spatial patterns are better visible in urban Rotterdam. The degree of urbanity must therefore be considered when doing research in this area.

Keywords: Web scraping I UGC I Data collection I Instagram I TripAdvisor I Spatial-temporal patterns I Hot spot analysis

Contents

I. List of abbreviations

API	Application Programming Interface
CDR	Cellular Data Records
CSV	Comma-separated values
DMO	Destination Marketing Organisation
DOM	Document Object Model
GDPR	General Data Protection Regulation
GeoJson	Geographic JavaScript Object Notation
GIS	Geographic Information Systems
HTML	Hyper Text Markup Language
IDC	International Data Corporation
IP	Internet Protocol
IT	Information Technology
Json	JavaScript Object Notation
Jsonl	JavaScript Object Notation - lines
NBTC	Netherlands Board of Tourism and Conventions
ToS	Terms of Service
UGC	User Generated Content
URL	Uniform Resource Locator
POI	Point of Interest

1. Introduction

When are people where? Where are they going? What are they doing, and when are they doing it? An answer to these questions would give valuable insights in the movements of people in a city. As cities get bigger and more crowded, questions arise on how to deal with the issues and challenges that come with it (Mark, 2019). Knowledge of the situation may help with getting answers to these questions. As people visit places, the ‘research objects’ are already there. The main challenge is finding a way to collect information of people at places.

People share their own pictures on Instagram and review the meals they had on Tripadvisor. By doing so, they leave a digital footprint on the locations they have visited. Sharing these activities through social platforms has become a central element of contemporary tourism (Van der Zee & Bertocchi, 2018). With many people sharing their pictures and reviews with the world, an incredible amount of data is being openly published and stored every day, keeping track of the activity of many individuals around the world. The wide range of social media and other applications where User Generated Content (UGC) is stored, results in an enormous amount of data, full of potential information. Currently, the challenge is to find how this UGC data can specifically be of use to investigate where people are when.

As Peter Sonder, the Senior Vice President of multi-billion-dollar IT company Gartner said: “Information is the oil of the 21st century and analytics is the combustion engine.” With information he refers to big data and big data analytics. ‘Big data’ is a colourful marketing phrase for the significant change that occurred over the last years in data capture, storage, retrieval, and analysis (Power, 2014). People generate data, whether this is through navigation applications, uploading pictures to websites, online purchases, posting, liking, tweeting, commenting and other activities. UGC of these applications on itself is not immediately useful. However, when combined with analytics, data could provide numerous opportunities (Miah et al., 2017).

UGC can for example give more opportunities for modelling reality (Albuquerque, Costa & Martins, 2018). The massive growth in social media usage makes it possible to get insights in the behaviour of people (Harrigan, Evers, Miles &

Daly, 2017). This could be beneficial for crowd management in historic city centres and other public places. As social media is currently used by over 3.6 billion monthly users, this development has increased the amount of available data on travel behaviour in cities (Giglio, Bertacchini, Bilotta & Pantano, 2019; Statista, 2020; Xiang & Gretzel, 2010). Instead of conducting surveys or interviews with people, the data can now be gathered from the actions on social media of users by retrieving UGC from these applications. UGC data can have the potential of altering behaviour and assist local governments, municipalities, and Destination Management Organisations (DMO’s) in data-driven decision making (Ganzaroli, De Noni & Van Baalen, 2017).

Besides regular UGC, data also often contains a geographical component. As Franklin & Hane already stated in 1992, eighty percent of all data has a spatial component. Nowadays, nearly thirty years later, the amount of data, including spatial data has significantly increased. Social media and UGC applications often concern a certain location. Pictures that are posted on social media can be geotagged, showing the location of where the picture was taken. Also, restaurant reviews can easily be traced back to the exact place of the restaurant. This all entails geosocial data. Geosocial data is social media data including a spatial component. While an individual is sharing its experiences, a data trail is left behind, enabling even more possibilities with this geosocial data (Heaton et al., 2019).

Another interesting component of geosocial data is the timestamp that is often provided with a post or review. This timestamp caters for the possibility to exactly capture the moment that a person was at a place. Among many other functionalities, this temporal component makes it possible to keep track of when there are people at certain locations. Doing so, gives the opportunity to investigate bottlenecks and peak months, weeks, days, or even hours. With both location information and temporal information, the question on where people are, could possibly be answered.

Unfortunately, geosocial data also comes with a few downsides. Obtaining the data could be a tough job, especially since it must comply with the General Data Protection Regulation (GDPR), the European privacy law that has come into effect in 2018. Also, several UGC applications prohibit the process of content collection through web crawlers, as it cannot guarantee the privacy of its users. Despite that, there are ways to successfully obtain geosocial data from websites, as it is often not the identity of a user which is of interest, but just the content, location and/or timestamp of an anonymous user to investigate when people are where, where people are going, and what they are doing when.

The goal of this thesis is therefore to create a proof-of-concept on finding both spatial and temporal patterns of people with the help of UGC data. Hereby, there will also be looked at the ifs and buts that occur during the process, and how representative the results are compared to reality.

1.1 Scientific relevance

Numerous studies have already shown that geosocial data can be translated to useful information to predict the behavior of individuals (Brandt, Bendler & Neumann, 2017; Chua, Servillo, Marcheggiani & Moere, 2016; Giglio et al., 2019; Vu, Li, Law & Zhang, 2018; Wu, Huang, Peng, Chen & Liu, 2018). Despite of these research projects being useful, each of these studies investigates one social media platform. Where Chua et al. (2016) and Brandt et al. (2016) make use of Twitter data, Giglio et al. (2019) and Wu et al. (2018) make use of Flickr data. The similarity between these research projects can be found in UGC with a geospatial component that can be traced back (anonymously) to individuals in space. Vu et al. (2018) have a different approach. Their research retrieved location data of venues where people have checked in. This could be either a restaurant or another Point of Interest (POI). The more check-ins a location has, the more crowded a place is in general. The obtained results could then give insight in the distribution of people across different locations (Ganzaroli et al., 2017). As can be derived from previous research projects, there are two ways to investigate the moving behavior of people. This could be either (1) using the exact locations of individuals or (2) using the exact locations of multiple POI's where individuals have posted a message or review.

What all mentioned research papers have in common is the use of geosocial data. While the results of these papers give a useful insight in crowd prediction, one might question the representativeness of reality. The aforementioned research papers have made use of one data source to base their results on.

The novelty of this particular research project is that it will shed light on new ways of data collection and analysis by making use of two main sources instead of one. Also, there will be looked at both the individual level and on POI level. One of the aims of this research project is to use multiple sources in such way that a better representation of reality can be simulated. To achieve this goal, geosocial data will be collected from two different web applications containing UGC data.

1.2 Societal relevance

Aside from the scientific relevance and technical possibilities, there is also a societal relevance to this research project. Gaining insight in where people are when has already been a challenge for the past years (Miah et al., 2017; Milano et al., 2019; Weber et al., 2017). The Netherlands Board of Tourism and Conventions (NBTC) published a document explaining their 2030 perspective on tourism in The Netherlands. While their focus is particularly on tourism, their research and vision on where people are in places is just as relevant for this research project. Recent discussions in scientific literature are debating the term ‘crowdedness’ (Koens, Postma & Papp, 2018). Negative effects of crowded areas could be criminality, less traffic flows for emergency services, and a lower mental health status of users of a place (McDonnell & MacGregor-Fors, 2016). However, what is crowdedness? Where does it come from? In the everyday discussion, crowdedness is a hot topic. But in order to understand crowdedness, there is a need to first understand how and when a place is being used by its users.

One of the research questions of the NBTC is how to measure tourist behaviour and how to determine the factors that decide whether areas are overcrowded with tourists or not. This research project will provide guidance to their question by investigating whether the analysis of geosocial data can be used to investigate underlying patterns in the behaviour of people.

Further reasons why this research is currently relevant is the global COVID-19 pandemic. The topic of crowdedness has gained increased attention over the last couple of months. People are staying home more often, and general urban tourism has experienced a massive decline. When having a closer look, it becomes clear that while urban tourism is experiencing a decrease in numbers, rural tourism is experiencing an increase (Nepal, 2020). These patterns can also be observed with the help of the collected geosocial data, providing answers to where the pandemic causes people to move to.

2. Research objectives

The main scope of this research project will be to investigate different methods to visualise spatial and temporal patterns in where people are when. This research will give a proof-of-concept of this in both an urban area (Rotterdam) and a rural area (Veere). First, the main research question will be formulated. After that, this chapter will state the concrete sub-questions and also elaborate on the research limitations.

Research question:

How can different sources of user generated geosocial data be used to visualise spatial and temporal patterns of individuals in Rotterdam & Veere?

The aim is to answer the research question by answering the sub-questions below.

Sub-questions:

SQ1: What methods can be used to obtain geosocial data and what geosocial data is available and usable?

SQ2: How can geosocial data be analysed to visualise spatial and temporal patterns of people?

SQ3: To what extent can geosocial data be used to display an accurate model compared to reality?

2.1 Research limitations

Before answering these research questions, it is important to also state the research limitations by specifying the scope of the project. The main goal is to visualise spatial and temporal patterns of individuals in places with the help of geosocial data, where geosocial data is a broad term that might require some extra elaboration.

The idea behind the use of geosocial applications, is that there exists a certain similarity between the different sources of geosocial data. Geosocial data is simply location-based social media data. It is content (often pictures and/or text) that is produced by people on social media platforms. The reason for choosing two different sources for this research project is that two data sources might give a more accurate result than just one source, as the results can be used to validate both sources. Also, the aim is to make use of a geosocial application that contains data of places, where people have checked in (i.e., a restaurant or POI), and an application that contains data directly from individuals. This assures the validity of the eventual model, by using a variety of methods to collect and analyse data on the same topic.

3. Theoretical framework

In this theoretical framework, the theoretical basis will be given for the research project. First, the rise of big data availability will be explained throughout the years. Second, there will be further elaborated on how data helps with decision making. Third, literature on maintaining the real-world representation while not having access to all data will be explained in further detail. To conclude this chapter, a conceptual model will be drafted, to structure the further course of the thesis.

3.1 The rise of big data availability

Surveys, books, papers, records. All of these are different types of data. Data has already been present for a long time and has only increased ever since. The invention of the printing press around the year 1440 introduced the 'era of mass communication', where books and documents could be copied and shared (Eisenstein, 1980). Since 1440, the revolution of data is still ongoing. Fast forwarding to the year 1983, a 'network of networks' was assembled and in that way the internet was founded. When, shortly thereafter, Tim Berners-Lee invented the World Wide Web in 1990, everyone with a computer could create, copy, store, and send data into the world. Currently, in the year 2021, the internet that we know is just over thirty years old. Numerous people have access to a computer and have access to a portable computer: the smartphone (Tstetsi & Rains, 2017). In the meantime, data has increased in a large scale over the past years. According to a report by the International Data Corporation (IDC) in 2018, the volume of created and copied data in the world was around 33 zettabytes, which is equal to 33 trillion gigabytes. The IDC also estimated that the expected number of zettabytes in 2025 will reach 175.

This is where the term big data comes in to use. As big data itself is an abstract concept, it is important to specify what exactly big data is in this context. Apart from masses of data, big data also has several other characteristics. In general, the term 'big data' is used for datasets that could not be perceived, acquired, managed, and processed by traditional IT and software/hardware tools within a tolerable time (Chen, Mao & Liu, 2014). A more comprehensive explanation is given through the five V's by Power (2014). The first and most obvious of all being that big data is characterised by its Volume, as it is much bigger than traditional datasets. Second, Velocity is given for the rapid speed with which the data is produced. Third, there is the Variety of formats, which forms both challenges and opportunities when it comes to data processing. Fourth, extending on the Variety, big data is Variable as it changes over time and it may originate from different sources. Lastly, as fifth, big data is also Volatile, meaning that the levels of production of the data are inconsistent. On some days there might be more data produced than on other days. The influences of the five V's combined all together results in the complex puzzle that big data is.

Knowing what big data is, is one thing, understanding big data comes next. Therefore, this paragraph will delve deeper into the origin of big data. While there are numerous sources capable of providing 'big data'-worthy datasets, this research project will focus on one specific area of big data, which is UGC (Power, 2014). Just like other big data sources, the amount of UGC grew massively over the last years. This is mostly due to mobile technology advancements, providing individuals with the power to leave a 'digital footprint' wherever they perform online tasks (Lu & Stepchenkova, 2015). This could be anything from posting a picture to reviewing a hotel. Most forms of UGC can be found on social media applications such as Facebook, Flickr, Instagram, Tripadvisor and Twitter. Social media is defined as "a generic term for social interactions built on a multitude of digital media and technologies, which allows user to create and share content and to act collaboratively" (Schoder, Gloor & Metaxas, 2013).

Social media has already proven to be suitable for research in the field of tourism (Brandt, Bendler & Neumann, 2017; Chua et al., 2016; Giglio et al., 2019; Lu & Stepchenkova, 2015; Vu et al., 2018; Wu et al., 2018; Xiang & Gretzel, 2010). Despite the fact that this research project is taking the users of the city in a broader sense than just for tourists, the scientific literature on the use of UGC to discover spatial patterns of where tourists are also applies to more users in the city (Koens, Postma & Papp, 2018). A reason for this is that the use of the terminology 'crowdedness' is often associated with tourists, especially in urban areas and historic city centres (Van der Zee & Bertocchi, 2018). The further literature that will be used in this theoretical framework will therefore mainly have a tourist perspective on the use of UGC data. Often, this can be generalized to other users of a city as well, as a local that walks through the streets while posting a picture of a building, or enjoys a coffee at a café, is just as much part of the number of people in a city (Koens et al., 2018).

Social media itself is filled with more information than just reviews, posts, and comments. Multiple social media applications also include some sort of geotags and timestamps to data that has been generated by a user. With the timestamp, user generated entries are provided with the time at which they were uploaded. The so called "geotag" is an electronic tag that assigns a geographical location to the data. Therefore, a social media user can show where a picture was taken or where a post was written (Shelton, 2017). This is the moment where the use of GIS comes into place. With the addition of the geotag and timestamp, spatial-temporal analysis becomes possible with UGC data (Heaton et al., 2019).

With all the earlier mentioned aspects of UGC data, it becomes more valuable. Users are not only posting the content, but also their location and timestamp in an application. Most applications have therefore a built-in option to keep your data private, or to a select group of followers, but other data is publicly available online (Puebla, 2018). The progress of collecting this data is nearly impossible to perform manually, and therefore several methods for web scraping have been invented to automate the data collection. Still, researchers need to be aware that content owners or applications may have web scraping/data harvesting restrictions (Jennings & Yates, 2009).

Numerous scientific articles argue that web scraping is a grey area when it comes to ethics or even legality (Alaei, Becken & Stantic, 2019; Bruns, 2019; Chen et al., 2014, Fiesler, Beard & Keegan, 2020; Johnson, 2012; Krotov & Silva, 2018; Smith, Szongott, Henne & Von Voigt, 2012). A literature research concludes that the privacy of users of UGC applications should always be guaranteed in the first place (Smith et al., 2012). Currently, the debate on ethics and legality mainly has its focus in the web scraping area. UGC applications often have a non-legally binding Terms of Service (ToS) included in their terms and conditions. Fiesler et al. (2020) investigated over a hundred ToS pages and their findings showed that while provisions on web scraping are common, they are also ambiguous, inconsistent, and lack context. Fiesler et al. (2020) concludes that by the nature of these provisions, ethical decision-making is required that moves beyond the ToS, also considering contextual factors, the data source, and the methodology of the research.

3.2 Data-driven decision making

Collecting and understanding big data for research on where people are when is quite the challenge already. What is important, is to also understand the goals that can be reached by big data analysis. This paragraph will therefore focus on how big data might assist local governments, municipalities, and DMO's with data-driven decision making. It will also examine the current body of scientific knowledge on data collection and analysis and elaborate on the specific research projects that already have taken place in the field of tourism and can be applied to UGC data studies in general.

Current issues in tourism often concern tourism overcrowding, or overtourism, as one of the biggest problems in cities. Most of the overcrowding issues occur in already crowded urban areas like Venice, Amsterdam and Barcelona (Araya Lopéz, 2020; Brandajs & Russo, 2019; Milano, Novelli & Cheer, 2019; Van Leeuwen, Klerks, Bargeman, Heslinga & Bastiaansen, 2020; Weber et al., 2017). Already in 1989, O'Reilly wrote an article on the concept and issues of the tourism carrying capacity. It is known that tourism overcrowding results in the destruction or near-destruction of historical landmarks and even of the natural environment (O'Reilly, 1986). O'Reilly advocated to include the tourism carrying capacity of an area in the planning of tourism by governments and DMO's, "despite the difficulties in tourism measurement". As mentioned before, it is difficult to research overcrowding by performing observations or conducting surveys on the street. With the rise of big data availability, there are now more advanced possibilities to research overcrowding.

To this day, the statement by O'Reilly still holds partially because of two reasons. First, the problem of overcrowding is still relevant, perhaps even more relevant than ever. There are however multiple articles questioning the frequent use of the term 'overtourism', while the main problem of places is mainly overcrowding in general (Koens et al., 2018; Robertson & Feick, 2016). Tourists clearly form a part of the issue; however local inhabitants should also be included in research on crowdedness. A second change that occurred over the years since the paper by O'Reilly (1986) is the growth in methodological possibilities (Huang et al., 2017). If data from the past can be mapped and analysed, crowdedness in the future could possibly be predicted (Huang et al., 2017). This knowledge can then be used for decision-making and planning purposes, as has been proven by Xia, Zeephongsekul & Arrowsmith (2009). In their research, a significant outcome was that DMO's can successfully be assisted by the information provided by the model which they created based on UGC data.

The goal of research on overcrowding is to eventually come up with patterns in the behaviour of individuals. This quest for finding similarities between UGC applications and places can be aided by big data (Giglio et al., 2019). Currently, many scientists are pioneering in the field of tourism with UGC data (Brandt et al., 2017; Chua et al., 2016; Ganzaroli et al., 2017; Giglio et al., 2019; Li, Goodchild & Xu, 2013; Miah et al., 2017; Van der Drift, 2015; Van der Zee & Bertocchi, 2018; Vu et al., 2018; Wood et al., 2013; Wu et al., 2018). On a methodological level, there has been a technical revolution resulting over the past years in data collection and new analytical tools in the field of geo information in general.

The outcomes from proceedings in the field of tourism can be adapted to reveal patterns of where people are when, and possibly even reveal who is at a certain location, revealing even more information on the background of users (Robertson & Feick, 2016). Therefore, the main challenge is to investigate the best way to analyse big datasets retrieved from UGC applications and how to translate the analyses to actual policy design and decision making. The aforementioned research projects are just a grasp from the many pioneering projects that have taken place in the field of geo information and human behaviour. The research projects use different methods and techniques to try and get a grasp on the behaviour of people. Table 3.1 shows the main information on each research project, which will also be further elaborated in the next paragraph.

Table 3.1: Authors, sources, datatypes, places and analysis methods used in several research projects.

Authors	***Source & Year***	***Datatype***	***Place***	***Analysis***
Ganzaroli, De Noni, Van Baalen	TripAdvisor 2017	Locations of restaurants with number of reviews	Venice, Italy	Descriptive, Temporal,
Chua et. al.	Twitter 2016	Geotagged Tweets (points)	Cilento, Italy	Descriptive, Flow map, Temporal
Wood et. al.	Flickr 2013	Geotagged Flickr photos (points)	Worldwide	Descriptive, Temporal (different timeframes), ANCOVA
Brandt, Bendler, & Neumann	Twitter 2017	Geotagged Tweets (points)	San Francisco, USA	Descriptive. Temporal, Kernel Density
Albuquerque, Costa, & Martins	Tourism databases 2018	Point data of hot spots	Portugal	Descriptive
Van der Zee & Bertocchi	TripAdvisor 2018	Locations of restaurants with reviews and profiles	Antwerp, Belgium	Clustering, Relational approach

Van der Drift	Flickr 2015	Geotagged Flickr photos (points)	Amsterdam, Netherlands	Descriptive, Temporal, Hot spot analysis, kernel density, clusters
Miah et. al.	Flickr 2017	Geotagged Flickr photos (points)	Melbourne, Australia	Temporal (Trends, Seasons), Descriptive
Huang, Zhang, Ding	Baidu/Google 2017	Temporal search data to predict overcrowding	China	Descriptive, Temporal, Granger causality
Giglio, Bertacchini, Bilotta, Pantano	Flickr 2019	Geotagged Flickr photos (points)	Six cities, Italy	Descriptive, Clusters, Content analysis

One of the first things that stand out when investigating table 3.1 is that most studies performed (at least) a descriptive analysis. Descriptive statistics can already show multiple patterns and shape an introduction to any further analysis, as can be seen in Van der Drift (2015), Ganzaroli, De Noni & Van Baalen (2017), and Brandt et al. (2017). All research projects show the amount of posts, in this case Tweets, TripAdvisor reviews and Flickr photos, per hour, per month, and over a whole year. Further analyses mostly concern a type of clustering analysis.

Another interesting find is that Flickr is frequently used as a data source. Reason for this could be their open Application Programming Interface (API), where users of this API can obtain data relatively easily. Twitter also has an API, however it might be less frequently used as opposed to Flickr. The cause for this may be that

Twitter has quite some noise in its data, as it is a frequently used application used by many different users, such as companies or news sources (Brandt et al., 2017; Chua et al., 2016). The location in the world does not seem to have an influence on the success of big data analysis. While there are certainly differences in results between places, big data research on the behaviour of people can happen anywhere, if there is enough geosocial data coverage on that place.

3.3 Modelling a real-world representation

Developing a model of reality is inevitable when researching tourism behaviour. Understanding what really happens in a certain place and getting to know where tourists are in a city is otherwise only possible by performing observations throughout several years and conducting short interviews on where every tourist in every city will go next. This methodology would require a lot of resources, which is why following the digital footprint of a tourist may be the best alternative to get an answer to the question of where tourists are in a city (Heaton, 2019). The downside of this, is that reality is being deducted to a more simplified model. Therefore, this paragraph will discuss the downsides and challenges that emerge when using UGC to map tourist flows.

Miah et al. (2017) explain that the observational methods that are closest to give a real-world representation rely on sensor data retrieved from cellular data records (CDR). Different information may be collected with the help of this technology. It contains information of cell phone activities which are created by cellular base stations (Gao, Liu, Wang & Ma, 2013). A location of a cell phone, which often is equal to one individual, can then be approximated by the particular base station (Miah et al., 2017). Although CDR might retrieve different types of information, there is always a location of the cell phone included and a timestamp, which is crucial information when investigating the question on 'when' people are 'where' in a certain place (Vu et al., 2018). This is also where the challenging part of CDR comes into place. Because of the detailed information, obtaining the data is most often impossible due to privacy regulations (Ghahramani, Zhou & Wang, 2020).

More distant from a real-world representation, but more realistic in terms of data collection and usage is where the aforementioned geosocial data comes into practice. Instead of collecting all CDR data from a wide range of applications, the data is being collected from a single data source. As geosocial applications with UGC data also include information on timestamps and locations, they offer an alternative to CDR data (Giglio et al., 2019). There are however multiple downsides to the use of datasets based on UGC.

The first issue is that there are multiple concerns on the background of users that post information on UGC applications (Albuquerque, 2018; Johnson, 2012; Van der Zee & Bertocchi, 2018). It is unclear what the motives are for posting on UGC. Due to privacy restrictions, the demographic background on users that are active on geosocial applications is not included in the collected data. At the same time, not every individual makes use of UGC applications. Even in the collected data, some users might post parenthetically, whilst others are considered frequent contributors (Alaei et al., 2019).

Second, because of the size and richness of the data, it is hard to translate it into a clear policy recommendation (Lu & Stepchenkova, 2015). The bigger a dataset is, the more possibilities there are to perform analysis and gain insights in tourist behaviour. However, at the same time, it will also get more technically and methodologically complex to collect and process the data (Van der Zee & Bertocchi, 2018).

Lastly, in contrary to CDR, where one is able to track nearly every individual that leaves a digital trace behind, singular UGC data is reduced to just one specific source (e.g. Facebook, Flickr, Instagram, Tripadvisor) of geosocial data. This results in numerous other fruitful sources being excluded in a research simply for methodological reasons (Johnson et al, 2011). An on the first hand simple looking solution to this could be to use multiple geosocial sources for an improved real-world representation. However, it is the challenge here that comparing apples to oranges is not possible. The collected data must include common methodological values to make analysis possible (Marti, Garcia-Mayor & Serrano-Estrada, 2020).

Concluding, it can be stated that UGC data can in fact be used to investigate tourist behaviour. However, researchers require to be cautious with bringing results forward concerning UGC data analysis, as one is always working with a simplified model of reality. A solution to this issue could be including expert opinions (e.g.

employees of local DMO's, municipality employees) to calibrate the results of the model to the actual world as opposed by Chua et al. (2016).

3.4 Conceptual model

The general challenge of researchers that can be derived from the literature is the analysis of where users of a city are, and when users of a city are at a place. Various research projects looked at this phenomenon in multiple ways. Not only does the vision between time and space vary, but also does the scale of the research. For example, the research by Giglio et al. (2019) and Van der Zee & Bertocchi (2018) explore the space where users are on a point scale. A point scale can be defined as an exact location, where every data point resembles a piece of data. Brandt et al. (2017) perform a mainly temporal analysis on both full city scale and point scale with the help of a density map.

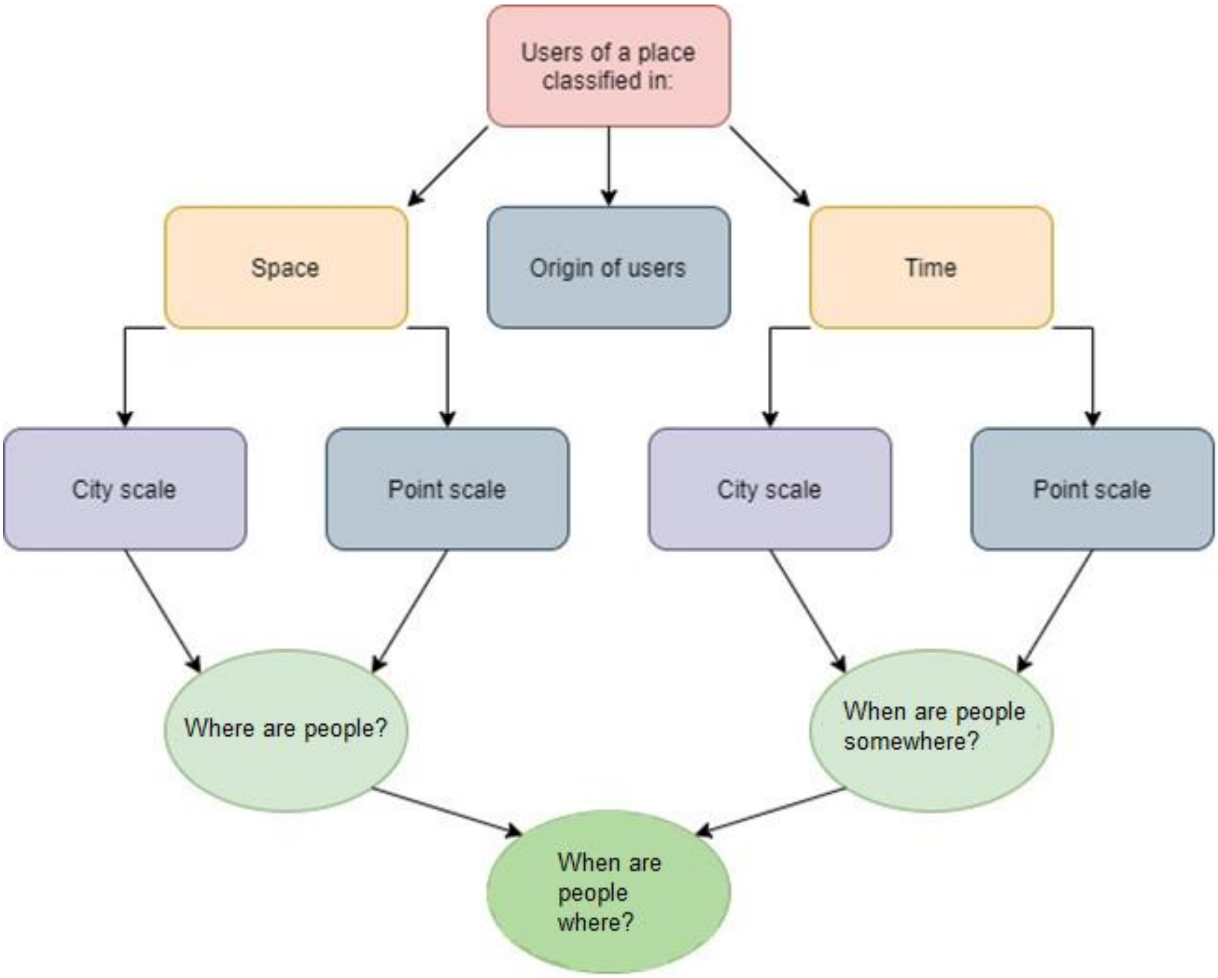


Figure 3.1: Conceptual model of this research project

There are also examples of research papers that describe both space and time on the city scale and on the point scale (Miah et al., 2017, Van der Drift, 2015). Another interesting element that already has been used by Van der Zee & Bertocchi (2018) is to also include the origins of users. These experiences altogether eventually form the conceptual model that is visualised in figure 3.1. All knowledge from the conceptual model combined is expected to give an answer to the proposed questions of this research project.

4. Methodology

The sub-questions that have been formulated will be operationalised in this chapter. These have been set out here again for readability purposes:

SQ1: What methods can be used to obtain geosocial data and what geosocial data is available and usable?

SQ2: How can geosocial data be analysed to visualise spatial and temporal patterns of people?

SQ3: To what extent can geosocial data be used to display an accurate model compared to reality?

To answer SQ1, multiple types of geosocial data will be investigated and tested on usability. Besides that, the data collection process will be briefly explained and eventually further elaborated in chapter 5 on the data preparation. For SQ2, the different types of analysis that have been performed will be explained for both the temporal and spatial part of this research. In order to answer SQ3 sufficiently, an extensive part of the results and discussion section will be dedicated to the ifs and buts that are paired with the performed research steps. Also, an expert opinion will be consulted on the outcomes of the study. Answering all three sub-questions will provide a complete answer to the main question:

"How can different sources of user generated geosocial data be used to visualise spatial and temporal patterns of individuals in Rotterdam & Veere?"

As aforementioned, this quantitative research is a proof-of-concept of using geosocial data. Rotterdam and Veere have been chosen as the study area of this research project. First, the choice behind these places will be further elaborated. Second, the research steps to provide answers to the sub-questions have been schematically drawn and are further explained after that. Third, there will be explained on how the results of this research project have been validated.

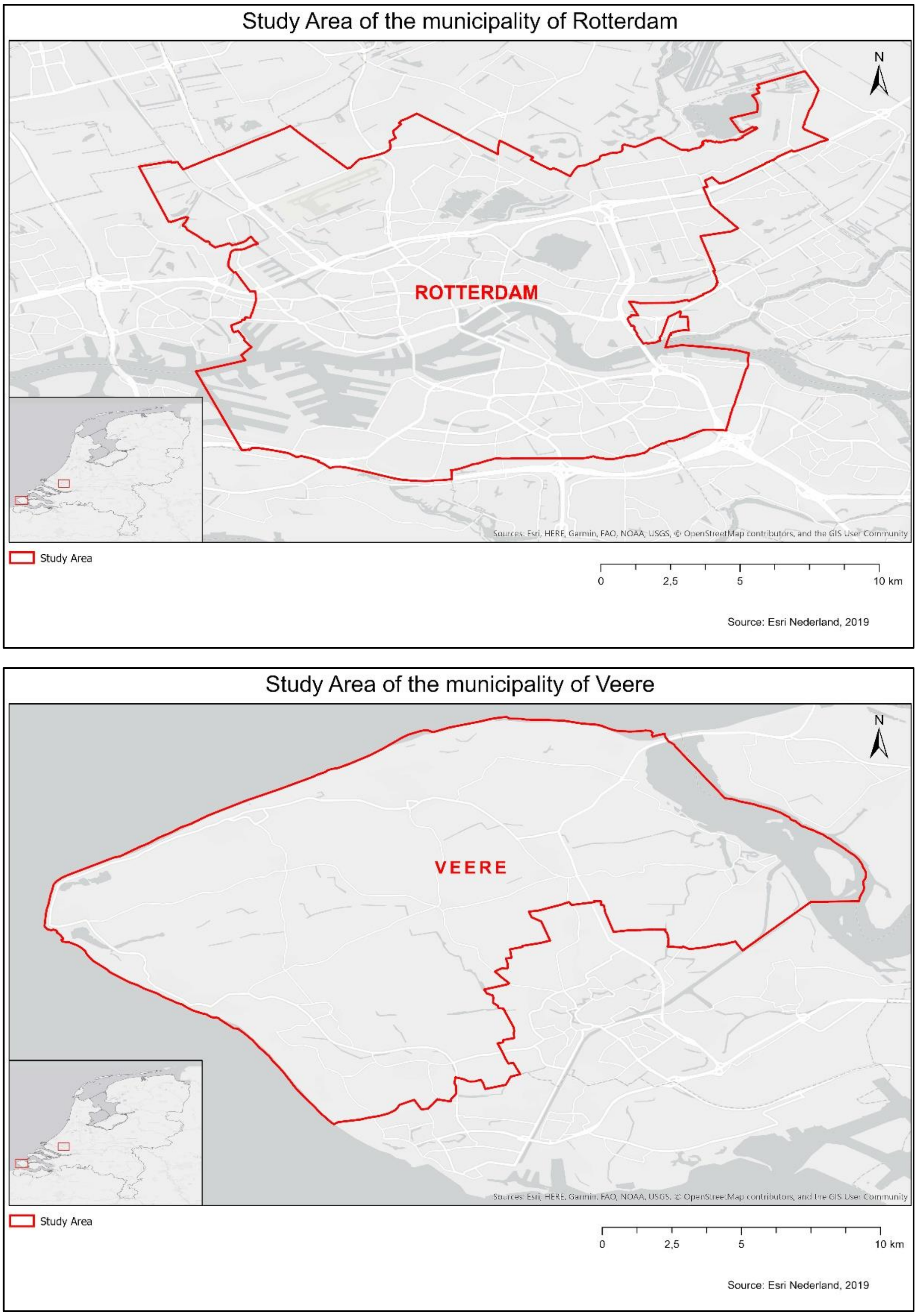

Study Area of the municipality of Rotterdam
N
ROTTERDAM
Sources: Esri, HERE, Garmin, FAO, NOAA, USGS, © OpenStreetMap contributors, and the GIS User Community
Study Area
0
2,5
5
10 km
Source: Esri Nederland, 2019
Study Area of the municipality of Veere
N
VEERE
Sources: Esri, HERE, Garmin, FAO, NOAA, USGS, © OpenStreetMap contributors, and the GIS User Community
Study Area
0
2,5
5
10 km
Source: Esri Nederland, 2019

Figure 4.1: Study Areas of this research project. (Esri Nederland, 2019)

4.1 The study area

The research project will be a proof-of-concept on using geosocial data to visualise spatial and temporal patterns. Eventually, the goal is to see whether analysis with geosocial data is possible for more places. The study area for this proof-of-concept research will consist of the city part of the municipality of Rotterdam (province of Zuid-Holland) and the municipality of Veere (province of Zeeland) (Figure 4.1). The reasoning for the choice of these municipalities is the difference in urbanity, where Rotterdam is the urban area and Veere the rural area. Rotterdam has about 650.000 inhabitants, is known as an international harbour city and is the second largest city of The Netherlands (CBS, 2020; Nientied, 2020). Veere has just 23.000 inhabitants. With villages like Zoutelande, Westkapelle and Domburg, Veere is mainly known for its coastal tourism (CBS, 2020; Kruizinga, 2016).

4.2 Research steps

Figure 4.2 shows the schematic overview of the research steps that will be followed during this project. Along the research steps the three sub-questions (SQ's) presented in the research design chapter will be answered.

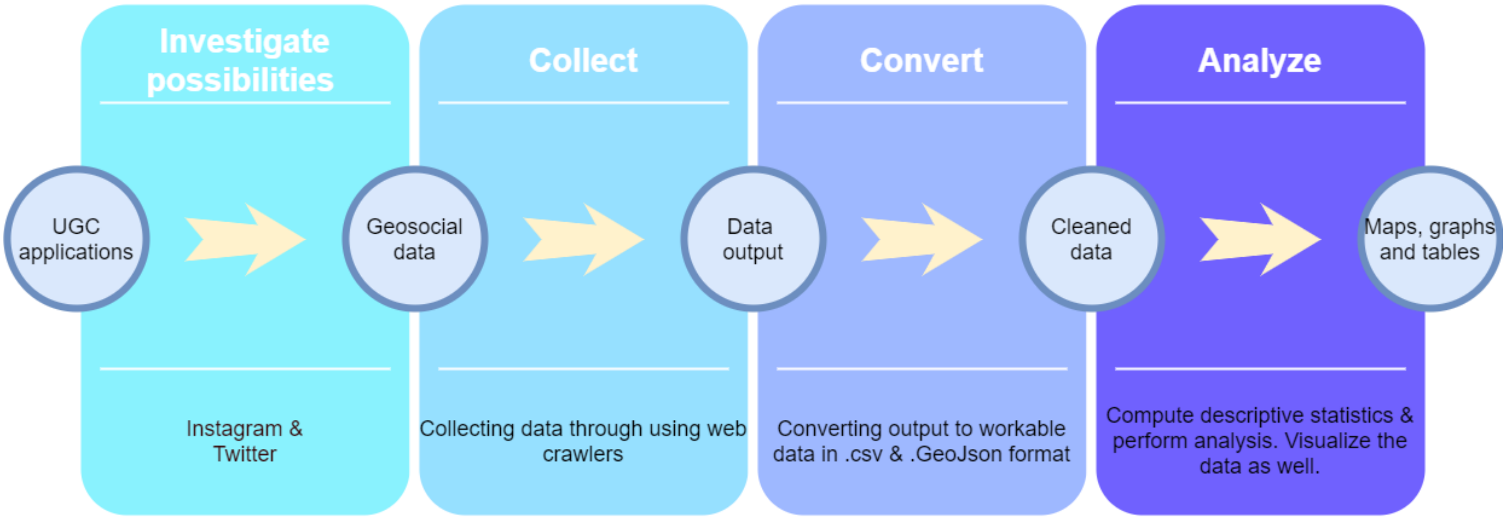

Investigate possibilities
Collect
Convert
Analyze
UGC applications
Geosocial data
Data output
Cleaned data
Maps, graphs and tables
Instagram & Twitter
Collecting data through using web crawlers
Converting output to workable data in .csv & .GeoJson format
Compute descriptive statistics & perform analysis. Visualize the data as well.

Figure 4.2 Schematic view of the research steps in this project.

4.2.1 Investigating data

There are numerous applications containing geosocial data. Previous research shows that UGC data has the potential to explain spatial and temporal patterns of individuals and could therefore contribute to answering the main question of this research project (Brandt et al., 2017; Chua et al., 2016; Ganzaroli et al., 2017; Giglio et al., 2019; Miah et al., 2017; Van der Drift, 2015; Van der Zee & Bertocchi, 2018; Vu et al., 2018; Wood et al., 2013; Wu et al., 2018). As there is an interest in both the spatial and the temporal aspect, it is important for the data to contain both a location and a timestamp. From the literature review it can be derived that many applications have already been used in similar research projects. The aim of SQ1 is therefore partially to find available and suitable datasets for this research project.

Flickr is an application that is used quite often in research with geosocial data (Giglio et al., 2019;, Miah et al., 2017; Van der Drift et al., 2015; Wood et al., 2013). Flickr data is relatively easy to obtain, as they have an API. Also, the actual data itself is useful as it contains both detailed location and timestamp information. A downside to Flickr is the relatively small user base of 100 million registered users in 2018 (The Verge, 2018) compared to Instagram with 1 billion users, a tenfold of the amount of Flickr users (Statista, 2020). Both applications serve in general the same purpose, as they are used for sharing pictures online. Instagram however has not been used often in studies (Tenkanen, Di Minin, Heikinheimo, Hausmann, Herbst, Kajala & Toivonen, 2017). Reasons for this could be that they do not have an official functional API as of now, which makes it more challenging to obtain data from their website. A reason for Instagram not being used often in studies might be because of the geolocation of posts that cannot be exactly traced back in the data. Therefore, the geographic component in the data cannot give a very detailed image of where posts have been posted.

The application Instagram lets users select a location of where the picture that they have posted was taken. This can occur on village or city scale, which is most common, but one can also select a location on neighbourhood, street, or even POI scale. While numerous Instagram posts are blocked from the public due to personal privacy settings, the time element is able to stand out in their data due to the high number of active users on the platform, as opposed to other geosocial data sources.

Also, users post pictures of many things they are undertaking, as opposed to Flickr users who generally upload pictures of specific landscapes or buildings.

Another type of social media that is frequently used in research with geosocial data is Twitter (Brandt et al., 2017; Chua et al., 2016; Tenkanen et al., 2017). In contrast to the Flickr posts, Tweets are, like Instagram posts, also posted as life updates on what people are undertaking. Currently, Twitter has around 330 million monthly active users (Statista, 2020b). Recent research that has been conducted by Tenkanen et al. (2017) is one of the few research projects that used multiple types of geosocial data (Instagram, Twitter, and Flickr). One of the conclusions was that the results from the Instagram data was more representative than the results from both Flickr and Twitter data. Therefore, also considering the fact that Instagram is not often used as a data source, it is interesting to use Instagram in this research project to get to possibly new and representative insights.

As this research project will make use of two data sources to answer the main question with extra validity, another type of application than a regular social media application will also be used. While Instagram can be greatly used for the time component of this research project, it still lacks the spatial component. This is because the data can only be collected by geotags in general, and not on point level. TripAdvisor has been frequently used for various research projects (Ganzaroli et al., 2017; Van der Zee & Bertocchi, 2018). TripAdvisor contains several types of UGC data and can therefore serve multiple functions. TripAdvisor is a public web application where people can upload reviews of Restaurants, Cafes, or POI's that they paid a visit to. These reviews can be used for qualitative research in the form of a content analysis but have also been proven to be effective for quantitative research, where the number of reviews of a place generally gives an indication of the number of visitors in a restaurant or café.

Another level of TripAdvisor data which can be used is the user profile of each reviewer. However, this information needs to be handled with care as it includes personal data, but as long as the data is anonymized, there are no ethical issues (Fiesler et al., 2020). Data on the user profile of reviewees can be used to identify what the country of origin is per individual that is visiting an eatery location (Van der Zee &

Bertocchi, 2018). Both Instagram and TripAdvisor will be used in this research project, which forms a partial answer already to SQ1.

The applications Instagram and TripAdvisor both contain UGC data and are also a form of geosocial data, as they contain a geographic component. A short summary of important aspects of both applications is given in table 4.1. The Instagram data is however lacking an exact geographical location, which does not imply that the two data sources could not be compared. A similarity between the two data sources is the presence of a timestamp in the collected data. Where the timestamp of Instagram posts is exactly on the minute, creating opportunities for precise temporal analysis, reviews from TripAdvisor are precise on the day. In addition to that, both geosocial data sources contain a digital footprint of different social media user groups. The data can therefore add on to each other and reveal patterns on these users.

Table 4.1: Availability, users, and relevance of Instagram and TripAdvisor as geosocial data sources (Statista, 2020)

Source	Availability	Users	Relevance
Instagram	Scraping possibilities per location tag (i.e., municipality, city, neighbourhood, POI). Only posts available from public accounts.	Over 1 billion users, mainly the younger population of people. Mainly Western users.	Large group of users, good representation of the younger population. Very detailed temporal results.
TripAdvisor	Possibility to scrape all POI's in a place, including number of reviews and review profiles of users.	Approximately 450 million users, visiting hotels, restaurants, cafes, highlights, or activities. Mainly Western users.	Gives an insight in POI visits on the specific location of the POI. Therefore, useful for spatial analysis. Also, able to provide information about the origin of users.

Combining the data from both Instagram and TripAdvisor with the conceptual model of chapter 3.4, results in an operationalized conceptual model that can be found in figure 4.3. The main challenge is to find out where and when users are located somewhere. According to the operationalized conceptual model, there might be an answer provided to that question when there is coverage of UGC data in that area. This is described by the arrows above in figure 4.3. The UGC data of Instagram will be mainly used to analyse spatial and temporal patterns in detail on city level, where UGC data of Tripadvisor will be mainly used for the specific analysis on location and time on point scale and also provide information on the origin of users.

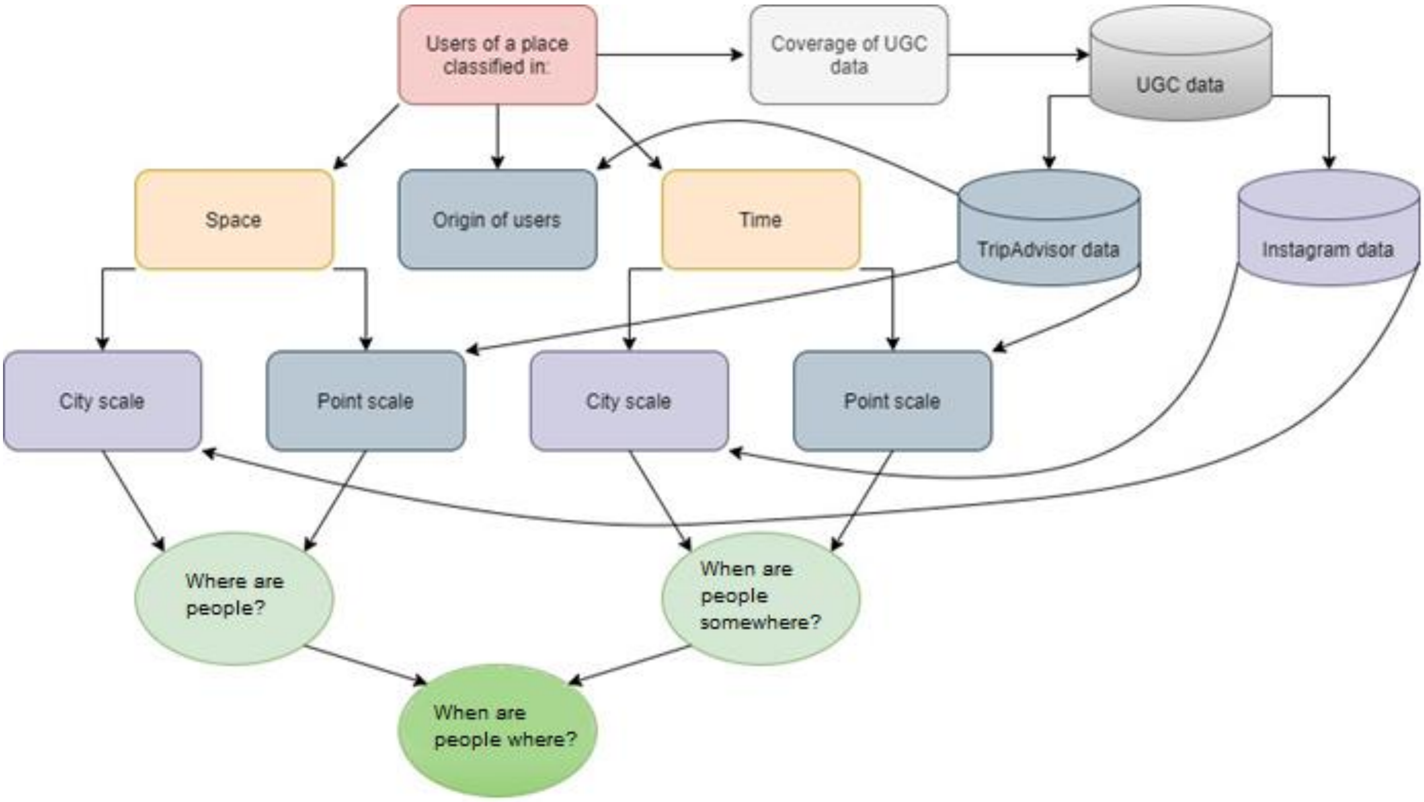


Figure 4.3: Operationalized conceptual model

4.2.2 Collecting & Converting data

After finding the correct data sources to harvest the geosocial UGC data from, the second stage, which is the actual web scraping commences. As there is no functional API available to retrieve Instagram- and TripAdvisor data, the code to retrieve the data needs to be written in Python. There is some web scraping documentation already available on GitHub, however several adjustments need to be made to obtain the correct workable data. After the data collection, the data must be carefully inspected and cleaned. More elaborated information on the web scraping process will be given in the operationalization chapter.

After the collection process, the data also needs to be converted to a workable .csv format to eventually perform analysis to. The Instagram data will be collected in Jsonl format, which stands for JavaScript Object Notation – Lines. It is a standard text-based format for representing structured data based on the JavaScript object syntax. The data conversion will be done with the help of a Pandas data frame in Python, that can eventually be converted to a .csv format. The TripAdvisor data is being collected in Hyper Text Markup Language (HTML). This data will be parsed and whatever data is required for this research project will be extracted in Json format. Chapter 5 on operationalization will also delve deeper into the details of the conversion process of both Instagram and TripAdvisor.

4.2.3 Analysing data

In the final step, the retrieved and cleaned geosocial data can be used to obtain descriptive statistics and perform multiple types of spatial and temporal analyses to get to an answer of SQ2. Eventually, the goal of this research project is to find out how both spatial and temporal patterns can be visualised with the help of geosocial data. For this, multiple approaches can be used, which will be explained in this section.

For the temporal results of both Instagram and TripAdvisor, it is already valuable to display the daily number of posts and reviews throughout the year. Therefore, the data is converted into .csv format, after which several pivot tables have been created. The data is visualised over the study year (01-09-2019 until 31-08-2020). For a comparison between an urban and a rural area, the relative daily Instagram posts of both Rotterdam and Veere have also been plotted in a graph. Further temporal analysis has been performed with the hours in which the Instagram users posted. The TripAdvisor data is not very suitable for detailed temporal analysis, as the body of data is smaller, giving a less stable image, and reviews are usually posted a while after an eatery location has been visited by a user.

However, due to the locational factor present in the TripAdvisor data, patterns can be visualised in maps by using heatmaps. However, as the methodology behind heatmaps is not always as clear and transparent, also some more advanced statistics will be applied to the data. For an insight in spatial autocorrelation the Moran's I has been calculated and visualised (Moran, 1950). To add on to that, a hot spot analysis is performed as well (Getis & Ord, 1992). With this type of analysis, the goal is to find significant hot spots in the study area. This has been done with using the number of reviews of each eatery location as a population field, as the number of reviews of the past year is likely to be in ratio with the number of customers of the past year. In previous research projects, aforementioned methods in comparable studies resulted in valid outcomes (Brandt et al., 2017; Chua et al., 2016; Giglio et al., 2019; Van der Zee, Bertocchi & Vanneste, 2020).

4.3 Validation of the results

Data analysis with the use of geosocial data will unfortunately always be a downgraded model of reality (Heaton, 2019). Therefore, to answer SQ3 on the accuracy of the analysis and the comparison to reality, there has been looked at other datasets to calibrate the data, such as weather datasets and data on public holidays. Also, the use of two datasets makes it possible to calibrate the two datasets with each other. After the multiple analyses have been performed, the results have also been compared to results of other research projects. Lastly, an expert opinion has been consulted on the results of study, to either confirm or doubt the analysis outcomes.

5. Data preparation

This chapter will elaborate on the graphic explanation of figure 4.3 and explain the actual web scraping progress for both the Instagram and TripAdvisor data. First, the Instagram data retrieval and preparation will be explained. Afterwards, the same will be done for the TripAdvisor data.

5.1 Instagram data

The UGC application Instagram is a platform where individuals can share pictures and videos with each other. An important aspect of these posts which is relevant for this research project, is the ability to include a geotag with the post. This can be any geographically definable location like a country, city, village, neighbourhood, street, or even a single building. In the explore section of Instagram, one can look for posts in a specific location. The Uniform Resource Locator (URL) includes a specific identifier for that location (e.g., Rotterdam = 213226541, Veere = 250640993). As Veere is a bigger municipality with multiple villages in it, the choice has been made to web scrape all villages in the municipality of Veere instead of the municipality of Veere as a whole. These villages are: Aagtekerke, Breezand, Domburg, Grijpskerke, Koudekerke, Meliskerke, Oostkapelle, Veere, Vrouwenpolder, Westkapelle, and Zoutelande. For all locations, the identifier will eventually be used to retrieve all Instagram posts in every area.

5.1.1 Web scraping

The complete Instagram scraper that is used for this research project can be found in Appendix B. Some code snippets will be highlighted in the following paragraphs as an explanation of what the main parts of the code are destined to do.

```
BASE_URL = "https://www.instagram.com/explore/locations/{location_id}/?__a=1&max_id={max_id}"
```

Code snippet 1: Code that sets the base URL

Before the start of the web scraping, several packages and tools need to be installed. Examples of used packages/tools are ArgParse, Backoff, and Requests. After these have been imported, code snippet 1 sets the base URL for the actual website where the scraping will happen from.

```
@backoff.on_exception(wait_gen=backoff.fibo, max_tries=MAX_TRIES,
                      exception=(requests.exceptions.HTTPError,
requests.exceptions.ConnectionError))
def pull_json(location_name, end_cursor):
    URL      =      BASE_URL.format(location_id=METRO[location_name],
max_id=end_cursor)
    logging.debug(URL)
    r = requests.get(url=URL)
    if r.status_code == 200:
        data                                                        =
r.json()['graphql']['location']['edge_location_to_media']
        return data
    else:
        raise requests.exceptions.HTTPError
```

Code snippet 2: Code that feeds the base URL from code snippet 1

```
def save_jsonl(data, dst='./'):
    path = Path(dst)
    assert isinstance(data, list)
    if path.exists():
        f = codecs.open(dst, "ab", 'utf-32')
    else:
        path.parent.mkdir(parents=True, exist_ok=True)
        f = codecs.open(dst, 'wb', 'utf-32')
    f.writelines("%s\n" % json.dumps(s) for s in data)
```

Code snippet 3: Code that writes the collected data to a .jsonl file

```
min_date  = "2019/09/01"
max_date = "2020/08/31"
location  = "Rotterdam"
```

Code snippet 4: Manual input for the scraper

Code snippets 2 and 3 are used to pull the Instagram data from the web server to a .jsonl file. The eventual web scraping occurred at a rate of approximately 3.5 posts per second. This is a rather slow collection process, and the Instagram servers are in no way harmed by this operation, which is important to consider (Fiesling et al., 2020). A common problem that occurred was a soft block of the program after scraping around 4000-5000 posts. This happened relatively quick as a year of data in Rotterdam results in over 400.000 posts, that would be obtained with inserting code snippet 4 in the scraper. Code snippet 4 also shows the novelty of the scraper, as instead of only scraping the latest number of posts, or all posts, the choice can be made for a minimum data and a maximum data where between data harvesting should take place.

A way to circumvent the issue of Instagram blocking the Internet Protocol (IP) address is to make use of Google Collaborations, which gives the ability to restore the coding environment and therefore makes it able to continuously collect data. Instead of using a local workspace, a clone is made from a local GitHub repository where the scraper is located, as can be seen in the first part of code snippet 5.

```
from pathlib import Path
from google.colab import drive
drive.mount('/content/drive', force_remount=True)
directory = Path('/content/drive/My Drive/Instascrape/')
mypath = directory/'Method_1'
mypath.mkdir(parents=True, exist_ok=True)
! cd "$mypath" && git clone https://github.com/BramR123/insta-graphql-scraper


! cd "$mypath"/insta-graphql-scraper && python scraper.py --restore-cursor --dir "$mypath" --max $max_date --min $min_date --location $location
```

Code snippet 5: Collecting data through Google Collaborations & restoring the cursor

Another issue that comes to the attention is that one would need to restart with the same date input over and over again when the scraper would get blocked or scrape separately by each day and create 365 data files over the year. To tackle this issue, a text file containing the latest end cursor, a code that refers to the latest post that has been scraped in a run, is added together with the scraping result. Therefore, the scraper keeps track of the last post that has been collected. With the use of the second part of code snippet 5, the previous end cursor location is restored, and the scraping continuous where it stopped.

5.1.2 Data conversion

After the collection of the data, the conversion progress takes place. For Rotterdam, a .jsonl file with a size of 5.5 gigabytes is created, and needs to be converted to a .csv file. First, as can be seen in code snippet 6, the .jsonl file needs to be loaded in and decoded.

```
path = r"C:\Users\bramr\Documents\Workplace\Instascrape\Method_1\Rotterdam_2019-09-01_2019-08-31.jsonl"

datas = []
with open(path, 'r', encoding='utf-32') as f:
    for line in f:
        datas.append(line)

jsons = []
for lines in datas:
    decoded_data=codecs.decode(lines.encode(), 'utf-8-sig')
    data = json.loads(decoded_data)
    jsons.append(data['node'])
```

Code snippet 6: Naming the path and decoding the .jsonl files

```
df = pd.DataFrame(jsons)

likes = []
for i in range(len(df['edge_liked_by'])):
    likes.append(df['edge_liked_by'][i]['count'])
df = df.assign(Number_of_likes = likes)

df['datetime'] = pd.to_datetime(df['taken_at_timestamp'], unit="s")

df_for_export = df[['ID', 'Number_of_likes', 'datetime']]
df_for_export.to_csv('Insta_Rotterdam_1_9.csv')
```

Code snippet 7: Creating the data frame and exporting a .csv file

A pandas dataframe is created to eventually load in all the columns of the data that will be used. As can be seen in code snippet 7, the columns and belonging data that is added are the ‘likes’ and the ‘datetime’. The column with likes contains the number of likes on every post, while the datetime column contains a timestamp that will be converted to a date-time format which is accurate to the second. After that, the data frame is exported to a .csv file, making the Instagram dataset ready for analysis.

5.2 TripAdvisor data

TripAdvisor.com is a website where users are able to leave reviews on places that they have visited. These places could be restaurants, cafeteria’s, hotels, and even attractions/landmarks. The reviews produced by visitors can be used by new potential visitors when deciding on where to eat, or what places to visit. The result of this is a big database filled with UGC on where people have been, which therefore makes it interesting to look at for this research project. As this research project is a proof-of-concept, only restaurant data will be considered for now. The review data originates from restaurants that are in the municipalities Rotterdam and Veere, as the Instagram data has also been collected in these municipalities.

5.2.1 Web scraping

For collecting the TripAdvisor data, it is important to first understand the websites' infrastructure. All sorts of information are stored on the TripAdvisor website. When looking at restaurants in for instance Rotterdam, the reviews are not visible immediately as they are nested in the page of every restaurant itself. Therefore, the first task is to collect all the restaurant URLs in both the municipalities. This is displayed in code snippet 8. A challenge here is that TripAdvisor sorts on restaurants in cities/villages instead of municipalities, which is similar to the Instagram geotags. For Rotterdam, this is not an issue as all restaurants are located in both the city and municipality of Rotterdam. For Veere however, the data needs to be collected individually for each village. When the pages are collected, the links to the restaurants can be collected, which is shown in code snippet 9.

```
page = "https://www.tripadvisor.nl/Restaurants-g652353-Vrouwenpolder_Zeeland_Province.html"


def getAllResults(firstPage, city):
        firstPage = firstPage + "#EATERY_LIST_CONTENTS"
        allLinks = [firstPage]
        p = requests.get(firstPage)
        s = bs(p.content, 'html.parser')
        idx = firstPage.index('-%s' % city)
        try:
            div = s.find('div', {'class': 'pageNumbers'})
            children = div.findChildren(recursive=False)
            numberofP = len(children)
        except AttributeError:
            numberofP = 1


        for i in range(1, numberofP):
            print("Creating URL for page "+str(i))
            template = '-oa%s0' % (i*3)
            finalLink = firstPage[:idx] + template + firstPage[idx:]
            allLinks.append(finalLink)
            print(finalLink)
        return allLinks
allthePages = getAllResults(page, "Vrouwenpolder")
```

Code snippet 8: Collecting the pages where the restaurants are listed

```
def getURLs(searchResultPages):

    searchResults = {}

    for idx, i in enumerate(searchResultPages):

        r = requests.get(i)

        soup = bs(r.content, 'html.parser')

        a = soup.find(id='EATERY_SEARCH_RESULTS')

        list_items = a.find_all("div", { "data-test" :
```

…

```
       return searchResults


alltheURLs = getURLs(allthePages)


with open('URLs_Vrouwenpolder_final.json', 'w') as fp:

    json.dump(alltheURLs, fp)
```

Code snippet 9: Collecting the URL of every restaurant

At this moment, a .json-file is created with links to each restaurant. By now, it is time to collect the actual data of the reviews for each restaurant. While the collection of the URLs has been done with the python requests package, the reviews are collected with the help of Selenium. Selenium is a web driver that automates browsers. After identifying several locations and buttons on a website (e.g. the ‘next page’ button), Selenium can crawl through websites once the Document Object Model (DOM) has loaded (Selenium.dev, 2021). The following paragraphs will highlight some of the important code snippets in the remainder of the data collection. The actual code for the web scraper can be found in Appendix C.

The main function of code snippet 10 is to make sure that every review page is scrapable. The TripAdvisor base URL is combined with a URL from the list of restaurants, and ‘all_lang’ is added on to the link to ensure that reviews in all languages will be shown, instead of only the reviews in the native language of the browser. Besides that, the ‘Do you agree with the cookie policy?’ needs to be clicked most of the time the page refreshes. A delay is also added in between the different steps to make sure all details are loaded before the data is collected.

```
base_url = "https://www.tripadvisor.com"

all_lang = "?filterLang=ALL"

url = base_url + restaurant_urls[key]["URL"] + all_lang

        driver.get(url)

        time.sleep(0.5)

        driver.refresh()

        time.sleep(2.5)

        try:

            driver.find_element_by_xpath(".//button[@class='evidon-banner-acceptbutton']").click()

        except:

            pass
```

Code snippet 10: Configuring the base URLs

After that, some try/except loops have been written to collect data from each restaurant. The try/except loop is required as some data fields do not exist for every review (e.g. not every restaurant has reviews, not every review has a ‘…more’ button). The location of the restaurant is obtained by finding the coordinates from the Google Maps insert on the main page of the restaurant. Besides the location, the total number of reviews is also obtained. After that, some modifications are done in such way that the reviews can be collected. When a review contains more than a certain amount of words, the remainder of the review is hidden behind a '…more' button. If there are any, they are expanded so that the whole review can be obtained.

For every review, the user information is also being collected. Some users have specified their country and/or place of origin in their profiles, which could get an insight in the home country of the restaurant visitors. Therefore, the user identification number (UID) of each reviewer is collected to use later. This is shown in code snippet 11.

```
try:
    uid_src                                                          =
container[j].find_element_by_xpath(".//div[@class='memberOverlayLink
clickable']").get_attribute("id")
    matches = re.match(r"^UID_([0-9a-fA-F]+)-SRC_(\d+)", uid_src)
        review["uid"] = matches.group(1)
        review["src"] = matches.group(2)
except:
    pass
```

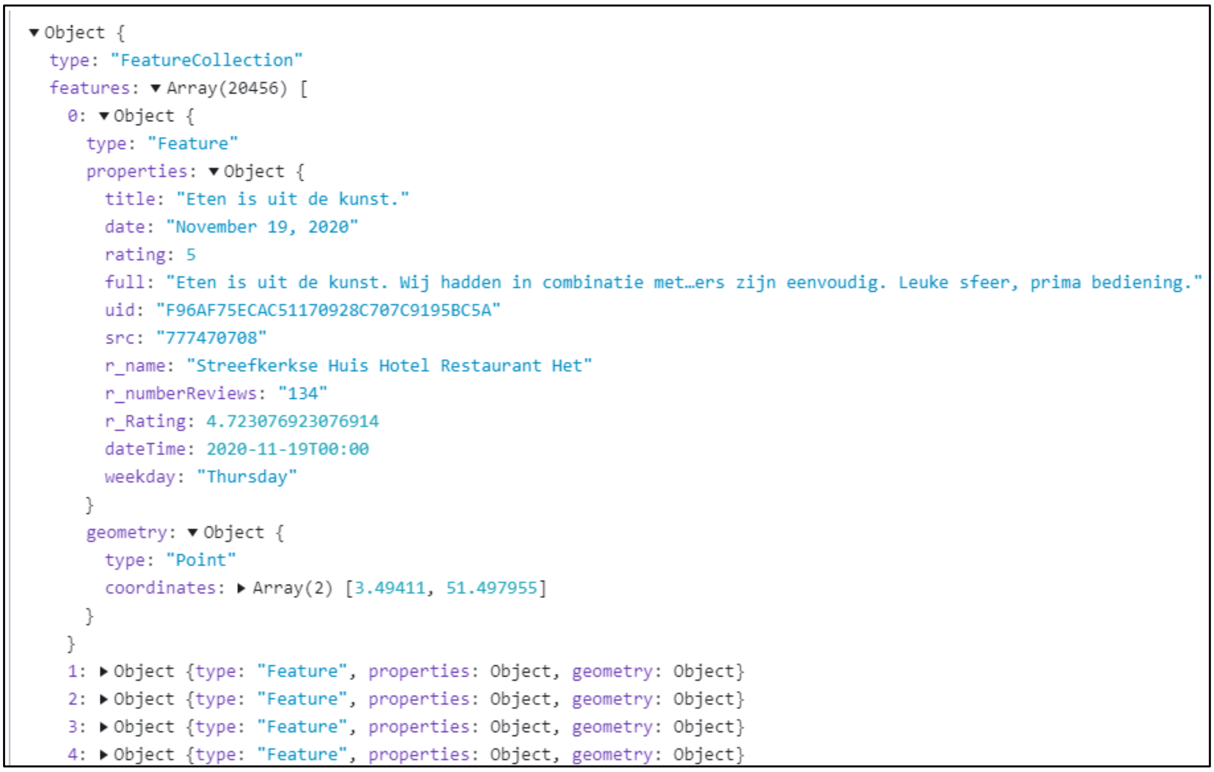
Code snippet 11: Collecting the User ID's

After completion, the data is stored into a .json-file. For a quick data overview and data conversion, the .json-file with all the reviews has been loaded into an Observable notebook. This is a JavaScript notebook, and therefore suitable to work with data in .json-format. However, as the files with reviews are relatively large, the data first needs to be uploaded to Dropbox and after that being shared with an open link. The Observable notebook gets the data from this link, after which array mapping and reducing can begin (Code snippet 12). Array mapping is done to apply a transformation on every item in an array, while array reducing summarizes multiple items into one value.

After the changes, the data is flattened in such way that every unique review has information stored. Every review consists of, besides the restaurant name, also the review date, the UID and the location of the restaurant. After that, the data is converted to .geojson-format to be able to successfully map the data in a GIS. An insight in one individual data point is shown in figure 5.1.

```
Final = Object.keys(data3)
  .map(i => data3[i])
  .map(item =>
    Object.keys(item.reviews).map(k => ({
      ...item.reviews[k],
      rating: Number(item.reviews[k].rating[0]),
      lon: item.lon,
      lat: item.lat,
      r_name: item.Name,
      r_numberReviews: item["Number of reviews"],
      r_Rating: Object.keys(item.reviews)
        .map(review => Number(item.reviews[review].rating) / 10)
        .reduce(average, 0),
      dateTime: d3.timeParse("%B %d, %Y")(item.reviews[k].date),
      weekday:
        days[
          getWeekDay(
            d3
              .timeParse("%B %d, %Y")(item.reviews[k].date)
              .getDay()
          )
        ]
    }))) .flat()
array_toGeoJSON(final)
```

Code snippet 12: Array mapping & reducing



```
▼Object {
  type: "FeatureCollection"
  features: ▼Array(20456) [
    0: ▼Object {
      type: "Feature"
      properties: ▼Object {
        title: "Eten is uit de kunst."
        date: "November 19, 2020"
        rating: 5
        full: "Eten is uit de kunst. Wij hadden in combinatie met…ers zijn eenvoudig. Leuke sfeer, prima bediening."
        uid: "F96AF75ECAC51170928C707C9195BC5A"
        src: "777470708"
        r_name: "Streefkerkse Huis Hotel Restaurant Het"
        r_numberReviews: "134"
        r_Rating: 4.723076923076914
        dateTime: 2020-11-19T00:00
        weekday: "Thursday"
      }
      geometry: ▼Object {
        type: "Point"
        coordinates: ▶Array(2) [3.49411, 51.497955]
      }
    }
    1: ▶Object {type: "Feature", properties: Object, geometry: Object}
    2: ▶Object {type: "Feature", properties: Object, geometry: Object}
    3: ▶Object {type: "Feature", properties: Object, geometry: Object}
    4: ▶Object {type: "Feature", properties: Object, geometry: Object}
```

Figure 5.1: TripAdvisor data structure after conversion

5.2.2 Data conversion

After creating a readable .geojson-file from the data, it is possible to load the data into a GIS. First, the data has been loaded into QGIS, where it was converted to a shapefile. Due to the size of the .geojson-file this could not be done in ArcGIS Pro. As TripAdvisor also looks for other restaurants in the area when there is not a lot of supply in the city/village that it is looking for, the data has been collected from multiple villages throughout Veere. After that, the data was merged into one dataset. Due to this method, all restaurants of the area have been scraped, but among them were a high number of duplicates. Filtering based on the full review and the review date made sure that all duplicates were removed from the dataset.

Due to this method of data collection, there were also some restaurants scraped outside of the study area. The shapefile was loaded into ArcGIS Pro, whereafter a selection based on location was performed to remove any restaurants outside the research area. After this process, the data has been carefully collected, converted and cleaned, and is therefore ready for analysis.

6. Results

This chapter will elaborate on the results and different analytical methods that could be applied to the collected and converted data from both Instagram and TripAdvisor. There are numerous types of analysis possible with the data. As the main focus of this research project concerns the question on how different sources of UGC data can help to visualise spatial and temporal patterns of individuals, the methodology and interpretation of the results are more important than the analysis itself. Therefore, this chapter will contain several forms of analysis, mainly aimed at understanding and debating the ifs, ands and buts behind the data.

For the consistency of data visualisations, there has been chosen for a visualisation period of a year. The period starts the 1st of September 2019 and ends the 31st of August 2020. The choice of this timeframe has been made so that the data is (1) as recent as it can be, and (2) might show differences during the different phases of the first lockdown concerning the COVID-19 pandemic.

The results are structured according to the conceptual model from chapter 4. The aim is to see if it is possible to answer the two questions inside the red circle visible in figure 6.0. First, a brief introduction will be given for both datasets that have been used. After that, a paragraph will shed light on the temporal aspect of this study. In the subsequent paragraph, the spatial part of this study will be further elaborated. Finally, the origin of users will be investigated, which is highlighted with the blue circle in figure 6.0.

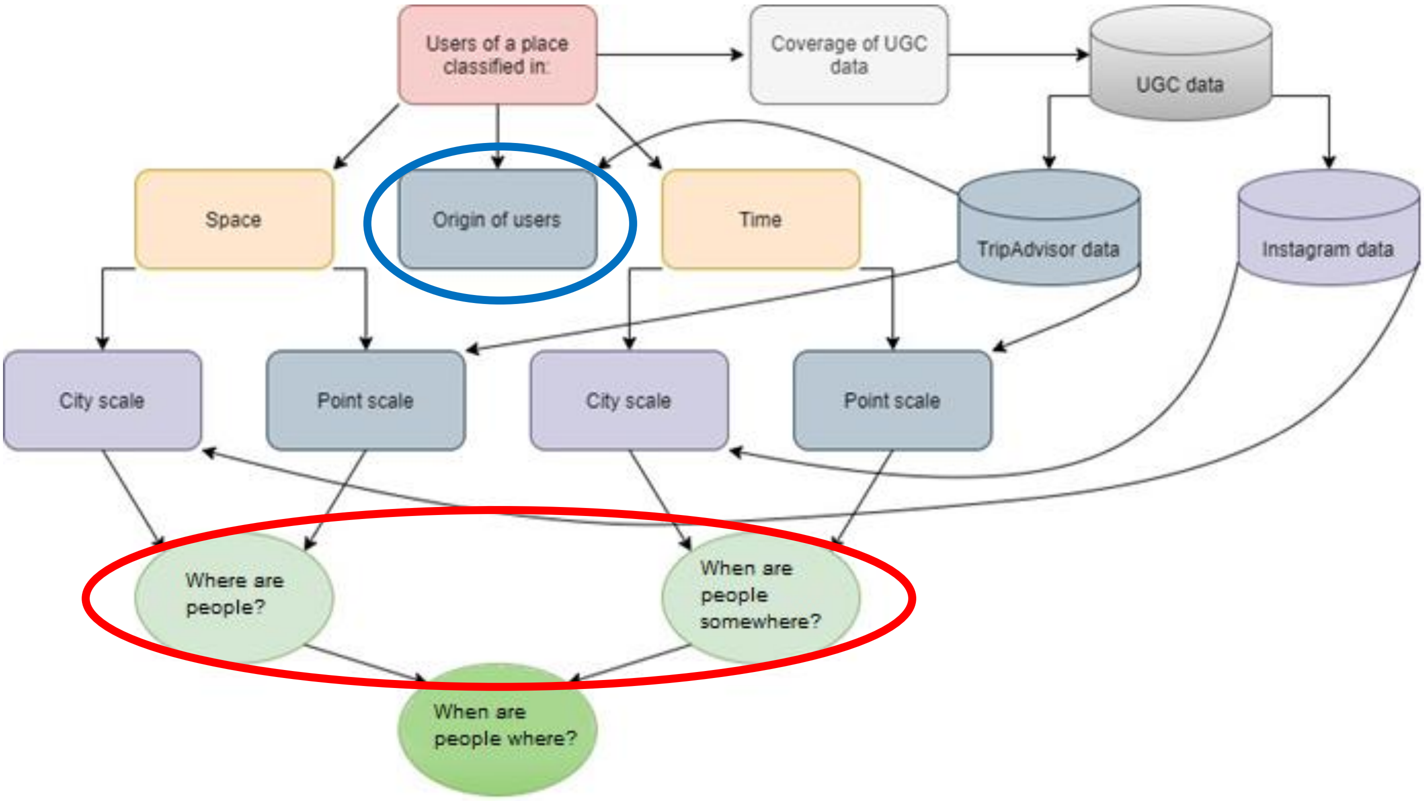


Figure 6.1: Conceptual model applied to the results

6.1 Data exploration

Both the TripAdvisor dataset and the Instagram dataset comprise of rich data. Through the data collection, much information was obtained per post or review. For Instagram, 445.928 posts were collected for Rotterdam and 37.518 posts were collected from Veere. Compared to the population of both study areas would this be 0.69 posts per capita for Rotterdam and 1.63 posts per capita for Veere. Every post contains information of the post itself, the UID, the exact time and date of the post, the amount of likes on a post, and the caption that has been written below the post. These characteristics enable multiple types of detailed temporal analyses.

Where the Instagram data mainly comprises of temporal data, the TripAdvisor data is mainly able to shed more light on spatial data. A downside of the data is that it is not that extensive in a quantitative way as opposed to the Instagram data. Where the Instagram data reached the tens of thousands for Veere and even hundreds of thousands for Rotterdam in a year time, the number of TripAdvisor reviews stays in the thousands. For Veere, 2.468 reviews were collected during the year, and for Rotterdam this number was 7.324. On the other side however, the quality of the data is higher, as more information is gathered for each datapoint. The TripAdvisor reviews of restaurants and cafes have been collected with X and Y coordinates and can therefore, in contrary to the Instagram data, also be plotted on a map.

6.2 Temporal results

Both the TripAdvisor data and Instagram data contain timestamps, which enables multiple forms temporal analyses. A first impression of the Instagram data in Rotterdam is displayed in figure 6.2, where all Instagram posts that have been collected in Rotterdam are shown over the year.

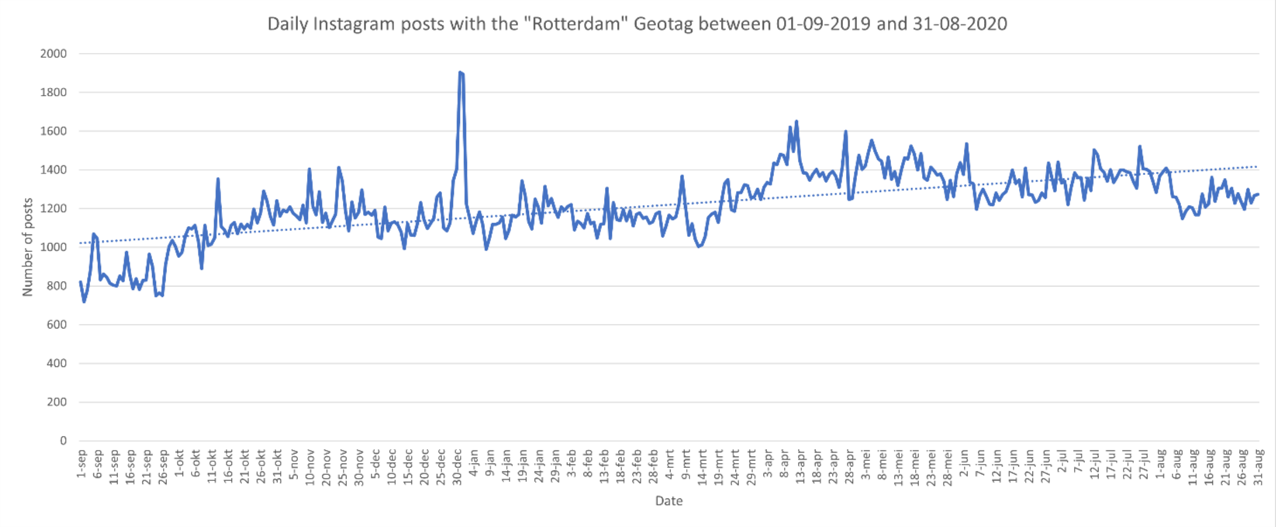


Figure 6.2: Daily Instagram posts with the “Rotterdam” geotag between 01-09-2019 and 31-08-2020. Source: Instagram (2020)

A first thing that is immediately visible is the peak of posts around the 31st of December 2019 during New Year’s Eve. Another observation is the rise in number of posts from the lockdown onwards around the 12th of March, when it was announced on national television (Government of the Netherlands, 2020). From this fact can be concluded that Instagram cannot be used in this way to determine the presence of tourists in a city, as there were few to none after the announcement of the lockdown. Furthermore, the number of posts has a positive trend throughout the year, meaning that the amount of Instagram posts per day has been increasing since September 2019.

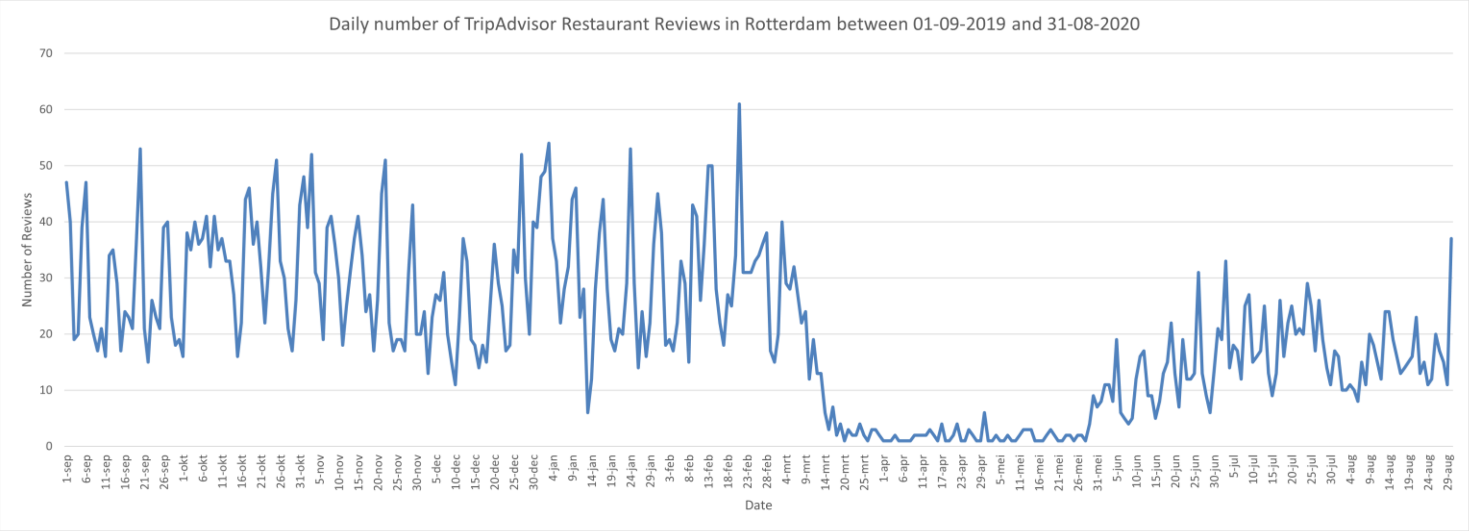


Figure 6.3: Daily number of TripAdvisor reviews in Rotterdam between 01-09-2019 and 31-08-2020.

When looking at the temporal TripAdvisor data in Rotterdam in figure 6.3, the initiated lockdown from the 12th of March is more visible. This makes sense, as restaurants and cafes were closed during the lockdown. The few reviews that have been posted were likely from users that made use of the takeaway option, or from users that were late with posting their review. Besides the lockdown, there is also a clear pattern with ups and downs visible for the rest of the year. TripAdvisor reviews mainly peak during weekends in Rotterdam, while this is not so much the case for the Instagram posts. Also, when looking at the y-axis on both figure 6.2 and 6.3, the number of daily TripAdvisor reviews is way lower than the amount Instagram posts per day. To add on to that, a TripAdvisor review may be placed a couple of days after the visit to the restaurant or cafe, while a picture can be posted quite soon after it was taken. The Instagram data therefore gives a more substantiated insight in the temporal patterns in Rotterdam.

For Veere, the same type of analysis has been carried out. The number of daily Instagram posts is shown in figure 6.4. Compared to the Instagram posts of Rotterdam, Veere contains a similar pattern, with several outliers during holidays and during summer. Veere also has more visible ups and downs. This is most likely due to the lower number of posts per day, resulting in a not so static data representation as compared to Rotterdam. The standalone Instagram data from both Rotterdam and Veere already give useful insights on the temporal aspect of both municipalities. To be able to compare both municipalities in a better way, the daily number of posts have been divided through the total sum of each municipality, giving an insight in the relative number of daily Instagram posts in Veere and Rotterdam.

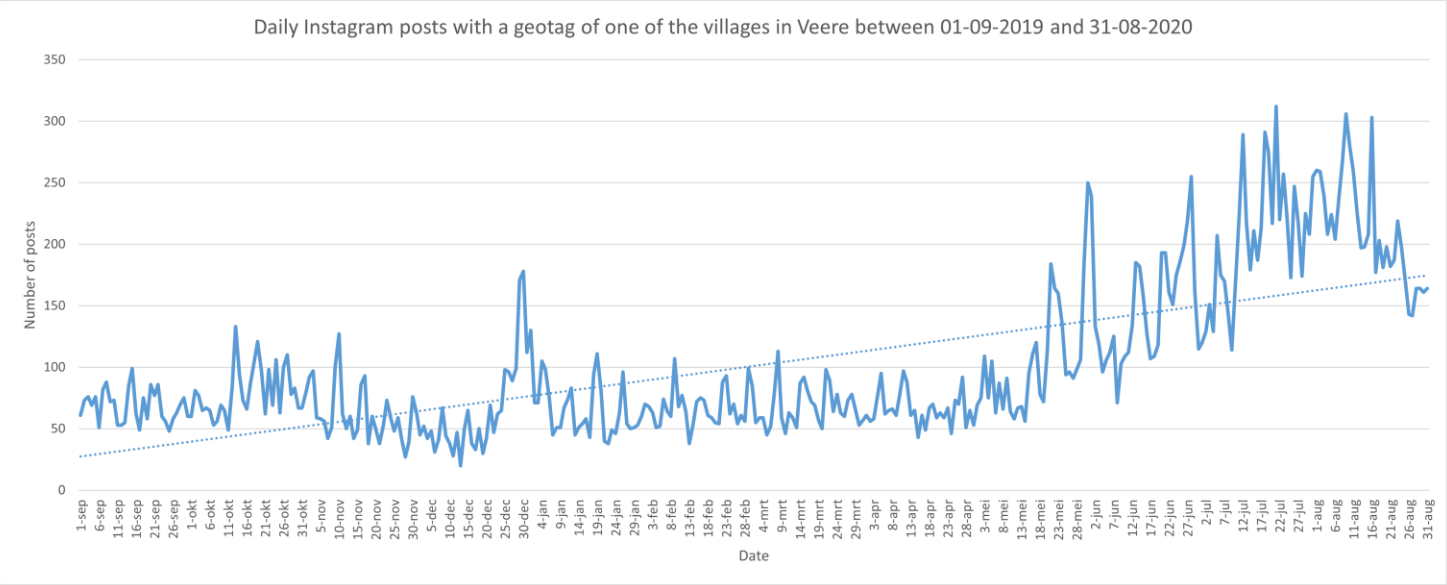


Figure 6.4: Daily Instagram posts with a geotag of one of the villages in Veere between 01-09-2019 and 31-08-2020. Source: Instagram (2020)

Figure 6.5 shows the relative number of daily Instagram posts in both Veere (grey) and Rotterdam (blue). The relative number of posts has been calculated by dividing the number of posts on a day by the total number of posts per place.

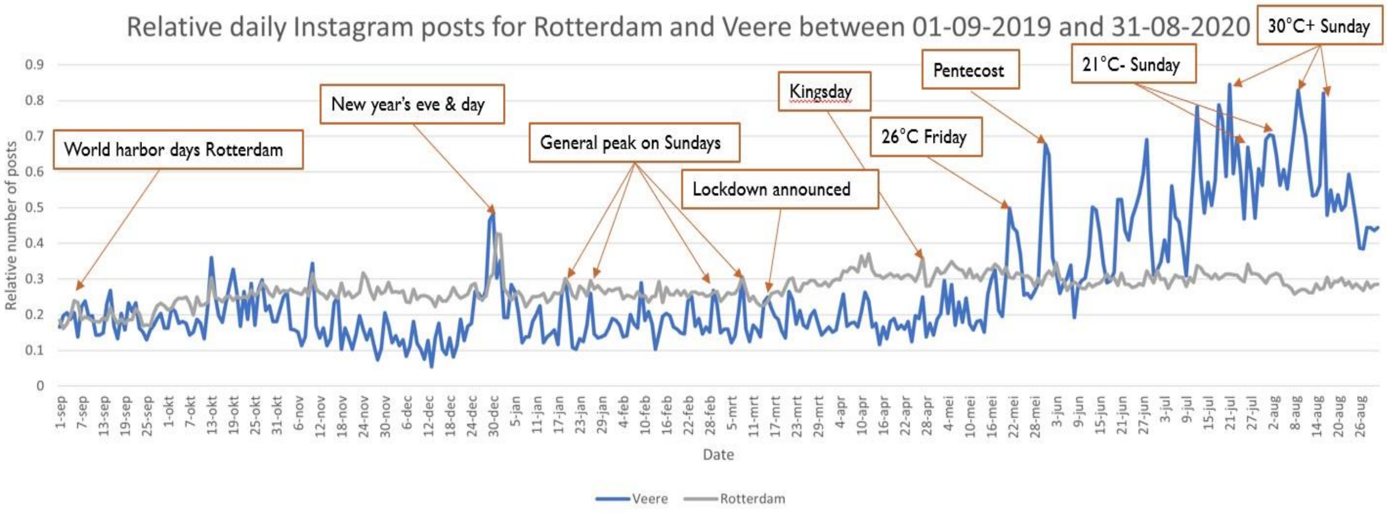


Figure 6.5: Relative Instagram posts for Rotterdam and Veere between 01-09-2019 and 31-08-2020

The posts in Rotterdam do not seem to differ a lot for each day in the year, while the posts in Veere see a clear increase during summer. To get a better insight in the data, research has been done on some important days in the period 01-09-2019 –-

31-08-2020. A small bulge at the beginning of September might be explained by the occurrence of the World harbour days in Rotterdam. Furthermore, during Christmas but most certainly during New Year's Eve & Day, there is a peak in the number of posts both in Rotterdam and in Veere. For the data of Veere, a clear pattern is visible on weekend days. The Sunday is a recurring busy day. Especially in summer, when temperatures are high, the number of Instagram posts increases in the coastal municipality.

A COVID-19 related result is the lockdown that has been announced. This happened the 12th of March. After this date, the amount of Instagram posts in Rotterdam rose, as compared to before the initiated lockdown. For Veere, a clear effect of the lockdown cannot be noticed from figure 6.5.

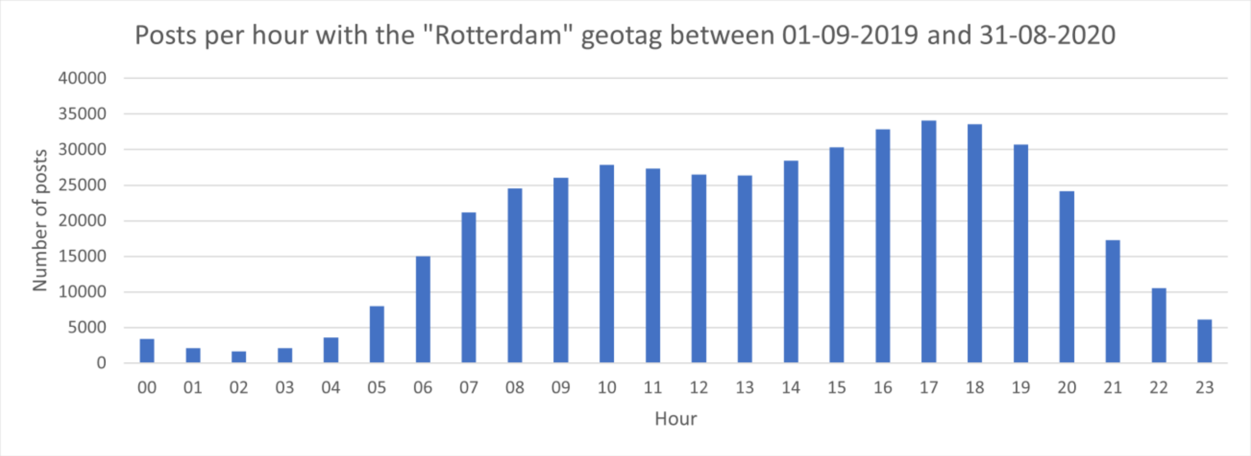


Figure 6.6: Posts per hour with the "Rotterdam" geotag between 01-09-2019 and 31-08-2020. Source: Instagram (2020)

With the Instagram data, an even more thorough analysis is possible. As the timestamp in the collected data is detailed to the very second, the number of posts for every hour can be set out. This has been done for Rotterdam in figure 6.6 and for Veere in figure 6.7. A clear and very similar pattern can be seen in the number of posts per hour for both municipalities. During night-time, the number of posts is low, while throughout the day the number of posts goes up again, with a slight drop between 11:00 and 14:00.

Looking at the hours that posts were uploaded to Instagram in Veere, the pattern leans more towards the right compared to the uploading hours in Rotterdam, meaning that relatively more pictures are posted during the late afternoon and the beginning of the evening (Figure 6.7). Also, compared to Rotterdam, 'nightlife' in the municipality of Veere does not appear to be as vivid as in Rotterdam, due to the low number of posts at night. For Rotterdam, the number of posts at night is still relatively low as compared to daytime, but more users are still active during these times.

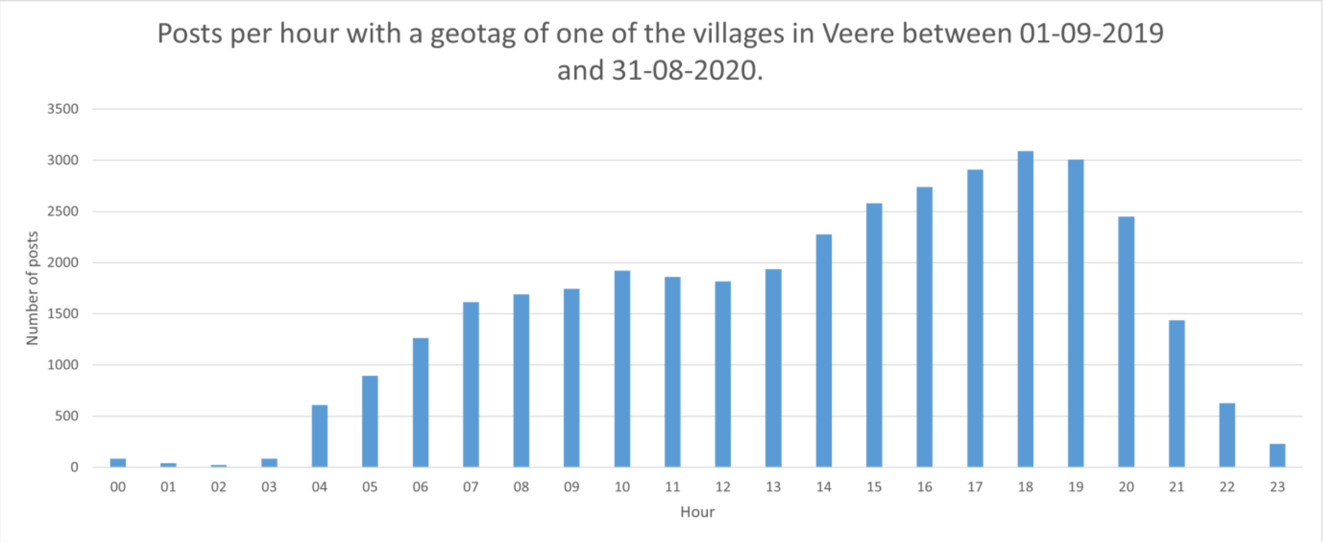


Figure 6.7: Posts per hour with a geotag of one of the villages in Veere between 01-09-2019 and 31-08-2020. Source: Instagram (2020)

As the complete timestamp is known for Instagram, it is also possible to look at what weekday a post has been uploaded. Through the TripAdvisor data, this is also possible for the reviews. Both the weekdays of Instagram posts and TripAdvisor reviews are displayed for both municipalities in figure 6.8. What is visible is that the TripAdvisor reviews for both Rotterdam and Veere and the Instagram posts in Veere show a similar pattern when it comes to weekdays. The Instagram posts in Rotterdam show an unexpected increase in post on Wednesday. A good reason for this could be that both Christmas and New Year’s Eve fell on a Wednesday in 2020. Another look at figure 6.2 confirms that during this time, the amount of Instagram posts was indeed higher than usual.

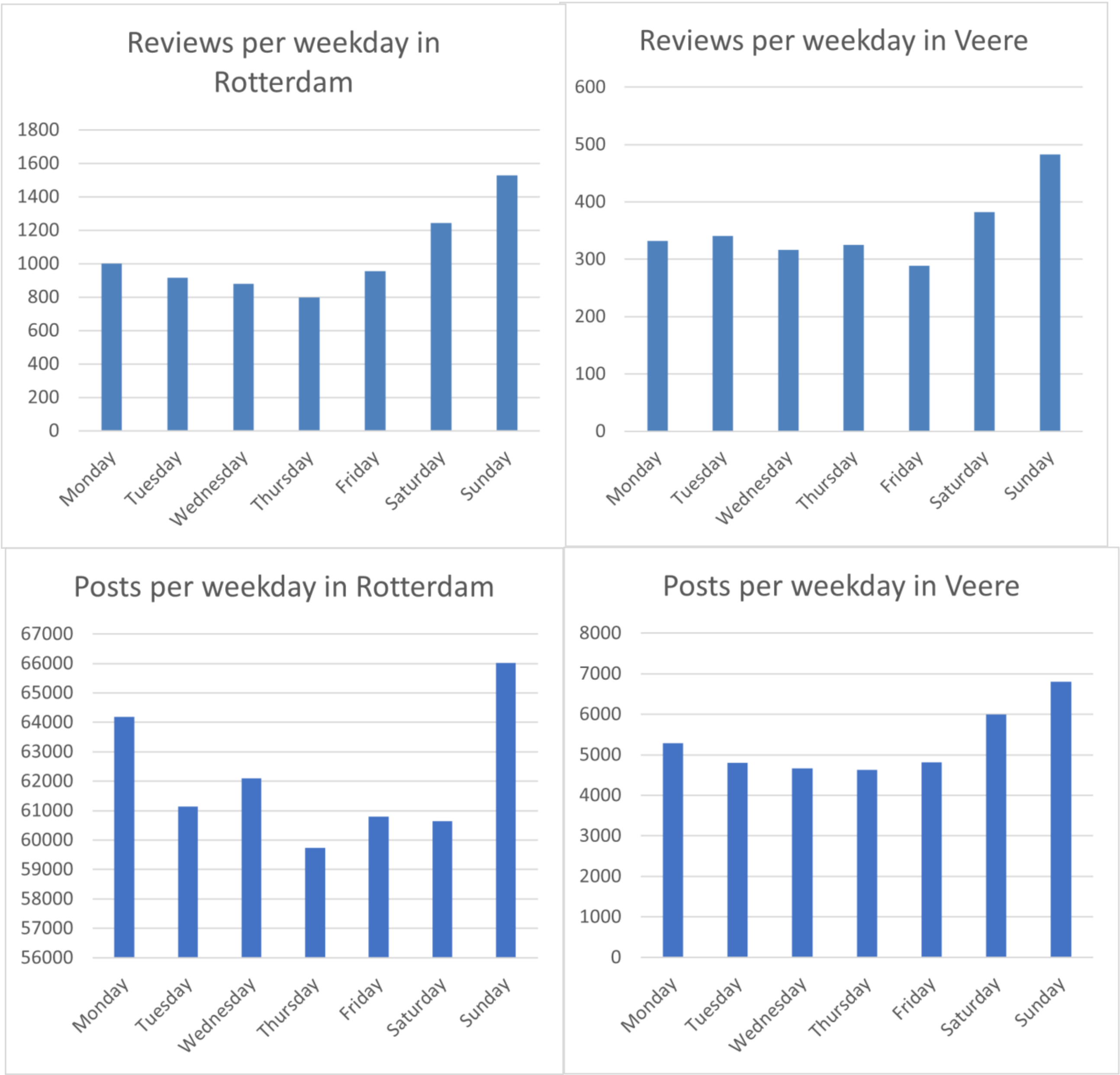
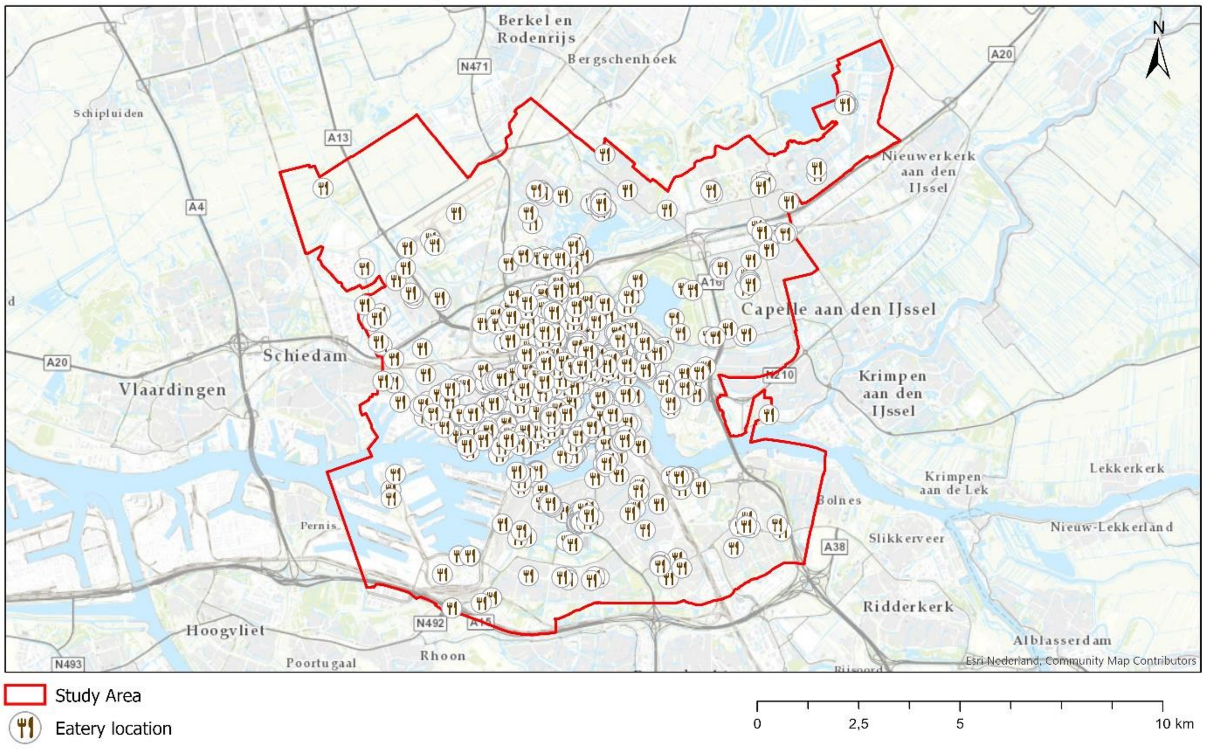


Figure 6.8: Instagram posts and TripAdvisor reviews per weekday in Rotterdam & Veere

6.3 Spatial results

While the Instagram data has very detailed temporal aspects, the spatial segment of the data is limited for now to the municipality as a whole. Therefore, the TripAdvisor data will mostly be used for spatial analysis. The X and Y coordinates have been displayed on maps for Rotterdam and Veere and are shown respectively in figures 6.9 and 6.10. The distribution of restaurants and cafes of these two figures embrace the differences in degree of urbanity between the two research areas. In Rotterdam, the eatery locations appear to be mainly clustered near the city centre, while in Veere smaller clusters appear in the different coastal towns and villages.

TripAdvisor eatery locations in the municipality of Rotterdam



Sources: Esri Nederland, 2019; TripAdvisor, 2020

Figure 6.9: TripAdvisor eatery locations in the municipality of Rotterdam.

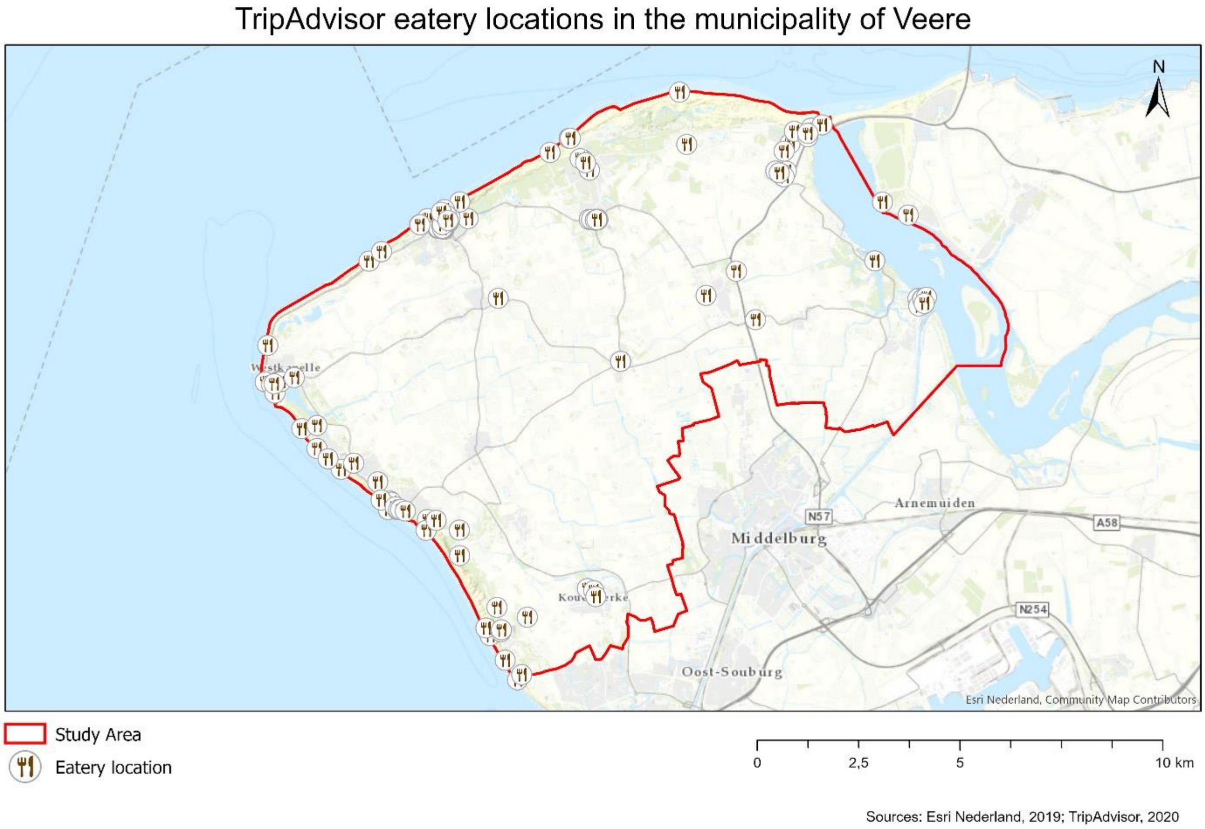

TripAdvisor eatery locations in the municipality of Veere
N
Westkapelle
Koudekerke
Middelburg
Arnemuiden
Oost-Souburg
N57
A58
N254
Esri Nederland, Community Map Contributors
Study Area
Eatery location
0
2,5
5
10 km
Sources: Esri Nederland, 2019; TripAdvisor, 2020

Figure 6.10: TripAdvisor eatery locations in the municipality of Veere.

While figures 6.9 and 6.10 already give an impression of where clusters of eatery locations appear, the cluttered point data does not give a clear overview of the distribution of restaurants and cafes. Therefore, a heatmap is applied as a type of data visualisation to identify the different clusters. This is visible for Rotterdam in figure 6.11. The amount of eatery locations in Veere is too small for a decent data visualisation through a heatmap, as the distribution of data points is unequally spread over the study area. The eatery locations in the municipality of Veere tend to follow the patterns of population density, as opposed to centrality in Rotterdam. The figure of Veere can be found in Appendix D.1.

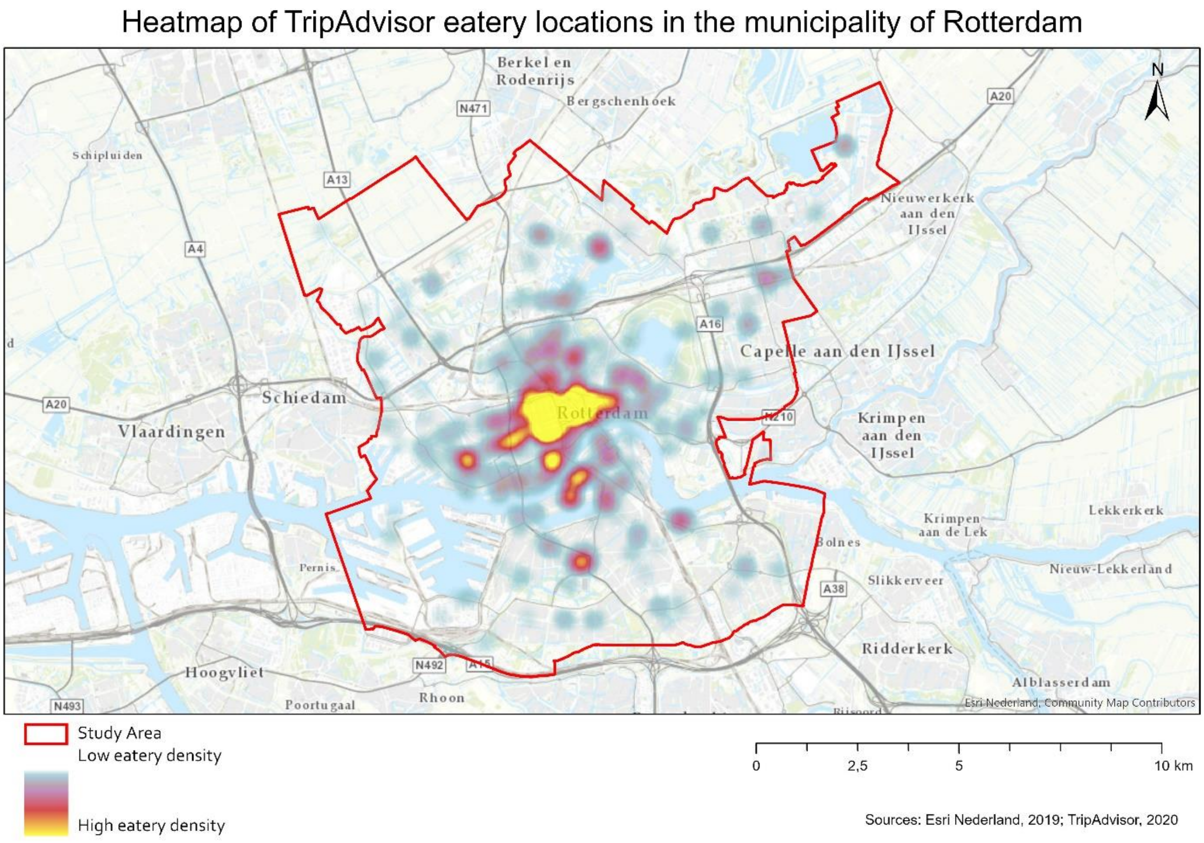
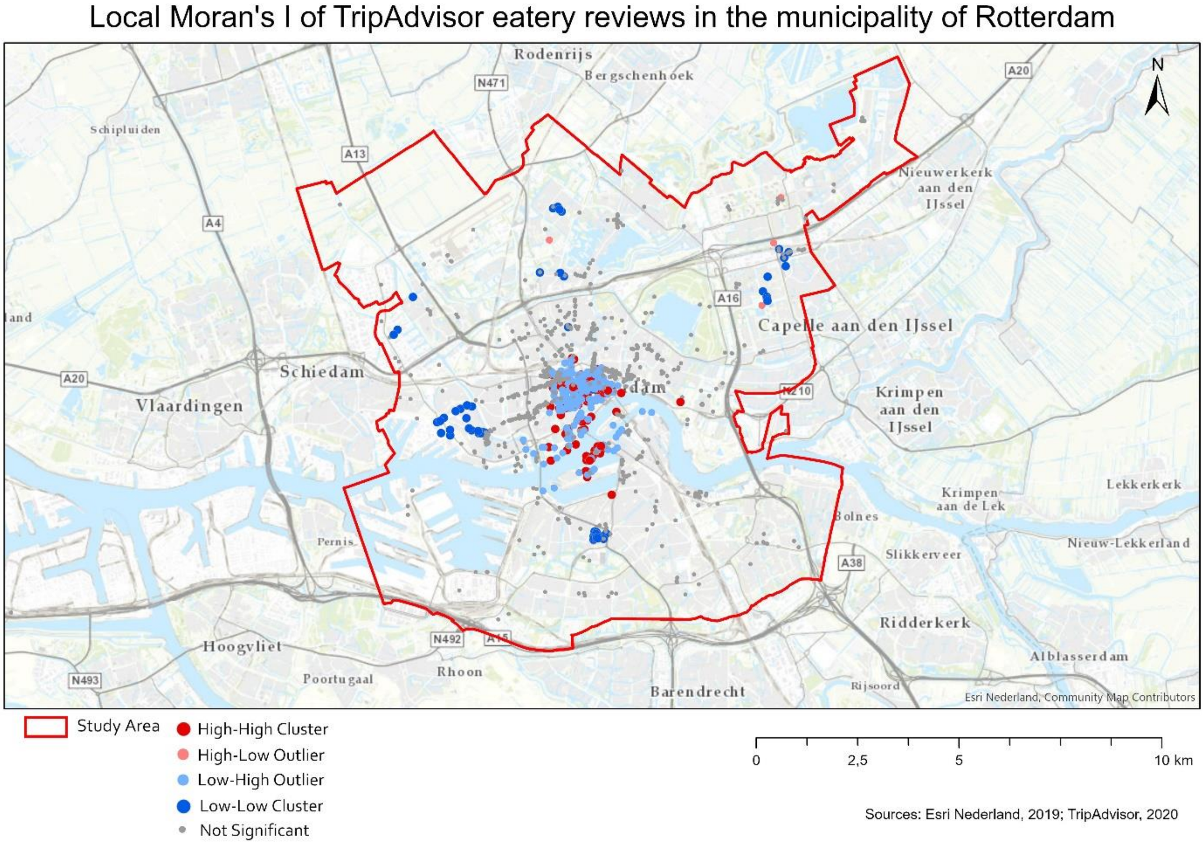


Figure 6.11: Heatmap of TripAdvisor eatery locations in the municipality of Rotterdam.

While a heatmap provides a quick overview in terms of visualisation, its methodology is more of a blackbox, as a heatmap does not take the complexity of big geo-datasets like these into account. For a more statistical substantiated methodology, a cluster analysis is performed through calculating the Moran’s I. This is done with the help of a so called ‘population field’, which in this case stands for the reviews each restaurant has had over the research period. The aim for this method is to get an insight in the most visited restaurants by looking at the most reviewed restaurants, which has also been done in earlier research (Ganzaroli, De Noni & Van Baalen, 2017; Van der Zee & Bertocchi, 2018).

The result of the Moran’s I cluster analysis is shown in Figure 6.12 and gives an overview of the eatery clusters in Rotterdam. The complete city centre shows up as a high-high cluster or as a low-high outlier. The high-high cluster points are explaining that most of the restaurant reviews are placed in the city centre. The low-high outliers indicate that there are also some restaurants in the city centre with not that many reviews, while the other surrounding restaurants do have a lot of reviews.

The cluster appears to be located mainly above the Maas, but also in Rotterdam Zuid several high-high clusters can be found, indicating that people also cross the Maas when looking for a place to eat or drink. This is in line with the public policy to increase the potential of Rotterdam Zuid, that is already being encouraged by the government for over ten years (Nationaal Programma Rotterdam Zuid, 2021).



Sources: Esri Nederland, 2019; TripAdvisor, 2020

Figure 6.12: Local Moran's I of TripAdvisor eatery reviews in the municipality of Rotterdam.

Figure 6.13 shows the same Moran's I cluster analysis, but now only from eatery locations in and around the city centre of Rotterdam. Here, it is even better visible that the restaurants and cafes in Rotterdam Zuid are performing well. The number of high-high clusters in “De Kop van Zuid” as well as a part of Katendrecht (both located below on the map in the middle) is relatively higher than in the city centre, explaining that on average the places in Rotterdam Zuid are performing very well. In the city centre itself there are still numerous low-high outliers to be found. This does not necessarily mean that these places are not performing well. It could also be the case that the eatery location is relatively new, or that its customers are not very active on TripAdvisor. Also, the contours of once central hot spot are visible already.

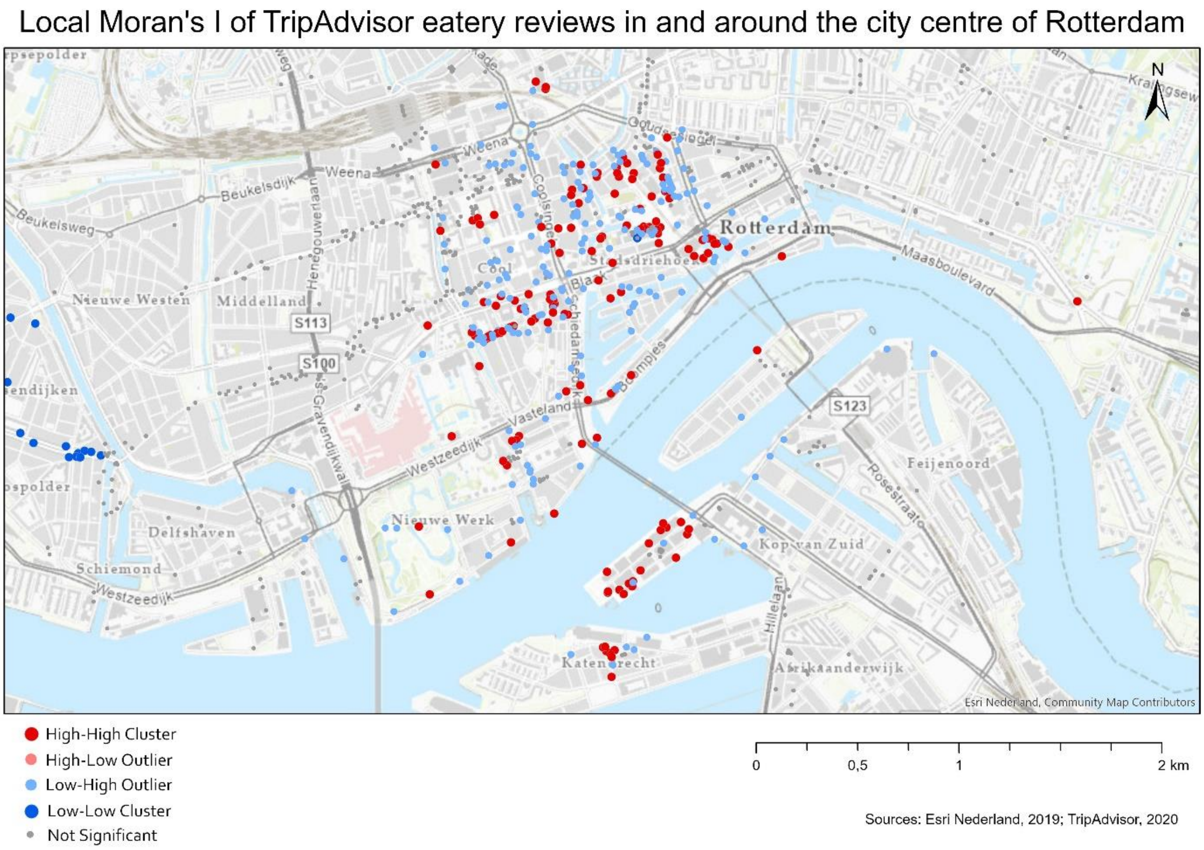


Figure 6.13: Local Moran's I of TripAdvisor eatery reviews in and around the city centre of Rotterdam.

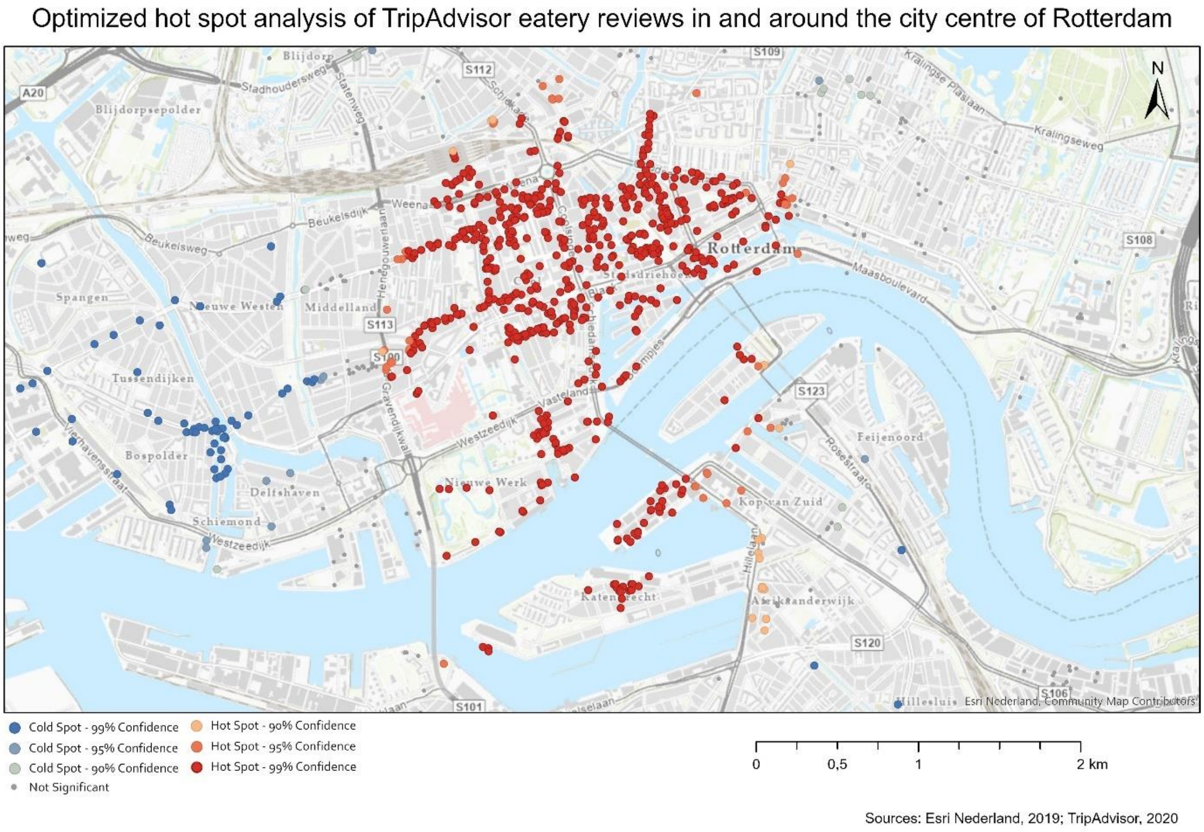


Figure 6.14: Optimized hot spot analysis of TripAdvisor eatery reviews in and around the city centre of Rotterdam.

The Moran's I result can be used to perform an optimized hot spot analysis. This gives the result of figure 6.14, where the pattern in the city centre of figures 6.12 and 6.13 is confirmed as a hot spot with 99% confidence. The figure of the complete municipality of Rotterdam can be found in Appendix D.2. As the hot spot is clearly located in and around the city centre, the choice has been made to focus the map on this location. The cluster appears to be mainly bounded by the Central Station, some main roads circling the city centre (Weena, 's Gravendijkwal, Goudsesingel) and the river Maas, apart from Noordereiland, Kop van Zuid, and Katendrecht located in Rotterdam Zuid.

6.4 Origin of people

While the answers to questions on when people are where are being managed, the question regarding where people are from is still unanswered. From the review profile of TripAdvisor, the origin can sometimes be extracted, depending on whether a user chose to fill this in or not. A sidenote that should be made is that for example The United States of America is not included in this analysis, as these are listed per state on the TripAdvisor website, resulting in a division of states instead of one country. For visualisation purposes, the choice has been made to include the first seven countries in the pie chart, and the rest is of the countries are placed in the 'other' category.

Figures 6.15 and 6.16 show the origin of TripAdvisor users for respectively Veere and Rotterdam. While the list of countries visiting Veere and Rotterdam is quite similar, the ratio surely differs. While around half of the people visiting the two places originates from The Netherlands, it can be said that Rotterdam welcomes relatively more 'native tourists' than Veere based on eatery reviews. Another remarkable result is that 29% of the reviewers are located from Belgium or Germany in Veere, while for Rotterdam this is just 7%. Contrastingly, the 'other' category for Veere is 21%, while for Rotterdam this number is at 35%. This might be an indication for Veere receiving more regional visitors than Rotterdam, as Rotterdam might be receiving more international visitors from all over the world.

While the two figures seem to give a clear overview of the exact visitor rate, it needs to be said that some countries might not be familiar with the TripAdvisor application, and that users can decide for themselves if they include their origin in their user profile. The results might therefore give a skewed image on the origin. However, this is the best method possible with the obtained datasets.

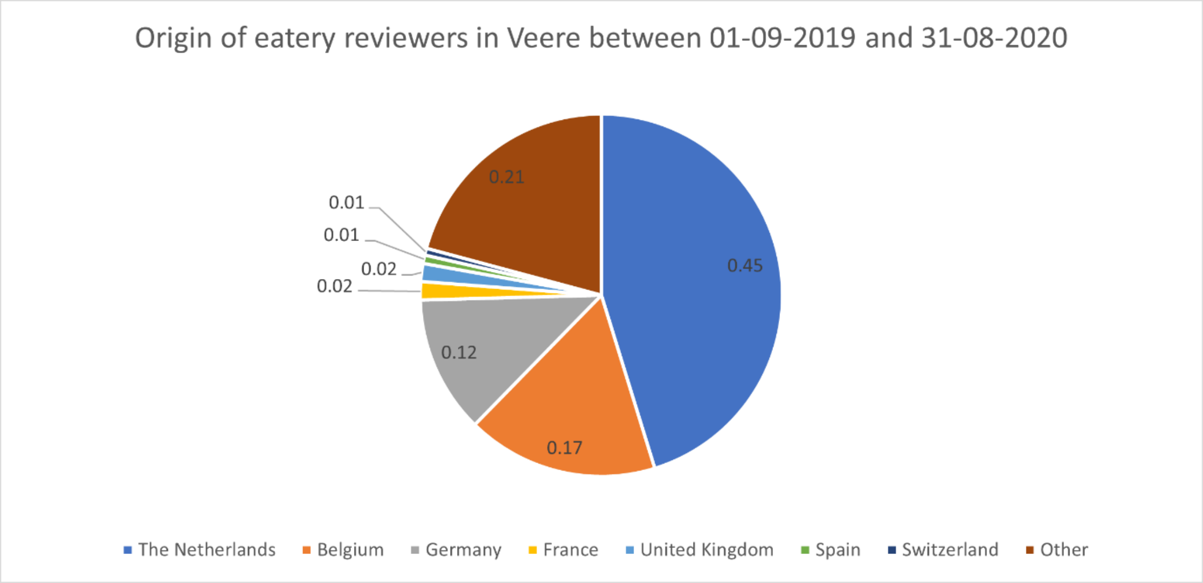


Figure 6.15: Origin of eatery reviewers in Veere between 01-09-2019 and 31-08-2020.

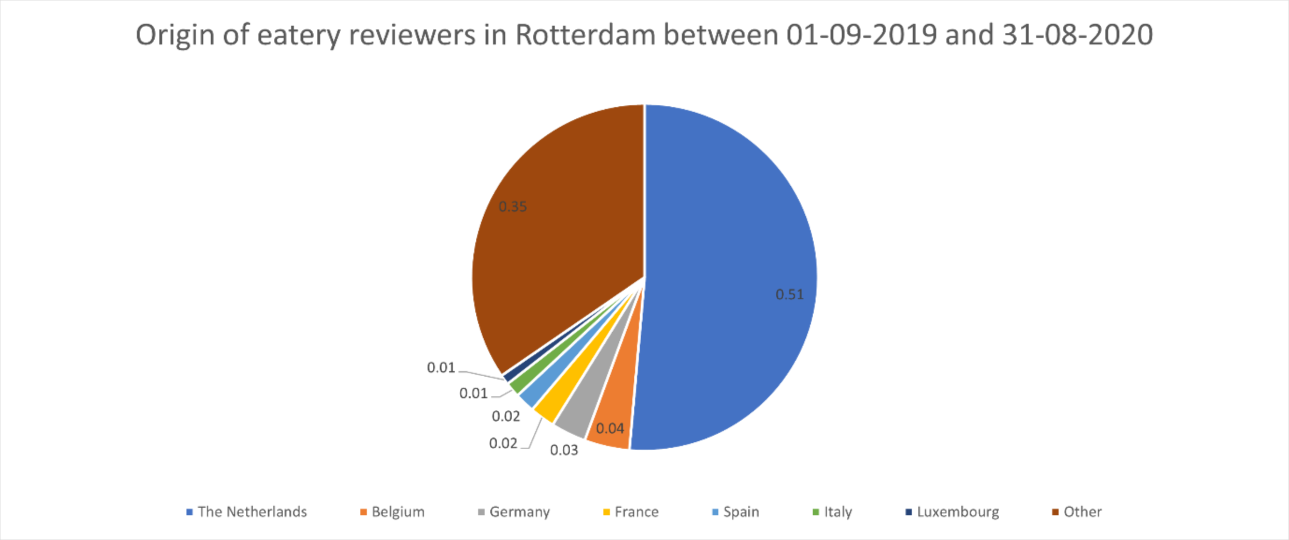


Figure 6.16: Origin of eatery reviewers in Veere between 01-09-2019 and 31-08-2020

7. Discussion

This research project has been successful and provided various insightful results. However, it should be stated that the specific choices that have been made during this project have influenced the detected spatial and temporal outcomes. This section discusses the data, methods used and provides a reflection on the obtained results. In the reflection process, an expert opinion has been consulted to discuss the obtained results. This has been done in an interview with R. Bron, the CEO and co-founder of Resono, which is a data company specialised in giving insights in data on human behaviour. Furthermore, a strong focus is given to the quality of the dataset and the developed methodology.

7.1 Obtained results

The results that were obtained during this research project give interesting insights in both temporal and spatial patterns of people. This paragraph will delve deeper into the meaning behind these results by explaining what can be concluded from the several analysis forms. First, the temporal results will be discussed, second, the spatial results will be further elaborated and lastly, the results based on the origin of people will be further discussed.

The temporal results of both Instagram and TripAdvisor came to good use for a better understanding of the time distribution of people in Rotterdam and Veere. Looking at the explanatory power of other scientific work with temporal methods used, like that from Li, Goodchild & Xu (2013), the results from this study can also be considered as useful. There has been made use of both an urban area and a rural area, which is also to be seen in the data. The Instagram data in Rotterdam shows, with exception of special events like New Year and Christmas, a continuous flow of people in the city throughout the whole year, without many big outliers. Even COVID-19 and its corresponding measures do not seem to influence the temporal data. This is likely the reason because, pandemic or not, there are always people in Rotterdam, which has also been confirmed by R. Bron (personal communication, February 24, 2021).

For Veere, the situation is different. Not only does the distribution look spikier, explaining the difference between weekdays and weekends, but also the data seems to be pattern seems to be more influenced by events that occur. This can easily be deducted from figure 6.5. The number of posts and people in Veere increases when the weather is nice, or when there are public holidays. A reason that these patterns show up is likely to do with the 'base-amount' of locals posting posts at their location. Further research could potentially use a more often used method to distinguish tourists from locals in the data. This has for example already been done by Li, Zhou & Wang in 2018 with Flickr data, and has proven to be successful.

For the spatial results, the Instagram data appeared to be not very useful in its current form, as its scale is not detailed enough. The TripAdvisor data however lends itself well for multiple types of spatial analysis, as seen in the previous chapter. One method that has been used is the Getis-Ord hot spot analysis, which has also been used previously in other scientific works (Colak, Memisoglu, Erbas & Bediroglu, 2018; Sánchez-Martín, Rengifo-Gallego & Blas-Morato, 2019; Van der Zee, Bertocchi & Vanneste, 2020). Applying this method to the city of Rotterdam, clearly shows an "invisible wall" with significant hot-spots on the inside and insignificant or even cold spots on the outside (figure 6.14). The "invisible wall" has also been confirmed by R. Bron. He explains that it is often the case that people tend to not cross a railroad or river, when not being nudged by POI's on the other side (personal communication, February 24, 2021). As this research project is a proof-of-concept, this method can eventually be applied to more cities and help policy makers give insights in the hot spots in their city.

The results on the origin of people have the ability to also provide useful information for policy makers. Easy visualization makes it possible to see whether there is a lot of 'local tourism' from inside the country and neighboring countires like in Veere, or that there is more internationally diverse tourism like in Rotterdam. Knowing who the people are in your city gives an advantage on how to market your city to either attract more tourists, or guide tourists to certain places. All insights together (where, when, who) can give a clear overview of the users of a city.

7.2 Used data

A thorough investigation resulted in using Instagram and TripAdvisor for this research project. Although the results appear to be showing relevant patterns, the choice could also have been made to use other datasets from other UGC websites or applications. The reason for using Instagram mainly originates from the fact that it has not been used very often in scientific research projects, while it is currently one of the biggest UGC/social media platforms. There were however also downsides that came paired with this choice.

A first big disadvantage of the Instagram data is the lacking exact geolocation in its data. This might therefore not give a detailed enough image for policy makers. There are however workarounds possible, but these also are accompanied with several other drawbacks. In this research, use has only been made of posts from the municipality of Rotterdam and the places in the municipality of Veere. Posts could also be collected on for instance neighbourhood level, but not that many Instagram users choose this as a geotag.

Besides the scale of the data, there is also an issue with regards to privacy. Users of the Instagram application can select the option to hide their profile for outsiders and only show the content of it to the user's own follower base. While it is unclear how many Instagram users have a private profile, it can be assumed that quite some data is missing from these accounts. To add on to that, it should be mentioned that the users with open accounts are also not obliged to choose a geotag for their post. This could also be left out of the post.

These disadvantages are however only taking the UGC application users themselves into account. There is however also another exclusion happening, namely that of people that are not even using geosocial applications. This is for example pointed out by Tenkanen et al. (2017), while doing research on using geosocial data to investigate nature park visitors. Here, some unvisited small nature parks showed high correlations while some highly visited nature parks gave a significant discrepancy between the actual visitor numbers and the social media statistics.

This however cannot be solved and counts for every application that works with UGC. One of the problems of this issue is for example the exclusion of some elderly people from the research, while they are also frequent visitors of restaurants and cities in general. Another group of users that is excluded might depend on the origin of people. While our Western society is familiar with applications like TripAdvisor and Instagram, this is not directly the case for African or Asian countries. Therefore, these people may also be at the exact same location as the Instagram or TripAdvisor users in this research, but they post their pictures to another type of social media than Instagram or write their reviews on a platform other than TripAdvisor.

While these challenges occur with every UGC data source, many scientists in literature argue that this is still the best practice to investigate spatial patterns of people (Brandt et al., 2017, Miah et al., 2017, Van der Drift, 2015). To add to that, Miah et al. (2017) argues that "…in time though, it is anticipated that older tourists will increasingly take photos and upload them to social media." Therefore, pioneering in the field of geosocial data analysis now, might reap benefits in the future.

7.3 Used methodology

During this proof-of-concept research, use has been made of several analysis techniques. These have been displayed in chapter 6 and are in line with the scope of the research project. There a however more types of analysis possible with the collected data. This paragraph will elaborate on the other types of analysis that could be conducted with the collected data but are not in the direct scope of this research project.

With the Instagram dataset for example, it is possible to perform a sentiment analysis. The captions underneath Instagram posts, which are pieces of text that are posted together with the pictures, can provide useful information about the sentiment of a post (Neri, Aliprandi, Capeci, Cuadros & By, 2012). Also, there is the possibility to perform analysis based on the hashtags (#) people use in their captions.

For the TripAdvisor reviews, a sentiment analysis can also be performed. While most of the reviews contain information about the food and/or drinks that have been consumed, the ambiance and underlying sentiment of the place can be extracted from the text. For Instagram, a sentiment analysis could also be performed based on the pictures posted. This has been proven a successful method by Wang & Li (2015), who have developed an Unsupervised SEntiment Analysis (USEA). Their approach exploits relations among visual content and relevant contextual information to bridge the semantic gap in the prediction of image sentiments. Extra care is however required with regards to privacy, as personal data in the form of pictures are used for analytical goals.

While for this research only eatery locations have been scraped from TripAdvisor, there is also the possibility to scrape POI's and their reviews. This can be done in a similar way and could potentially also provide useful answers to the questions where people are at what time of the year.

A downside of working with all types of TripAdvisor data is the inability to delve into the past. On the one hand, TripAdvisor did not exist (in its current form) before 2007. This implies that research can only be done back to the year 2007. However, when restaurants or cafes shut down due to reasons, their TripAdvisor profiles and all the reviews accompanied with it, will be deleted from the website. Due to this, and the fact that not that many people used TripAdvisor yet back in 2007, the data cannot be used for earlier research. To visualize this, the same temporal analysis has been performed as in figure 6.3, but now over the whole timeframe since 2007. This figure can be found in Appendix D.3

7.4 Reflection on results

While the results of this research project provided interesting insights, it is also important to check whether the obtained results give a sufficient representation of reality. On the one hand, the results certainly do. The temporal Instagram data gives an accurate image of the situation on busy beach days for example. Also, there is a clear spike during weekends and holidays visible in both the TripAdvisor and the Instagram data, which can also be deducted from the actual spikes in reality.

Not all the data does however consist of complete explanatory power. The lockdown effect due to COVID-19 is clearly visible with the TripAdvisor data, but the Instagram data does not show a dent. Contrastingly, the number of Instagram posts in Rotterdam do even rise after the 12th of March. This could also imply that Instagram represents the continuous flow of people in a city, which has also been confirmed by

R. Bron (personal communication, February 24, 2021). However, in that way, it cannot really explain the temporal patterns for big cities with the current analysis that have been performed, as the number of posts is at a constant high with a few outliers like New Year's Eve and Christmas.

With regards to privacy and obtainability of the data, there are limits to the collection process. As data is gathered from people, the data should always be handled with care in the first place (Smith et al., 2012). In this research for example, no part of the presented data can trace back to an individual or minority groups. Referring to privacy laws like the GDPR, the data collection process has been legal. The data that has been gathered is all open data, implying that all users agreed to openly share their data on the internet. The next step that had to be checked is whether or not the applications Instagram and TripAdvisor allow web scraping. Scientific works by Smith et al. (2012) and Fiesler et al. (2020) did not find any strict laws or regulations on this, however it is important to not make an endless stream of requests to a website or application, delaying the servers of it. During this study, the Instagram web crawler collected the information at a speed of about four posts per second. This is comparable to regular browsing in the application itself (which is done by millions), and therefore not harmful to the servers. The same applies to the TripAdvisor web crawler. It had a few seconds of sleep integrated in the code, and it therefore sent a regular number of requests from time to time to the servers, making it in no way any harmful.

Another issue, which is related to privacy, is that some websites do not like their information getting collected with an automated web crawler. With collecting the TripAdvisor data, no issues were encountered. Instagram however did block the IP address from accessing their website after scraping 4000-8000 posts. A workaround that was used to tackle this issue had to do with using an IP rotation system, to still collect all the posts. This made the collection process way slower than intended.

Collecting the data has been a time intensive challenge. Especially when taking into account that the data has only been collected from Rotterdam and Veere as a proof-of-concept. Collecting the data for the whole of The Netherlands for example would likely take several months. There are chances that the data collection system can be further developed to potentially decrease this time. However, concerning scraping ethics the limits on the number of posts or reviews collected per second will always stay the same.

7.5 Further research

This paragraph will give recommendations for further research in this study area. While working with UGC data is already done for quite a while, there are still many things to improve and/or take note of.

An improvement regarding the Instagram data could be to look at the number of accounts that posted per day instead of the number of posts per day. As the research is about people, and not about posts, this will probably give a better representation. A more general improvement with regards to the data that has been used could be to create a better synergy between the two datasets. Right now, comparing the results from reviews of restaurants and cafes and Instagram posts might sound like comparing apples to oranges. A possibility to solve this issue could be to perform a content analysis on both the Instagram captions and TripAdvisor reviews, to check for similarities or differences in the semantics.

While the results have proven to provide useful insights in where people are when and where they are from, it could be improved. The choice for a municipality (Rotterdam) or villages (Veere) from Instagram gave a good overview of the study areas in general. However, it is also possible to scrape posts with a tag of POI’s (like the Markthal or the Central Station) on Instagram. For TripAdvisor, the same is also possible. Instead of collecting reviews of eatery locations, the reviews from attractions/POI’s can also be collected. In this way, the spatial and temporal data could then have been combined in a better way.

8. Conclusion

Over the past years, there have been more and more possibilities with data. Since the introduction of the internet, the amount of available data has ever increased. In scientific literature on tourism, social media has already proven itself useful for research (Brandt, Bendler & Neumann, 2017; Chua et al., 2016; Giglio et al., 2019; Lu & Stepchenkova, 2015; Vu et al., 2018; Wu et al., 2018; Xiang & Gretzel, 2010). Especially in the years 2015-2019, the consensus was positive towards the endless opportunities that UGC data can provide. Nowadays however, more questions pop up on the representativity, ethics, validity, and the collection and processing methodology of the use of geosocial data (Alaei, Becken & Stantic, 2019; Bruns, 2019; Fiesler, Beard & Keegan, 2020; Krotov & Silva, 2018). The goal of this research project is to come up with a proof-of-concept on the collection and visualisation of both temporal and spatial patterns of individuals. This paragraph will aim to answer the drafted sub-question, and by doing so, also provide an answer to the main research question:

"How can different sources of user generated geosocial data be used to visualise spatial and temporal patterns of individuals in Rotterdam & Veere?"

8.1 Data investigation & collection

SQ1 concerns the data investigation and collection process and is posed as follows:

SQ1: What methods can be used to obtain geosocial data and what geosocial data is available and usable?

There are numerous websites and applications to choose from when it comes to collecting geosocial data. Previous research has already worked often with data originating from Flickr, TripAdvisor, Google Reviews, Twitter, and more (Table 3.1). Often, these research projects limit themselves to just one social media source to base their research on. The novelty of this research project lies in the choice that has been made for using two social media sources to see if both datasets are able to support each other, or if it is a matter of comparing apples to oranges. This is done to potentially increase the validity of the research.

For the choice of what geosocial media data source to use in this research project, there are several options available. In this case, the choice has been made to investigate the well-known and already widely used social media application TripAdvisor, as it provides both temporal and spatial data. In addition to that, the application also provides exact information on the origin of people. For the second data source, the choice has been made to use the more novel application Instagram, that has not been frequently used yet in previous research, but is used worldwide by more than a billion people (Statista, 2020). While this dataset does not comprise of very detailed spatial information, it is a rich source of temporal data. The big user database caters for many posts per day, giving a decent substantiation of the results.

The data preparation can be seen as the lion's share of this research project. While the internet is filled with examples and pieces of programming code, the data collection comprised of many exceptions on the rule. A reason for this is the ever-changing infrastructure of websites/applications, making it impossible to create a long-lasting sustainable web scraping script. However, in the end, all data has been collected successfully from both Instagram and TripAdvisor through self-written web crawlers. Therefore, it can be stated that while it is a time intensive task, it is possible to collect geosocial data using web crawlers. With this collected data, different types of analyses can be performed. This will be further explained with the help of SQ2.

8.2 Data analysis

SQ2: How can geosocial data be analysed to visualise spatial and temporal patterns of people?

The collection process is only a part of the work that needs to be done with the data to come to a good result. The obtained data comprises of several different characteristics, among which temporal data and spatial data. Therefore, the data can be used for multiple types of analysis, besides the regular descriptive statistics that already give valuable insights in the spatial and temporal patterns of individuals.

An example of such an analysis is the Getis-Ord hot spot analysis (Getis & Ord, 1992). This analysis is used in this research project to identify the locations of significant clusters, as has been done before in other studies (Colak et al., 2018; Sánchez-Martín, Rengifo-Gallego & Blas-Morato, 2019; Van der Zee, Bertocchi & Vanneste, 2020). The outcome of the results is comparable to the results in other works, meaning the data is of sufficient quality to perform analysis with. Also, the outcomes of the analysis give interesting insights, which will be further elaborated in the answer to SQ3.

8.3 Data representativity

SQ3: To what extent can geosocial data be used to display an accurate model compared to reality?

As aforementioned, the results of the performed analyses are similar to results from other scientific works, implying that this form of geosocial data can be used for analysis. Also, the data has been compared with other data like public holidays and the weather, giving the datasets more explanatory value. Also, as the geosocial data that has been used for this research is free and people post on voluntary basis, the data can be conceived as a clean data source. There are however ifs and buts to the results. Social media is not used by everyone and not evenly distributed among all age groups and ethnicities. Therefore, the resulting image is likely to not be an exact replica of reality. Nonetheless, the results give an adequate insight in where people are when, and where they are from. Table 8.1 gives an overview of what can and cannot be obtained from both the TripAdvisor and Instagram dataset.

Table 8.1: What can and cannot be obtained from both datasets?

	Instagram	TripAdvisor	Combination
Where are people?	This information can be obtained through collecting Instagram data; however, this highly depends on the scale. It is not possible to get the exact locations of people, as searching happens through a 'geotag'. This could be a municipality, village, neighbourhood or even a POI. The data collection method however does not lend itself to obtain all these kinds of data easily.	This information is obtainable on eatery level. Therefore, when a user visits a place, there is exact information on where this has happened. Therefore, TripAdvisor can be used for different forms of spatial analysis.	Instagram locations can be collected through different scales. TripAdvisor data can only be collected on restaurant level. It is usually comparing apples with oranges. It could be a possibility to transform the TripAdvisor data into the same scale levels as those from Instagram, but this might give some skewed results as to it just depends on where restaurants are located.
When are people somewhere?	Instagram lends itself greatly for temporal analysis, as posts are collected on the second precisely. Also, the quantity of Instagram data is sufficient to perform precise temporal analysis with the data.	The TripAdvisor data consists of just the upload day of the review. This is however not the biggest issue, as users might leave a review a day or even a few days after their visit to a restaurant or café.	While Instagram's temporal data is very detailed, the TripAdvisor data is not. Therefore, the two temporal datasets can be compared, but only on a one-day scale. This gives interesting and different views on the temporal part of the data. An example is the lockdown, where Instagram posts were still posted, but TripAdvisor reviews were close to zero.

Where are people from?	At this moment, Instagram data offers no possibility to obtain data on where a user is from. This could possibly be extracted by investigating all Instagram posts from a user, and thereby determining the ‘supposed home location’ of every user. This however is both questionable in terms of privacy, while also very time intensive.	TripAdvisor can be used to gather the origin of a user to a certain extent. Users can choose to show their location, which is often a city and/or a country. It is however questionable whether the collected data is a good representation of reality, as not all visitors from all countries use TripAdvisor.	As Instagram data does not offer locational information on users, the data can also not be combined.

8.4 Concluding remarks

To provide an answer to the main question, it can be stated that the era pioneering in the field of geosocial data is not over yet. The data collection process is still a challenge, especially once use is made of multiple social media data sources, multiple places, and/or multiple years. For now, as this research contains a proof-of-concept, the data preparation and collection method are time intensive. This, while only collecting two data sources from two places in one year. Therefore, the following can be concluded from this research project:

As long as multiple datasets can be compared with each other, it is useful to use multiple data sources. With regards to time, it is important to consider that a platform like Instagram or TripAdvisor also gains or loses popularity over time. Therefore, data that is presented for more than a year should be analysed with care. Relating to the study area of a research with geosocial data, one should keep in mind that there are differences in the explanatory power between analyses in urban areas and rural areas. Where temporal patterns are usually clearly visible in rural areas, the temporal patterns in urban areas tend to ‘drown’ in the steady flow of people in the city. On the other hand, does the spatial data explain clear patterns in urban areas, while in the rural areas the spatial patterns tend to follow a similar pattern as that of the population density. When doing research in this field, there should be taken care about the degree of urbanity of the research area.

9. References

Alaei, A. R., Becken, S., & Stantic, B. (2019). Sentiment analysis in tourism: capitalizing on big data. Journal of Travel Research, 58(2), 175-191.

Araya López, A. (2020). Policing the'Anti-Social'Tourist. Mass Tourism and'Disorderly Behaviors' in Venice, Amsterdam and Barcelona. PARTECIPAZIONE E CONFLITTO, 13(2), 1190-1207.

Albuquerque, H., Costa, C., & Martins, F. (2018). The use of geographical information systems for tourism marketing purposes in Aveiro region (Portugal). *Tourism management perspectives*, *26*, 172-178.

Ashworth, G., & Page, S. J. (2011). Urban tourism research: Recent progress and current paradoxes. *Tourism management*, *32*(1), 1-15.

Bakar, N. A., & Rosbi, S. (2020). Effect of Coronavirus disease (COVID-19) to tourism industry. International Journal of Advanced Engineering Research and Science, 7(4).

Brandajs, F., & Russo, A. P. (2019). Whose is that square? Cruise tourists' mobilities and negotiation for public space in Barcelona. Applied Mobilities, 1-25.

Brandt, T., Bendler, J., & Neumann, D. (2017). Social media analytics and value creation in urban smart tourism ecosystems. Information & Management, 54(6), 703-713.

Bruns, A. (2019). After the 'APIcalypse': social media platforms and their fight against critical scholarly research. Information, Communication & Society, 22(11), 1544-1566.

Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. Mobile networks and applications, 19(2), 171-209.

Chua, A., Servillo, L., Marcheggiani, E., & Moere, A. V. (2016). Mapping Cilento: Using geotagged social media data to characterize tourist flows in southern Italy. Tourism Management, 57, 295-310.

Colak, H. E., Memisoglu, T., Erbas, Y. S., & Bediroglu, S. (2018). Hot spot analysis based on network spatial weights to determine spatial statistics of traffic accidents in Rize, Turkey. Arabian Journal of Geosciences, 11(7), 1-11.

Eisenstein, E. L. (1980). The printing press as an agent of change. Cambridge University Press.

Fiesler, C., Beard, N., & Keegan, B. C. (2020, May). No robots, spiders, or scrapers: Legal and ethical regulation of data collection methods in social media terms of service. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 14, pp. 187-196).

Franklin, C., & Hane, P. (1992). An Introduction to Geographic Information Systems: Linking Maps to Databases [and] Maps for the Rest of Us: Affordable and Fun. *Database*, *15*(2), 12-15.

Ganzaroli, A., De Noni, I., & van Baalen, P. (2017). Vicious advice: Analyzing the impact of TripAdvisor on the quality of restaurants as part of the cultural heritage of Venice. *Tourism Management*, *61*, 501-510.

Gao, S., Liu, Y., Wang, Y., & Ma, X. (2013). Discovering spatial interaction communities from mobile phone d ata. Transactions in GIS, 17(3), 463-481.

Getis, A., & Ord, J. K. (1992). The analysis of spatial association by use of distance statistics. Geographical Analysis, 24(3), 189–206. doi:10.1111/j.1538-4632.1992.tb00261

Ghahramani, M., Zhou, M., & Wang, G. (2020). Urban sensing based on mobile phone data: approaches, applications, and challenges. IEEE/CAA Journal of Automatica Sinica, 7(3), 627-637.

Giglio, S., Bertacchini, F., Bilotta, E., & Pantano, P. (2019). Using social media to identify tourism attractiveness in six Italian cities. Tourism management, 72, 306-312.

Government of the Netherlands, (2020). New measures to stop spread of coronavirus in the Netherlands. Retrieved from: https://www.government.nl/latest/news/2020/03/12/new-measures-to-stop-spread-of-coronavirus-in-the-netherlands on 13-01-2021

Heaton, M. J., Datta, A., Finley, A. O., Furrer, R., Guinness, J., Guhaniyogi, R., ... & Lindgren, F. (2019). A case study competition among methods for analyzing large spatial data. *Journal of Agricultural, Biological and Environmental Statistics*, *24*(3), 398-425.

Huang, X., Zhang, L., & Ding, Y. (2017). The Baidu Index: Uses in predicting tourism flows–A case study of the Forbidden City. *Tourism Management*, *58*, 301-306.

Jennings, F., & Yates, J. (2009). Scrapping over data: are the data scrapers' days numbered?. Journal of Intellectual Property Law & Practice, 4(2), 120-129.

Johnson, P. A., Sieber, R. E., Magnien, N., & Ariwi, J. (2012). Automated web harvesting to collect and analyse user-generated content for tourism. Current Issues in Tourism, 15(3), 293-299.

Kádár, B. (2014). Measuring tourist activities in cities using geotagged photography. Tourism Geographies, 16(1), 88-104.

Koens, K., Postma, A., & Papp, B. (2018). Is overtourism overused? Understanding the impact of tourism in a city context. Sustainability, 10(12), 4384.

Krotov, V., & Silva, L. (2018). Legality and ethics of web scraping.

Kruizinga, P. (2016). Health tourism and health promotion at the coast. The Routledge handbook of health tourism, 386-398.

Li, D., Zhou, X., & Wang, M. (2018). Analyzing and visualizing the spatial interactions between tourists and locals: A Flickr study in ten US cities. Cities, 74, 249-258.

Li, L., Goodchild, M. F., & Xu, B. (2013). Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. Cartography and geographic information science, 40(2), 61-77.

Lu, W., & Stepchenkova, S. (2015). User-generated content as a research mode in tourism and hospitality applications: Topics, methods, and software. Journal of Hospitality Marketing & Management, 24(2), 119-154.

Mark, R. (2019). Ethics of Public Use of AI and Big Data: The Case of Amsterdam's Crowdedness Project. The ORBIT Journal, 2(2), 1-33.

Martí, P., García-Mayor, C., & Serrano-Estrada, L. (2020). Taking the urban tourist activity pulse through digital footprints. Current Issues in Tourism, 1-20.

McDonnell, M. J., & MacGregor-Fors, I. (2016). The ecological future of cities. Science, 352(6288), 936-938.

Miah, S. J., Vu, H. Q., Gammack, J., & McGrath, M. (2017). A big data analytics method for tourist behaviour analysis. *Information & Management*, *54*(6), 771-785.

Milano, C., Novelli, M., & Cheer, J. M. (2019). Overtourism and tourismphobia: A journey through four decades of tourism development, planning and local concerns.

Moran, P. A. (1950). Notes on continuous stochastic phenomena. Biometrika, 37(1/2), 17-23.

Nepal, S. K. (2020). Travel and tourism after COVID-19–business as usual or opportunity to reset?. *Tourism Geographies*, 1-5.

Neri, F., Aliprandi, C., Capeci, F., Cuadros, M., & By, T. (2012, August). Sentiment analysis on social media. In 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (pp. 919-926). IEEE.

Netherlands Board of Tourism & Conventions (NBTC), 2020. Relevante vraagstukken Perspectief 2030.

Nientied, P. (2020). Rotterdam and the question of new urban tourism. International Journal of Tourism Cities.

O'Reilly, A. M. (1986). Tourism carrying capacity: concept and issues. Tourism management, 7(4), 254-258.

Power, D. J. (2014). Using ‘Big Data’ for analytics and decision support. Journal of Decision Systems, 23(2), 222-228.

Puebla, J. G. (2018). Big data y nuevas geografías: la huella digital de las actividades humanas. Documents D’Anàlisi Geogràfica, 64, 195–217.

Robertson, C., & Feick, R. (2016). Bumps and bruises in the digital skins of cities: unevenly distributed user-generated content across US urban areas. Cartography and Geographic Information Science, 43(4), 283-300.

Sánchez-Martín, J. M., Rengifo-Gallego, J. I., & Blas-Morato, R. (2019). Hot spot analysis versus cluster and outlier analysis: an enquiry into the grouping of rural accommodation in Extremadura (Spain). ISPRS International Journal of Geo-Information, 8(4), 176.

Schoder, D., Gloor, P. A., & Metaxas, P. T. (2013). Social media and collective intelligence—ongoing and future research streams. KI-Künstliche Intelligenz, 27(1), 9-15.

Shelton, T. (2017). Spatialities of data: mapping social media 'beyond the geotag'. GeoJournal, 82(4), 721-734.

Smith, M., Szongott, C., Henne, B., & Von Voigt, G. (2012, June). Big data privacy issues in public social media. In 2012 6th IEEE international conference on digital ecosystems and technologies (DEST) (pp. 1-6). IEEE.

Statista (2020). Distribution of Instagram users worldwide as of July 2020, by age and gender. Retrieved from: https://www.statista.com/statistics/325587/instagram-global-age-group/ on 26-09-2020.

Statista (2021). Number of worldwide social network users. Retrieved from: https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/ on 13-02-2021.

The Verge (2018). The man behind Flickr on making the service 'awesome again'. Retrieved from https://www.theverge.com/2018/3/20/4121574/flickr-chief-markus-spiering-talks-photos-and-marissa-mayer on November 22, 2020.

Tenkanen, H., Di Minin, E., Heikinheimo, V., Hausmann, A., Herbst, M., Kajala, L., & Toivonen, T. (2017). Instagram, Flickr, or Twitter: Assessing the usability of social media data for visitor monitoring in protected areas. Scientific reports, 7(1), 1-11.

Van der Drift, S. (2015). Revealing spatial and temporal patterns from Flickr photography.

Van der Zee, E., & Bertocchi, D. (2018). Finding patterns in urban tourist behaviour: a social network analysis approach based on TripAdvisor reviews. *Information Technology & Tourism*, *20*(1-4), 153-180.

Van der Zee, E., Bertocchi, D., & Vanneste, D. (2020). Distribution of tourists within urban heritage destinations: A hot spot/cold spot analysis of TripAdvisor data as support for destination management. Current Issues in Tourism, 23(2), 175-196.

Van Leeuwen, M., Klerks, Y., Bargeman, B., Heslinga, J., & Bastiaansen, M. (2020). Leisure will not be locked down–insights on leisure and COVID-19 from the Netherlands. *World Leisure Journal*, 1-5.

Vu, H. Q., Li, G., Law, R., & Zhang, Y. (2018). Tourist activity analysis by leveraging mobile social media data. Journal of travel research, 57(7), 883-898.

Wang, Y., & Li, B. (2015, November). Sentiment analysis for social media images. In 2015 IEEE international conference on data mining workshop (ICDMW) (pp. 1584-1591). IEEE.

Weber, F., Stettler, J., Priskin, J., Rosenberg-Taufer, B., Ponnapureddy, S., Fux, S., ... & Barth, M. (2017). Tourism destinations under pressure. *Challenges and Innovative Solutions; Lucerne University of Applied Sciences and Arts: Lucerne, Switzerland*.

Wood, S. A., Guerry, A. D., Silver, J. M., & Lacayo, M. (2013). Using social media to quantify nature-based tourism and recreation. Scientific reports, 3(1), 1-7.

Wu, X., Huang, Z., Peng, X., Chen, Y., & Liu, Y. (2018). Building a spatially-embedded network of tourism hotspots from geotagged social media data. IEEE Access, 6, 21945-21955.

Xia, J. C., Zeephongsekul, P., & Arrowsmith, C. (2009). Modelling spatio-temporal movement of tourists using finite Markov chains. *Mathematics and Computers in Simulation*, *79*(5), 1544-1553.

Xiang, Z., & Gretzel, U. (2010). Role of social media in online travel information search. *Tourism management*, *31*(2), 179-188.

10. Appendices

A. Instagram Web Scraper Code

The metros.json file with id’s for each place that is scraped:

```
{
  "Rotterdam": "213226541",
  "Veere": "250640993",
  "Vrouwenpolder": "290002343",
  "Breezand": "301194391",
  "Oostkapelle": "142235",
  "Domburg": "254004334",
  "Zoutelande": "10233915",
  "Aagtekerke": "300413981",
  "Westkapelle": "266104309",
  "Meliskerke": "12381069",
  "Grijpskerke": "15678874",
  "SintLaurens": "264590442"
}
```

To obtain the data:

```
import argparse
import codecs
import json
import logging
import sys
from pathlib import Path

import backoff
import requests
from tqdm import tqdm

import dateToId

BASE_URL =
"https://www.instagram.com/explore/locations/{location_id}/?__a=1&max_id={m
ax_id}"
MAX_TRIES = 11

@backoff.on_exception(wait_gen=backoff.fibo, max_tries=MAX_TRIES,
                      exception=(requests.exceptions.HTTPError,
requests.exceptions.ConnectionError))
def pull_json(location_name, end_cursor):
    URL = BASE_URL.format(location_id=METRO[location_name],
max_id=end_cursor)
    logging.debug(URL)
    r = requests.get(url=URL)
    if r.status_code == 200:
        data = r.json()['graphql']['location']['edge_location_to_media']
        return data
    else:
        raise requests.exceptions.HTTPError


def save_jsonl(data, dst='./'):
```

```python
    """Saves the data to a jsonl file."""
    path = Path(dst)
    assert isinstance(data, list)
    if path.exists():
        f = codecs.open(dst, "ab", 'utf-32')
    else:
        path.parent.mkdir(parents=True, exist_ok=True)
        f = codecs.open(dst, 'wb', 'utf-32')
    f.writelines("%s\n" % json.dumps(s) for s in data)


def scrape(date1, date2, location, restore_cursor=None):
    try:
        maxid1 = dateToId.run(date1)
        maxid2 = dateToId.run(date2)
    except:
        logging.error("error in date format, make sure no following zeroes example: 2020/7/15")
        sys.exit(1)

    end_cursor = maxid1 if restore_cursor is None else restore_cursor
    logging.debug(f"end_cursor: {end_cursor}")

    pbar = tqdm(unit='post')
    while end_cursor and end_cursor > maxid2:
        data = pull_json(location, end_cursor)
        if data:
            filename = f"{PATH}/{location}_{date1.replace('/', '-')}_{date2.replace('/', '-')}"
            save_jsonl(data['edges'], filename + '.jsonl')
            end_cursor = data['page_info']['end_cursor']
            open(f"{filename}_CURSOR.txt", "w").write(str(end_cursor))
            pbar.update(len(data['edges']))
        else:
            logging.warning(f"No data found at end_cursor: {end_cursor}")
    else:
        pbar.close()

def main():
    parser = argparse.ArgumentParser()
    parser.add_argument("--dir", nargs=1, default=['./data'])
    parser.add_argument("--max", nargs=1, required=True)
    parser.add_argument("--min", nargs=1, required=True)
    parser.add_argument("--location", nargs=1, required=True)
    parser.add_argument('--restore-cursor', action='store_true',
default=False)
    parser.add_argument('--log-level', type=int, default=20)

    args = parser.parse_args()
    global METRO
    global PATH
    global log
    PATH = args.dir[0]
    METRO = json.load(open("./metros.json", 'r'))
    logging.basicConfig(level=args.log_level)
    log = logging.getLogger(__name__)

    assert args.location[0] in METRO, 'Location not available in metros.json'

    if args.restore_cursor:
        logging.warning(f"--max {args.max[0]} will be ignored. Scraping as far back as {args.min[0]}")
        filename = f"{PATH}/{args.location[0]}_{args.max[0].replace('/', '-')}_{args.min[0].replace('/', '-')}"
        maxid1 = open(f"{filename}_CURSOR.txt", "r").read()
    else:
        maxid1 = None
        logging.info(f"Scraping between dates {args.min[0]} and {args.max[0]}")

    logging.info(f"PATH: {PATH}")
    scrape(args.max[0], args.min[0], args.location[0], maxid1)


def test():
    global METRO
    global PATH
    global log
    PATH = "./data"
    METRO = json.load(open("./metros.json", 'r'))
    logging.basicConfig(level=10)
    log = logging.getLogger(__name__)
    scrape("2020/07/16", "2020/07/15", "Memphis")



if __name__ == "__main__":
    main()
#
```

The scraping notebook:

```
! pip install backoff
```

```
#deciding path

from pathlib import Path

from google.colab import drive

drive.mount('/content/drive', force_remount=True)

directory = Path('/content/drive/My Drive/Instascrape/')

mypath = directory/'Method_1'

mypath.mkdir(parents=True, exist_ok=True)


min_date  = "2019/11/11"

max_date = "2019/11/19"

location  = "Rotterdam"


! cd "$mypath" && rm -r insta-graphql-scraper/

! cd "$mypath" && git clone https://github.com/BramR123/insta-graphql-
scraper


import os

os.chdir("/content/drive/My Drive/Instascrape/Method_1")


!python scraper.py --dir "$mypath" --max $max_date --min $min_date --location
$location


## re-run from last cursor, run this after getting error, after factory
resetting it.

! cd "$mypath"/insta-graphql-scraper && python scraper.py --restore-cursor
--dir "$mypath" --max $max_date --min $min_date --location $location
```

To convert the data:

```python
import json, codecs, csv, pandas as pd

path = r"C:\Users\bramr\Documents\Workplace\Instascrape\Method_1\Rotterdam_2019-09-01_2019-08-31.jsonl"

datas = []
with open(path, 'r', encoding='utf-32') as f:
    for line in f:
        datas.append(line)
    #datas.append(json.loads(line))

jsons = []
for lines in datas:
    decoded_data=codecs.decode(lines.encode(), 'utf-8-sig')
    data = json.loads(decoded_data)
    jsons.append(data['node'])

keyList = jsons[0].keys()
keyList = list(keyList).extend(['accessibility_caption', 'video_view_count'])

df = pd.DataFrame(jsons)

ids = []
for i in range(len(df['owner'])):
    ids.append(df['owner'][i]['id'])
df = df.assign(ID = ids)
```

```
likes = []
for i in range(len(df['edge_liked_by'])):
    likes.append(df['edge_liked_by'][i]['count'])
df = df.assign(Number_of_likes = likes)

df['datetime'] = pd.to_datetime(df['taken_at_timestamp'], unit="s")

df_for_export = df[['ID', 'Number_of_likes', 'datetime']]

df_for_export.to_csv('Insta_Rotterdam_1_9.csv')
```

B. TripAdvisor Web Scraper Code

To collect the URLs for all restaurants in a certain place:

```
from bs4 import BeautifulSoup as bs
import pandas as pd
import csv
import requests
import time
import re, random, json


page = "https://www.tripadvisor.nl/Restaurants-g652353-Vrouwenpolder_Zeeland_Province.html"


def getAllResults(firstPage, city):
        """ Gets all search results based on the first page. """
        firstPage = firstPage + "#EATERY_LIST_CONTENTS"
        allLinks = [firstPage]
        p = requests.get(firstPage)
        s = bs(p.content, 'html.parser')
        idx = firstPage.index('-%s' % city)
        try:
            div = s.find('div', {'class': 'pageNumbers'})
            children = div.findChildren(recursive=False)
            numberofP = len(children)
        except AttributeError:
            numberofP = 1


        for i in range(1, numberofP):
            print("Creating URL for page "+str(i))
            template = '-oa%s0' % (i*3)
            finalLink = firstPage[:idx] + template + firstPage[idx:]
            allLinks.append(finalLink)
            print(finalLink)
        return allLinks


allthePages = getAllResults(page, "Vrouwenpolder")


def getURLs(searchResultPages):
    """ Takes a list with URLs to all pages of search results and returns a list with links to each restaurant """
    searchResults = {}
    for idx, i in enumerate(searchResultPages):
        r = requests.get(i)
        soup = bs(r.content, 'html.parser')
        a = soup.find(id='EATERY_SEARCH_RESULTS')
        list_items = a.find_all("div", { "data-test" : re.compile(r"\d+_list_item")})
        for b in range(len(list_items)):
            currentidx = idx*30+b
            searchResults[currentidx] = {}
            c = list_items[b].find("a", {"class" : "_15_ydu6b"})
            d = c.get('href')
            e = c.get_text()
            searchResults[currentidx]['URL'] = d
            searchResults[currentidx]['Name'] = re.sub(r'^\d+. ', '', e)
        if i != 0:
            time.sleep(3)

    return searchResults
```

```
alltheURLs = getURLs(allthePages)

with open('URLs_Vrouwenpolder_final.json', 'w') as fp:
    json.dump(alltheURLs, fp)
```

To collect the data for each restaurant:

```
import sys, time
from selenium import webdriver
import selenium
import time, json, re, logging

def main():
    # configure logger
    logfile = 'TAScrape' + time.strftime("%Y%m%d-%H%M%S") + '.log'
    logging.basicConfig(filename=logfile,format='%(asctime)s  %(message)s',
datefmt='%m/%d/%Y %I:%M:%S %p', level=logging.DEBUG)
    logging.getLogger("requests").setLevel(logging.WARNING)
    # default path to file to store data
    path_to_file = "koudekerke.json"

    # open file with links and page numbers
    with open('URLs_Koudekerke_final.json', 'r') as fp:
        restaurant_urls = json.loads(fp.read())

    # default number of scraped pages
    num_page = 10
    # default tripadvisor website of restaurant
    base_url = "https://www.tripadvisor.com"
    all_lang = "?filterLang=ALL"

    # Import the webdriver
    driver = webdriver.Chrome()

    # production code
    # for key in restaurant_urls:

    # test code
    # for key in ['60']:

    for key in restaurant_urls:
        logging.info("Scraping restaurant: " + restaurant_urls[key]["Name"])

        url = base_url + restaurant_urls[key]["URL"] + all_lang
        driver.get(url)
        time.sleep(0.5)
        driver.refresh() # refresh necessary to avoid bug that prevents all reviews from showing
        time.sleep(2.5)
        try:
            driver.find_element_by_xpath(".//button[@class='evidon-banner-acceptbutton']").click()
        except:
            pass
        try:
            img = driver.find_element_by_xpath(".//span[@data-test-target='staticMapSnapshot']/img").get_attribute("src")
            location = re.match(r".*center=(\d{2}.\d*),(\d{1}.\d*)&", img)
            restaurant_urls[key]["lon"] = location.group(2)
            restaurant_urls[key]["lat"] = location.group(1)
        except:
            logging.warning("Whoops, couldn't fetch GMaps image @ URL " + url)
            address = driver.find_element_by_xpath(".//span[@class='_2saB_OSe']").text
            restaurant_urls[key]["address"] = address

        # look for review container and get number of reviews
        try:
            number_of_rev = driver.find_element_by_xpath(".//span[@class='reviews_header_count']").text[1:-1]
            restaurant_urls[key]["Number of reviews"] = number_of_rev
        except:
            restaurant_urls[key]["Number of reviews"] = 0

        # Get the reviews
        try:
            num_page = driver.find_element_by_xpath(".//a[@class='pageNum last ']").text # int(restaurant_urls[key]["numberofPages"])
        except:
            logging.info("Only one page of reviews at restaurant " + restaurant_urls[key]["Name"])
            num_page = 1

        if restaurant_urls[key]["Number of reviews"] != 0:
            restaurant_urls[key]["reviews"] = []

            for i in range(0, int(num_page)-1):
                # expand the reviews
                time.sleep(2)
                try:
                    # click to expand all reviews
                    expand =
driver.find_element_by_xpath("//span[@class='taLnk ulBlueLinks']")#.click()
                    driver.execute_script("arguments[0].click();", expand)
                except:
                    logging.info("Comments not expandable on page " + 'i' +
"on " + restaurant_urls[key]["URL"])
                time.sleep(1)
                container =
driver.find_elements_by_xpath(".//div[@class='review-container']")
                print("Number of reviews on page " + str(i) + ": " +
str(len(container)))

                # for each review...
                for j in range(0,len(container)):
                    review = {}
                    try: # in case there are no reviews
                        review["title"] =
container[j].find_element_by_xpath(".//span[@class='noQuotes']").text
                        review["date"] =
container[j].find_element_by_xpath(".//span[contains(@class,
'ratingDate')]").get_attribute("title")
                        review["rating"] =
container[j].find_element_by_xpath(".//span[contains(@class,
'ui_bubble_rating bubble_')]").get_attribute("class").split("_")[3]
```

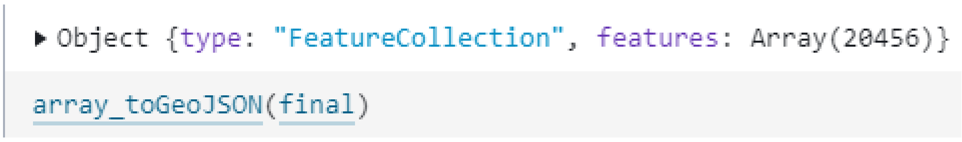
```
                        review["full"] = container[j].find_element_by_xpath(".//p[@class='partial_entry']").text.replace("\n", " ")
                    except:
                        logging.info("No reviews at restaurant number " + key)
                    try:
                        uid_src = container[j].find_element_by_xpath(".//div[@class='memberOverlayLink clickable']").get_attribute("id")
                        matches = re.match(r"^UID_([0-9a-fA-F]+)-SRC_(\d+)", uid_src)
                        review["uid"] = matches.group(1)
                        review["src"] = matches.group(2)
                    except:
                        logging.info("Profile unavailable at restaurant number " + key)
                    restaurant_urls[key]["reviews"].append(review)

                try:
                    driver.find_element_by_xpath('.//a[@class="nav next ui_button primary"]').click()
                except:
                    logging.info("Page not changeable at " + restaurant_urls[key]["Name"] + ", page " + str(i))

    with open(path_to_file, 'w') as fp:
        json.dump(restaurant_urls, fp)

    driver.close()
    logging.info("Scraping successful! Shutting down...")
```

```
if __name__ == '__main__':
    main()
```

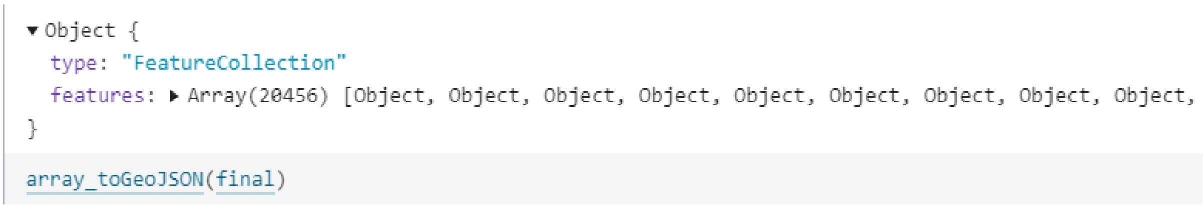
Data conversion and flattening in Observable notebook: (JS)

```
data2 = await fetch(

"https://dl.dropboxusercontent.com/s/5jwr86e5vmj79fn/zoutelande.json?dl=0"
).then(data => data.json())


data3 = Object.keys(data2).map(i => ({
  ...data2[i],
  reviews: { ...data2[i].reviews }
}))


final = Object.keys(data3)
  .map(i => data3[i])
  .map(item =>
    Object.keys(item.reviews).map(k => ({
      ...item.reviews[k],
      rating: Number(item.reviews[k].rating[0]),
      lon: item.lon,
      lat: item.lat,
      r_name: item.Name,
      r_numberReviews: item["Number of reviews"],
      r_Rating: Object.keys(item.reviews)
        .map(review => Number(item.reviews[review].rating) / 10)
        .reduce(average, 0),
      dateTime: d3.timeParse("%B %d, %Y")(item.reviews[k].date),
      weekday:
        days[
          getWeekDay(
            d3
              .timeParse("%B %d, %Y")(item.reviews[k].date)
              .getDay()
          )
        ]
    }))
  )
  .flat()

array_toGeoJSON(final)
```

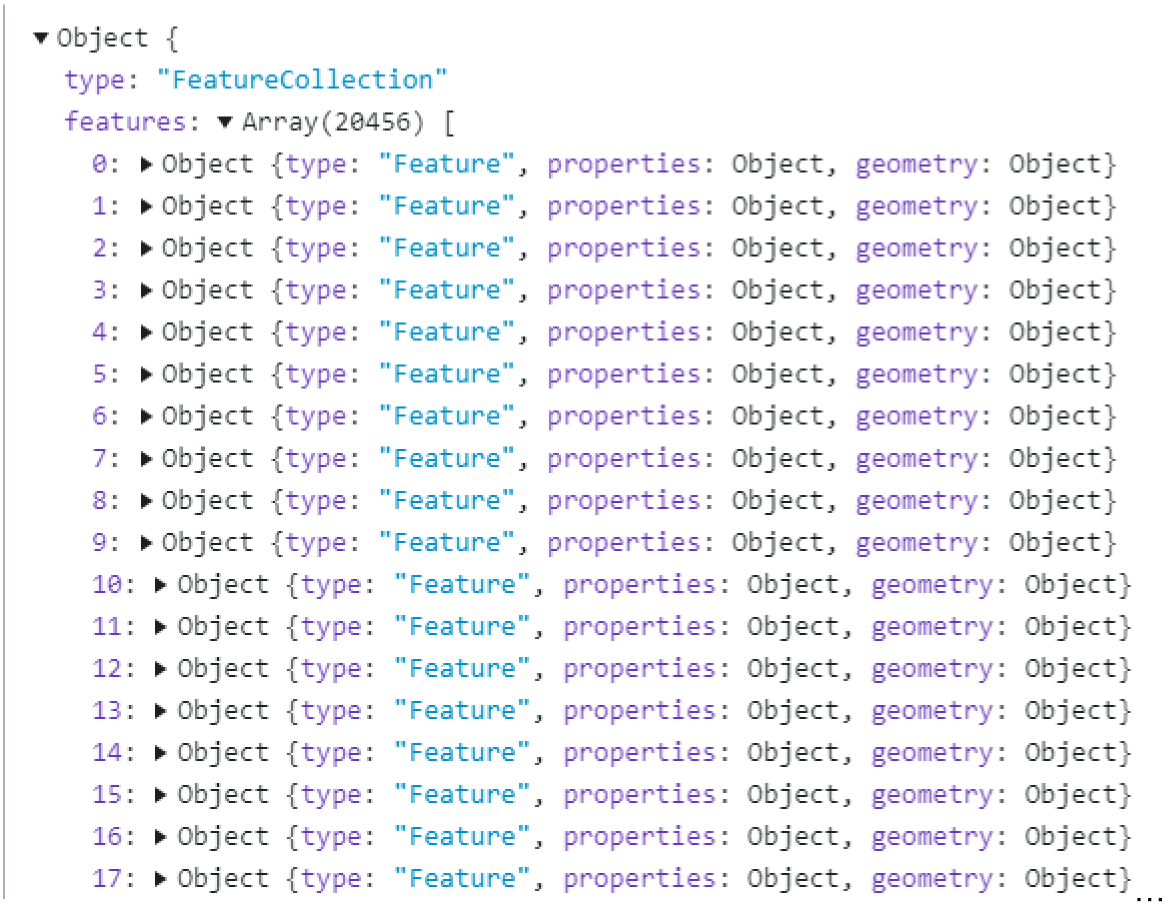
Observable snippets with the JSON structure:



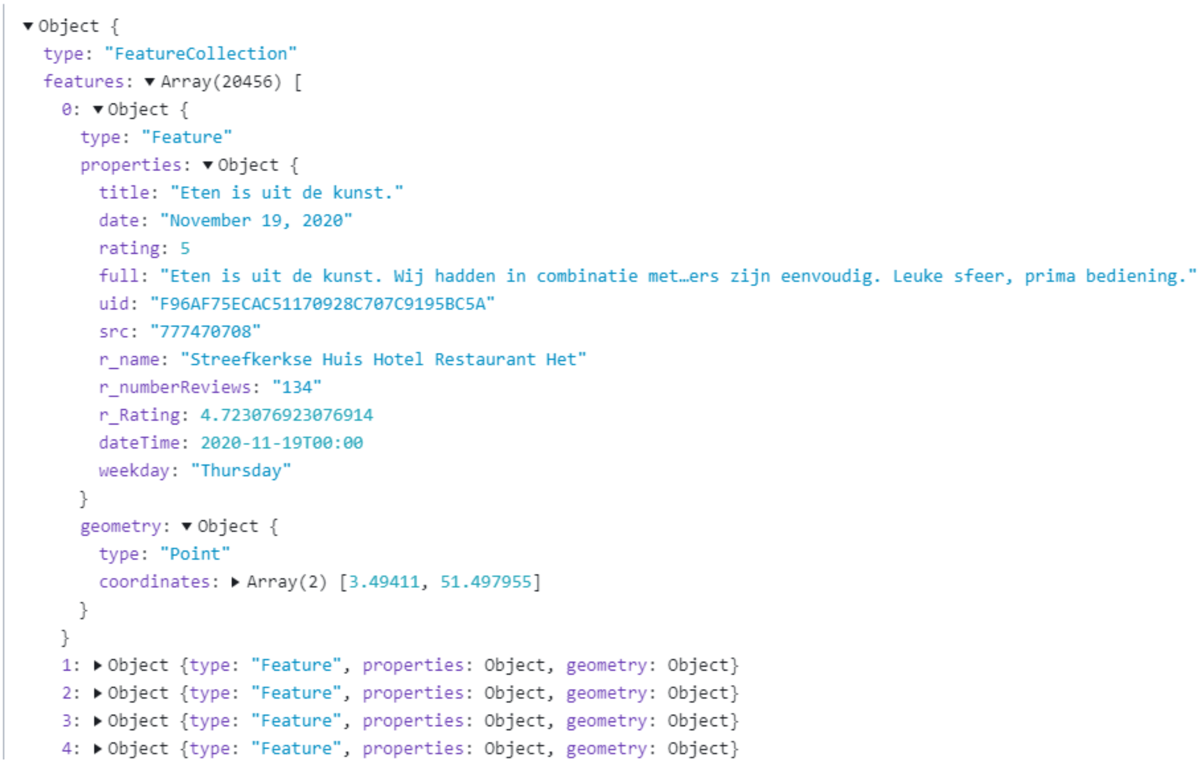
```
▶ Object {type: "FeatureCollection", features: Array(20456)}
array_toGeoJSON(final)
```



```
▼ Object {
  type: "FeatureCollection"
  features: ▶ Array(20456) [Object, Object, Object, Object, Object, Object, Object, Object, Object,
}
array_toGeoJSON(final)
```

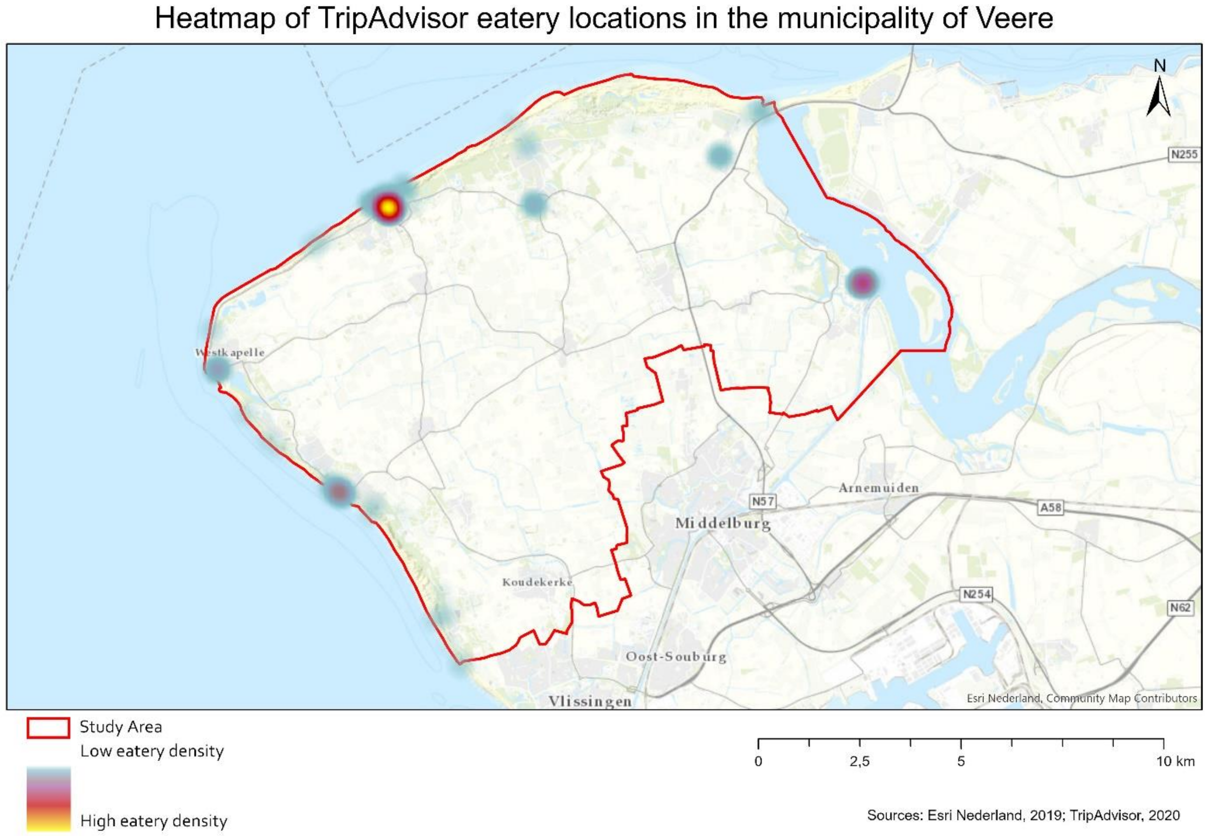


```
▼Object {
  type: "FeatureCollection"
  features: ▼Array(20456) [
    0: ▶Object {type: "Feature", properties: Object, geometry: Object}
    1: ▶Object {type: "Feature", properties: Object, geometry: Object}
    2: ▶Object {type: "Feature", properties: Object, geometry: Object}
    3: ▶Object {type: "Feature", properties: Object, geometry: Object}
    4: ▶Object {type: "Feature", properties: Object, geometry: Object}
    5: ▶Object {type: "Feature", properties: Object, geometry: Object}
    6: ▶Object {type: "Feature", properties: Object, geometry: Object}
    7: ▶Object {type: "Feature", properties: Object, geometry: Object}
    8: ▶Object {type: "Feature", properties: Object, geometry: Object}
    9: ▶Object {type: "Feature", properties: Object, geometry: Object}
    10: ▶Object {type: "Feature", properties: Object, geometry: Object}
    11: ▶Object {type: "Feature", properties: Object, geometry: Object}
    12: ▶Object {type: "Feature", properties: Object, geometry: Object}
    13: ▶Object {type: "Feature", properties: Object, geometry: Object}
    14: ▶Object {type: "Feature", properties: Object, geometry: Object}
    15: ▶Object {type: "Feature", properties: Object, geometry: Object}
    16: ▶Object {type: "Feature", properties: Object, geometry: Object}
    17: ▶Object {type: "Feature", properties: Object, geometry: Object}…
```



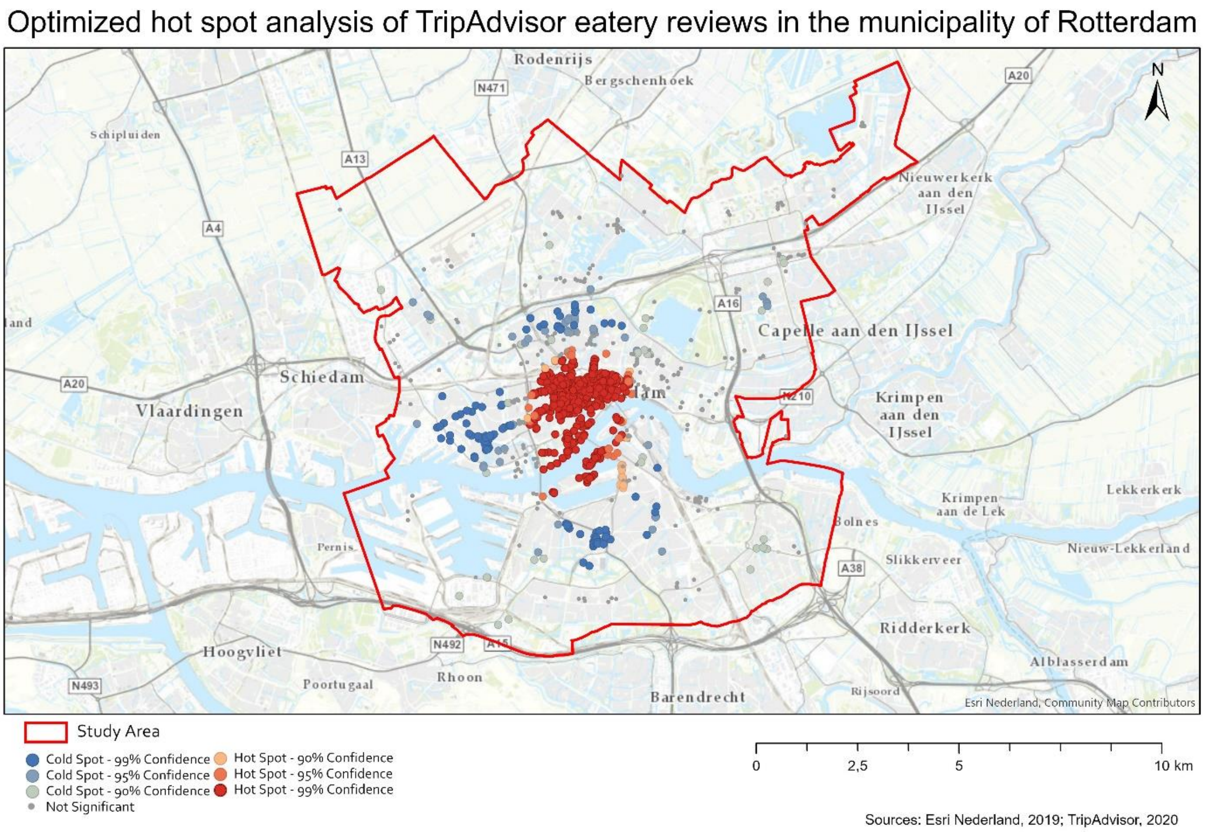
```
▼Object {
  type: "FeatureCollection"
  features: ▼Array(20456) [
    0: ▼Object {
      type: "Feature"
      properties: ▼Object {
        title: "Eten is uit de kunst."
        date: "November 19, 2020"
        rating: 5
        full: "Eten is uit de kunst. Wij hadden in combinatie met…ers zijn eenvoudig. Leuke sfeer, prima bediening."
        uid: "F96AF75ECAC51170928C707C9195BC5A"
        src: "777470708"
        r_name: "Streefkerkse Huis Hotel Restaurant Het"
        r_numberReviews: "134"
        r_Rating: 4.723076923076914
        dateTime: 2020-11-19T00:00
        weekday: "Thursday"
      }
      geometry: ▼Object {
        type: "Point"
        coordinates: ▶Array(2) [3.49411, 51.497955]
      }
    }
    1: ▶Object {type: "Feature", properties: Object, geometry: Object}
    2: ▶Object {type: "Feature", properties: Object, geometry: Object}
    3: ▶Object {type: "Feature", properties: Object, geometry: Object}
    4: ▶Object {type: "Feature", properties: Object, geometry: Object}
```

C. Extra figures



Sources: Esri Nederland, 2019; TripAdvisor, 2020

Figure D.1: Heatmap of TripAdvisor eatery locations in the municipality of Veere.



Optimized hot spot analysis of TripAdvisor eatery reviews in the municipality of Rotterdam

Sources: Esri Nederland, 2019; TripAdvisor, 2020

Figure D.2: Optimized hot spot analysis of TripAdvisor eatery reviews in the municipality of Rotterdam.

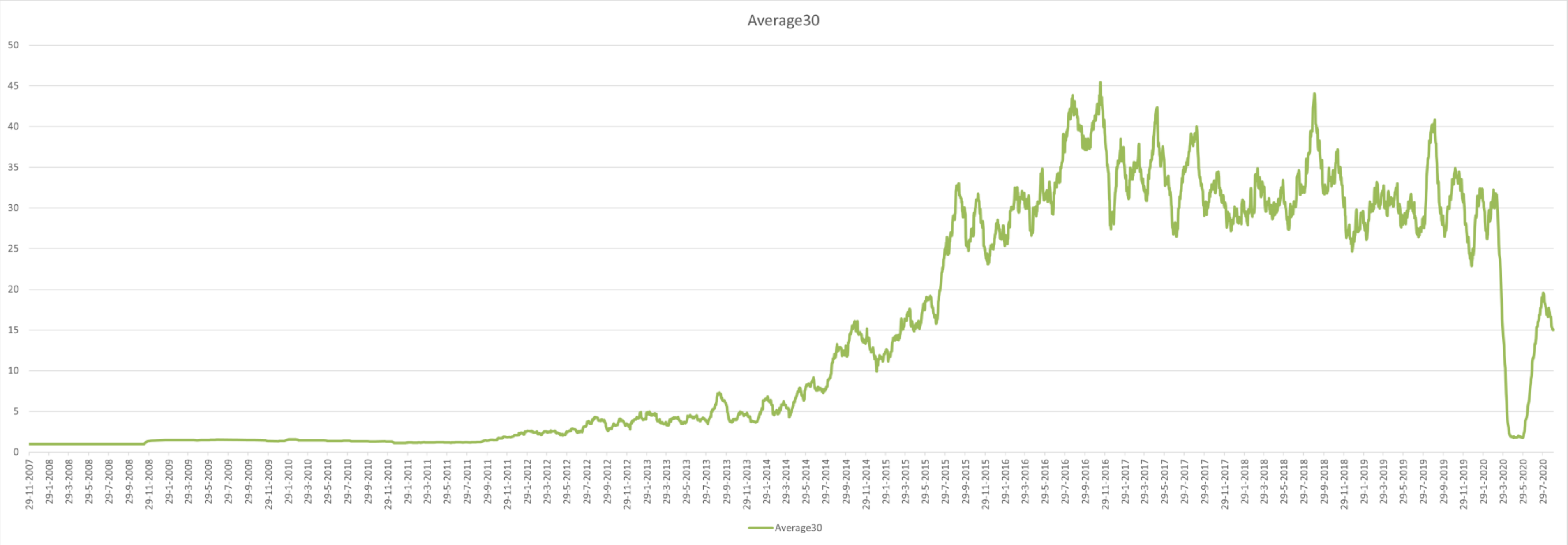


Figure D.3: Daily number of TripAdvisor reviews in Rotterdam between 29-11-2007 and 31-08-2020, displayed as a 30-day average