Utilizing volunteered geographical information for the benefit of city planners and urban science: Case of Rotterdam

Deniz Özağaç
Utilizing volunteered geographical information for the benefit of city planners and urban science: Case of Rotterdam

Graduation Thesis for

MASTER OF SCIENCE

in Complex System Engineering and Management

TUDelft

Faculty of Technology, Policy and Management

by

Deniz Özağac

Student number: 4627482

To be publicly defended on November 26, 2018

Graduation Committee

First Supervisor : D.Phil. Scott W. Cunningham, Multi-Actor Systems
Second Supervisor : Dr. Laurens Rook; Values, Technology and Innovation
It is quite fascinating to me how this thesis came to existence from great uncertainty. From not knowing the topic of it before it’s beginning to not being sure where it was headed at the early months, the feeling of uncertainty was with me the entire way. But I took a lesson from this journey. This entire process made me learn how to be in peace with uncertainty and how to trust the process. The process here meaning having the discipline to work at it. Of course, some days are slower than the others, but as long as your thesis at the end of the day looks more complete than the version in the morning, you will get there. I think this basic formula can be applied to most things at life: Just make sure that you finish the day being a better person than the one who started it.

There are numerous people who helped me on this journey. First, I would like to thank my first supervisor Scott Cunningham for meeting me weekly and unclogging my writer’s blocks with great discussion and suggestions. Similarly, I want to say thank you to my second supervisor Laurens Rook with the inspiration he provided to me about using social media data and teaching the most enjoyable course I had in my master’s education, high technology marketing.

I would also like to thank my friends here in the Netherlands who shared this journey with me and to the ones who are at far places, who accompanied me in long calls while I was on my bike. And finally, I want to say thank you to my family who helped me become a person who has the patience and discipline to write a hundred-page thesis.

I also thank you reader, for your attention and interest. I hope you leave this reading session with more knowledge than the person who started it!

Deniz Özağaç
Delft, November 2018
Summary

It is expected that 68% of the world’s total population will be living in cities by 2050, which is around 55% as of 2018 (World urbanization prospects, 2018). With the increasing population density of cities all around the world, it is no doubt that the systems and services that help operate these cities will get more challenging to manage. Therefore, it is necessary for domains of data science and urban science to have great synergy between each other to solve these challenges. This thesis project focuses on this bridge between these two domains. Supported by the literature review the focus is on the use of social media data with geolocational components, named as volunteered geographic information (VGI, since users almost always have the option to hide their location on social media yet choose to put out to the internet). Therefore, the initial question asked in this study is:

“How can one create value for the urban science domain with smart and effective use of VGI data?”

Consequently, with the aim of generating tangible results and demonstrating how VGI can be used to generate insights for a city, the City of Rotterdam is selected as the subject of this thesis. The goal of this study can, therefore, be stated as:

“Performing a case study on Rotterdam where social media data with geolocational components (VGI) is brought under the spotlight to assess its usefulness and actionability for the benefit of urban scientists and city planners.”

Leading to the model-analysis part of the document; a systematic comparison of Twitter, Foursquare and Instagram as social media data sources for research purposes is performed in chapter 4. The comparison in this section results in Twitter to be the selected as the best social media data source for urban research context in 2018, which is also selected as the data source for this study. Following that, data collection and cleaning are explained in a way that is empathetic to an urban scientist rather than a technical data person. This Twitter dataset which contains data in the form of Twitter activity of Rotterdam in April 2018 is also made available to the public.

In order to move a step further than simply collecting and visualizing the data and attaining an actionable statistical summary of the data, a model-based analysis is performed. A set of generative models are created after the model selection process in chapter 6, where various potential model types are compared to each other for fitness to the goal and context at hand. As a result of this model selection process, simple Markov models and mixture models are selected to perform analysis on the Twitter data. To my knowledge, this is a unique approach since this is the only example of this data-model combination in the literature. As for the results, Markov models and resulting transition matrices have the potential to help researchers understand mobility patterns within cities by providing a probabilistic explanation for behaviors. While simple Markov models are too abstracted from reality, they can give an idea of where an individual is likely to go looking at their current position in the city. Similarly, mixture models generate “chains” which are simply behavioral classification groups for the population in the dataset. With chain weights, it is possible to probabilistically assign new individuals to one of the existing chains, hence the generative nature of the models.
The findings of this thesis can be further improved by attempting to use more advanced model types such as hidden Markov models (HMM), mixture of Markov models and mixture of hidden Markov models. Furthermore, use of various other data sources has the potential to yield better results as social media data is eventually limited by the behavior of its users and may not reflect the situation of the population as a whole. Some of these alternative data sources are mentioned and recommended in the future research discussion in chapter 9.5.

To summarize, this thesis project is forming a bridge between urban science and data science, reaching out from the side of data science of the gap. For doing so, the societal relevance of using data to improve urban lives is discussed. It is followed by evaluation and systematic comparison of selected social media data sources regarding their usefulness for research in 2018. Then, Twitter data is collected and made available for the public. Following that model analysis is performed with simple Markov models and mixture models. Throughout this thesis, I tried communicating the results and findings in a way that would be easier for individuals with a non-technical background to find interesting. I did so by often leading with relatable questions. After all, the results of this thesis are aimed at city planners, urban scientist, decision-makers and researchers who seek to utilize publicly available social media data for the purpose of improving the cities and the lives of their residents.
Abstract

One of the main challenges of the recently popular data science field is establishing a common ground of understanding between technical methods and domain knowledge. Making smart and effective use of data is just as important for public organizations as it is for private organizations as our cities and their problems get more complex with increasing populations and their demands. Addressing this very issue, this thesis focuses on the connection between data science and urban science, mainly on how freely available social media data with geolocational components which is called volunteered geographic information (VGI) here, can be utilized for the benefit of urban science. For this purpose, 3 popular VGI sources; Foursquare, Instagram and Twitter API’s are inspected and compared for their usefulness for data driven urban research. At the first sections of this thesis, a literature review followed by a discussion part is presented about how the smart and effective use of data is beneficial for cities. Then, an event detection application is conceptualized which is used for deriving data and model requirements. This thesis paper sets itself apart by taking a semi-design-oriented approach with real social media data and testing the usefulness of two modeling styles -simple Markov and mixture models- that does not have prior representation in the literature in conjunction with VGI.

Key words: Social media data, Twitter data, urban science, volunteered geographic information, city planning, Markov models, mixture models
# Table of Contents

List of Figures vii  
List of Tables viii  
List of Acronyms viii  

1 Introduction 1  
1.1 Outline of the thesis report 2  

2 Literature Review 5  
2.1 Search technique 5  
2.2 Review of literature 5  

3 Methodology 15  
3.1 Approach 15  
3.2 Societal relevance 17  
3.3 User scenario 19  
3.3.1 User requirements 22  
3.3.2 Technical requirements 22  

4 Evaluation of Data Sources & Data Collection 25  
4.1 API and data collection 26  
4.1.1 Foursquare 26  
4.1.2 Instagram 27  
4.1.3 Twitter 28  
4.1.4 Decision for the data source 29  
4.2 Concerns regarding the data 30  

5 Initial Data Exploration 33  
5.1 Understanding the dataset 34  
5.2 Inspecting missing data 34  
5.3 Exploratory analysis 34  
5.3.1 Location of the tweets 34  
5.3.2 User data 36  
5.3.3 Language 38  
5.3.4 Time series analysis 38  

6 Model Selection 43  
6.1 Introduction to simple Markov model 43  
6.2 Introduction to mixture models 44  
6.3 Introduction to hidden Markov models 45  
6.4 Introduction to probabilistic PCA 46  
6.5 Model comparison and selection 47  

7 Data Preprocessing and Analysis 49  
7.1 Preprocessing the data 49  
7.2 Analysis 53  
7.2.1 Simple Markov model 53  
7.2.1.1 Expectation Maximization Algorithm 55  
7.2.1.2 Assigning new trajectories/ users to the chains 57  
7.2.2 Mixture model 58  
7.2.2.1 Random walks and Crowd management 59
## List of Figures

Figure 2.1: Basic model for usage of VGI ................................................................. 7  
Figure 2.2: The municipal borders (black) and livelihood clusters comparison .......... 8  
Figure 2.3: Neighborhood recommendations for New York SoHo region ............... 9  
Figure 2.4: Weekday check-in behavior ................................................................. 10  
Figure 2.5: Only 6 human accounts (blue) in top 30 most active tweeters ............... 11  
Figure 2.6: CRISP DM methodology ................................................................. 12  
Figure 2.7: Flow diagram of the approach ................................................................. 13  
Figure 2.8: Derivation of evaluation criteria from the requirements .......................... 14  
Figure 2.9: Weekday check-in behavior (Noulas et al., 2011) ................................. 15  
Figure 2.10: Estimation age distribution of Twitter users in UK............................. 16  
Figure 2.11: Number of tweets by day of April .................................................... 17  
Figure 2.12: Intended bounding box for API data collection .................................... 18  
Figure 2.13: Visualization of collected tweets as red dots ........................................ 19  
Figure 2.14: Geolocations of Tweets from Rotterdam ............................................ 20  
Figure 2.15: Number of tweets by day of April .................................................... 21  
Figure 2.16: Number of tweets by day of the week ............................................. 22  
Figure 2.17: Number of tweets by hour of the day ............................................. 23  
Figure 2.18: Tweet densities per part of the day on the City of Rotterdam ............... 24  
Figure 2.19: Bounding box used for data processing ............................................. 25  
Figure 2.20: Hotspots of Rotterdam (n=50) ....................................................... 26  
Figure 2.21: Hotspots of Rotterdam ................................................................. 27  
Figure 2.22: Hotspots of Rotterdam (n=9) ............................................................ 28  
Figure 2.23: Combination of mixtures with n=9 hotspots ...................................... 29  
Figure 2.24: Hotspot locations of the mixture model ............................................ 30  
Figure 2.25: Log-likelihood scores starting from a random guess ............................ 31  
Figure 2.26: The main trajectory matrix (top) and the 3 chains that form it (below) .... 32  
Figure 2.27: The transition matrix between hotspots ............................................ 33  
Figure 2.28: Example random walk with 5 steps ............................................. 34  
Figure 2.29: The transition matrix between hotspots ............................................ 35  
Figure 2.30: Top 3 most places that the users retweet from .................................... 36  
Figure 2.31: Top 5 most attractive hotspots of Rotterdam ..................................... 37  
Figure 2.32: Hotspots of Rotterdam according to the Mixture model ....................... 38  
Figure 2.33: Chain 1 accompanied by a map that shows hotspot locations ............... 39  
Figure 2.34: Chain 2 accompanied by a map that shows hotspot locations .......... 40  
Figure 2.35: Chain 3 accompanied by a map that shows hotspot locations .......... 41  
Figure 2.36: Hotspots of Rotterdam (n=22) ....................................................... 42  
Figure 2.37: Hotspots of Rotterdam (n=92) ....................................................... 43  
Figure 2.38: White and black box testing ............................................................... 44  
Figure 2.39: Only 6 human accounts (blue) in top 30 most active tweeters ............... 45  
Figure 2.40: Non-Stationary and non-human top tweeters and their trajectories ......... 46  
Figure 2.41: Google popular time graphics ............................................................. 47
List of Tables

Table 2.1: Search phrases ............................................................................................................. 5
Table 2.2: Matching GSMD to the current understanding of tourist flows ................................. 10
Table 3.1: Tangible use cases of data science in a city context .................................................. 18
Table 3.2: User requirements ........................................................................................................ 22
Table 3.3: Data requirements ........................................................................................................ 22
Table 3.4: Model requirements ..................................................................................................... 23
Table 4.1: Summary of data source evaluation ........................................................................... 29
Table 5.1: Most popular Twitter accounts in Rotterdam ............................................................ 37
Table 5.2: Accounts that tweeted the most in Rotterdam ............................................................ 37
Table 5.3: Top 5 used languages in the dataset ............................................................................. 38
Table 6.1: Upper-left piece of the transition matrix ...................................................................... 43
Table 6.2: Model comparison with respect to model requirements ............................................. 47
Table 8.1: Potential attraction points and nearby buildings of hotspots .................................... 67
Table 8.2: Classification of Hotspots according to satellite view .............................................. 67
Table 9.1: Observed trajectory lengths and corresponding frequencies ..................................... 74

List of Acronyms

VGI : Volunteered geographical information
CoSEM : Complex system engineering and management
API : Application programming interface
CRISP-DM : Cross industry standard process for data mining
PCA : Principal component analysis
HMM : Hidden Markov model
EM : Expectation maximization
1 Introduction

The emergence of data science as a popular field had an impact on the most aspects of our lives. Numerous major fields like engineering, psychology and other social sciences have benefited from emerging improvements in the data science field. Just like many other fields, urban and civic management have major potential benefits that can be attained with the smart and effective use of data. This potential mainly lies in improving decision making processes by making them data-driven: for the decision makers, understanding the situation of the physical infrastructure, the political environment and their interaction with the population is the key to effective and satisfactory management. The problems can potentially be detected, addressed and solved more rapidly and accurately if the cause and the result of the events are understood. Therefore, having a solid understanding of to what extent data science can be utilized in a field is crucial for staying on the cutting edge of innovation.

Nevertheless, with the rise of data science’s popularity, a problem of not knowing what to do with this new “tool” has emerged. For the context of urban planning, for example, questions arise such as: how can the tool that is data science can be made use of so that it provides the most value for urban scientists? How would a researcher know what methods to approach their data with, while avoiding choice paralysis in front of the large pool of possible answers? In order to address this problem, this thesis project brings the potential use of data science for the benefit of city planners, urban scientists, municipalities and policymakers under the spotlight. Consequently, the purpose of this thesis is to help to forming a bridge between the data science and urban science domains, initiated from the data science side of the gap. However, since a bridge is best completed in the middle, I put myself in the shoes of an urban scientist and aim to generate actionable and meaningful results for the decision makers. Finally, I refine my scope and focus on a specific type of data, volunteered geographical information (VGI) to be able to extensively investigate a certain topic within the time constraints of a thesis project. The expected result of this thesis is to provide answers to questions such as “How can one create value for the urban science domain with smart and effective use of VGI data?”, “What type of applications can be built with VGI?”, “What are the challenges that lie within using VGI’s for analysis?”. These answers are to be provided by data science, for urban scientists.

To address the not-knowing-what-to-do problem mentioned above, a structural way to define the application fields is found in the literature. Naaman (2011) summarizes some of the approaches that are useful for generating value for cities by utilizing social media data with a geolocational component. The application fields are based on better modeling of geographic areas, improving the understanding of city dynamics and how cities are used by individuals and communities. Moreover, the paper proposes a basic framework to divide these application fields, which this thesis project uses as a base to build upon. This framework divides the application fields into four main categories. These are: boundary definition and detections (District identification), computation of attractions (Landmarks identification), derivation and recommendation of paths (Mobility flow detection), evaluation of activities, interests and temporal trends.

Within this context, some key questions are where and how to get the data, how accessible the data sources are and what does the data contain. Each of these questions on their own
may form a respectable learning barrier in some fields since finding a useful and accessible data source might be a challenging task. It could be the case that most useful sources of data are privately owned or expensively paywalled. However, when it comes to the field of interest of this project - city planning and urban science - there exists a prominent data source which is accessible by everyone: Social media. Social media platforms collect data from their users in various formats and types and this data is often available for individuals without any special access requirements. A strong point of social media data is because the data is collected in the context of a service that is provided, geolocational data is often combined with other metrics such as text filled with sentiment and user’s demographic information. This is because the consumption and generation of the information are almost always structured around a functionality with online social networks (Naaman, 2011). This combination of information and being able to locate the exact source of the information is what makes the social media data a promising asset for the domain of urban science.

There are numerous ways to generate value out of social media data with a geolocational component. For example, with enough data entries, this data can be utilized to generate clusters on the city map that are grouped by various attributes such as the mentioned sentiment, demographic and context. Comparison of these clusters to the existing neighborhood delineations of a city can generate valuable information for city planners and urban scientists. Neighborhood types can be identified by dominant age group, nationality, economic class, ideas can be attributed to neighborhoods regarding various political or social phenomenon, existing neighborhood delineations can be improved, and the information may aid the further planning of the city. Altogether, just with this example, geolocational social network data has the potential to help city planners to divide a city in a more effective way while helping the decision makers to understand the neighborhoods and their residents to a better extent, which will result in better and more responsive policies.

1.1 Outline of the thesis report

The following chapter 2 is a literature review focusing on relevant prior papers and projects. The literature review starts by dividing the use case of social media data with geolocational components into categories and then further leans into these categories. The chapter is finalized by identifying the knowledge gap and presenting the main research question.

Chapter 3 starts by explaining the approach of the thesis project and presents a flowchart that is inspired by the CRISP-DM method. This is followed by a brief chapter where the use cases and possibilities of data driven problem solving are explored in a city context. The first step of the flowchart - the user scenario - is also presented at Chapter 3, which is also used for defining the user and technical requirements that serve as choice guidelines for the data source and model selection later.

Following that is chapter 4, which explains the process of data collection through the social media service API services. The compatibility of the data requirements with the acquired data will be discussed for each data source and advantages/disadvantages of the data sources will be mentioned. Next, chapter 5 introduces the datasets and guides the reader through its contents. Later in chapter 5, the data preparation & cleaning steps will be briefly touched, and the final cleaned dataset will be presented. The clean dataset and the Jupyter notebook for
collecting similar data for further research are also made available for the public use. Once the clean dataset is acquired, an initial exploration of the dataset will be performed. This aims to help the reader to be familiar with the contents of the data. Interesting results will be visualized and potential research questions that are generated from these will be briefly discussed. Once the exploration of the data is done and a solid understanding of the data is acquired, the paper proceeds into model selection in chapter 6, where the model requirements that were generated earlier are the decision criteria. The selection results in choosing a simple Markov model and a mixture model to run the analysis with. Later in chapter 7, the first part is dedicated to getting the previously cleaned data ready to be used as model input. This is followed by the second part where the models are built and executed. The results of the models are also briefly presented here and the discussion about the results are in chapter 8. Once the model section is over, there is an evaluation and validation chapter which discusses the shortcomings of the data, the modeling approach and potential improvements. Finally, urban theories of Milgram are re-visited with the learnings that were acquired from the thesis project in chapter 10. The thesis project is concluded in chapter 11.
2 Literature Review

This section starts by explaining how the literature is scanned by explaining the search phrases are used. The literature review itself focuses on the use cases of social media data for the domain of urban science. The structure for categorizing the use cases is also presented in the upcoming part. At the end of this literature review, a gap in the literature is addressed and the main research goal is formulated.

2.1 Search technique

The framework mentioned in the introduction is also used to create the structure of this literature review. As a reminder, the framework divides the application fields into four main categories: **boundary definition and detections** (District identification), **computation of attractions** (Landmarks identification), **derivation and recommendation of paths** (Mobility flow detection), and **evaluation of activities, interests and temporal trends**.

To inspect each of these application fields, the literature is scanned using pairs of keywords given below. The main phrases are used to ensure the type of data used the found papers would be relevant to the approach of this thesis. These main phrases are queried in combination with each of the secondary phrases, which define the use cases for the papers.

<table>
<thead>
<tr>
<th>Main phrases</th>
<th>Secondary phrases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social media data urban</td>
<td>Neighborhood delineation</td>
</tr>
<tr>
<td>Social media data city</td>
<td>District identification</td>
</tr>
<tr>
<td></td>
<td>Landmark identification</td>
</tr>
<tr>
<td></td>
<td>Mobility flow detection</td>
</tr>
</tbody>
</table>

2.2 Review of literature

As of 2018, 55% of the world’s total population lives in cities or urban areas. In 1950, only 30% of the population lived in rural areas. Since then, this percentage has been on the rise and it is expected to reach 68% by 2050 (World urbanization prospects, 2018). Cities are appealing to populations because they provide simulation senses and opportunities. In other words, cities make choices available for the individual. Compared to less densely populated areas, there is an exponential increase in possibilities of communication with other individuals in great cities (Milgram, 1970). Cities also have district atmospheres and specializations. Paris, for example, is a lot more attractive for individuals who pursue a career in art & fashion that Chicago. Individuals and organizations form these specializations together, the relationship has a reciprocal nature: successful organizations attract the skillful individuals and said companies desire to be close to the skill pool.

Stanley Milgram, one of the most prominent social psychologists approached the cities from a perspective of human communication and social networks. He even stated that humans were a part of a communicative web, in the 1970’s before the emergence of the internet.
(González-Bailón, 2013). Years before his conceptualization became the reality of our lives, he was able to perceive societal relations as a web. However, when Milgram was interested in the topic, the so-called webs and networks and their structures were intangible and not efficiently measurable since the internet was yet to exist. Later, with the emergence of the internet, researchers had a medium in which they could feasibly test the theories of Milgram about social communication and cities.

Milgram states that the individual’s perception of a city is directly affected by the nature of their purpose of stay. A tourist, a resident and a newcomer are bound to have different ideas about the same city. This idea of individuals having different ideas about the city can be extended to say that different lifestyles, background and demographics impact the way individuals perceive a city. Milgram called these cognitive maps of the city and believed that these maps had the potential to reveal valuable information about how a city is perceived by different social classes and groups if generated in a population level. Therefore, they have the potential to help design better policies and interventions that are uniquely formed for each city (Milgram, 1977). In our age, past the emergence of the internet, social media and mobile devices that provide an endless stream of data, it is possible to collect the information that Milgram conceptualized in the form of cognitive maps of the city.

The ubiquitous existence of mobile devices and the streaming data coming with it has changed our lives in many fronts. Improving the city planning was not one of the main goals of the mobile device developers, however, social network services such as Twitter, Facebook and Foursquare have generated and collected a vast amount of user data and made it available to the public. The domain of urban science also got its share of benefits from this new data source. Thanks to the GPS’s that are built into almost every mobile device, the provided data is often paired with the exact geographical data of the action of the user as well. While systematic analysis and a solid understanding of the collective value of this data may provide us with greater societal awareness of our cities, the discipline of urban planning may also greatly benefit from this understanding (Zhang et al., 2013).

Since the explosion of social media in 2004-2005. Many researchers have performed various studies to make use of big scale social media data from platforms and social networking services such as Flickr, Twitter, YouTube, Instagram and many other (Hochman & Manovich, 2013). Among the available data types from these social networks, Geotagged data such as from Twitter and Foursquare contains vast amounts of up-to-date or even real-time data for most locations worldwide. This combined with the constant availability, being freely accessible and the high granularity of the data leads to social media data having quite the value and the potential for studies in various fields and use cases (Chua et al., 2016). Goodchild coined the term volunteered geographical information (VGI) for social media data with geolocational component and the term is used in the rest of this paper to refer to social media data that contains geographical information. The term VGI is often used in the context of geological information and mapping communities, however, Goodchild’s description mentions it as data being created on various online sites to “satisfy a variety of needs within the government, industry and social networking communities.” (Coleman et al., 2009). While social media data is not specifically volunteered by users for the benefit of city planning, urban science and geographical understanding; the use of it fits the description in the context of this thesis project.
Dividing regions into neighborhoods has always been a challenge for the domain of urban social sciences and city planners (Arribas-Bel, 2014). Studies such as “Measuring spatial dynamics in metropolitan areas” (Rey et al., 2011) has been found which address the neighborhood delineation challenges. In this context, utilizing social network data may enable researchers to evaluate if the neighborhood delineation is in harmony with the social life of the city. A quite relevant example of using social media data to generate insights for cities is the “The livehoods project” by Cranshaw et al., which uses 18 million check-ins collected (scraped) from the Foursquare & Twitter and applies various machine learning techniques to generate location-based clusters of venues called “Livehoods” of a city. The term livehood (Not livelihood) refers to a subset of a city which contains similar residents. To clarify, people who go to the venues in this region also go to the other venues in the region quite often (Cranshaw et al., 2012). The study also compares these derived clusters to the neighborhoods on the map of the city as seen in figure 2.2. These visualizations have the potential to improve the neighborhood delineations of a city by revealing rooms for improvements and remapping. The idea of livehoods has not been widely accepted as a literature term, however, a great potential is seen in their approach.
Another relevant study with a similar approach to the delineation problem is “Hoodsquare: Modeling and recommending neighborhoods in location-based social networks” (Zhang et al., 2013). Their approach stands on the idea that the data gathered from location-based social networks can be aggregated to model neighborhoods within urban spaces. They state that many services can benefit from the knowledge about neighborhoods such as recommender systems which can define regions in a city and advise visits to various venues. A mentioned example of such is the Airbnb Neighborhoods project where the aggregated spatial data from Airbnb is used to provide insights to travelers about neighborhoods within their travel distances (“Airbnb Neighborhoods - Your Local Travel Guide). Similar to the livehood project, Zhang et al. derives data-based neighborhoods and presents them in visualizations such as in figure 2.3.
Another popular use for VGI is mobility flow detection within urban and touristic areas. This area of use is similar to how the traffic flow models emerged from the need for identifying congestion factors in transport infrastructures (Nagatani, 2002). In a similar way, understanding the flow patterns of urban areas is important for designing effective urban areas and policies (Sykora & Mulicek, 2014). Although the trajectory of an individual may not generate insights, once aggregated, the trajectories often reveal underlying patterns (Liu et al., 2014). Having a deep understanding of the travel routes is a prerequisite for designing effective policies that prevent capacity overload on the infrastructure (Prideaux, 2000). On the other side of the coin, the quality of the travel experience can also be improved for the tourists by solving the same issue. Moreover, the tourist attractions may be adapted to the needs and preferences of various tourist demographics (Lew & McKercher, 2006).

Chua et al., 2016 has a study with depth analysis regarding the use of VGI for understanding tourist flows. The paper focuses on Cilento region of Italy and has multiple perspectives where the social media data is compared to other data types available for the tourism industry to collect. The comparison of various data sources is reflected in tables and the requirements for data that is necessary for tourist flow analysis is aggregated in another table. Twitter is selected as the main VGI source because it provides freely accessible mechanisms (API's) that allows the researchers to focus on a specific time window and area. The paper also explains how tourist demographic classification is made and the methods that are used for flow analysis are explained in depth. In summary, the study responds to 3 main questions: “What are the meaningful tourist profiles in the region?”, “What are the valuable patterns of tourist flows in the region? the region and how do they differ?” Finally, the difference of generated insights via geotagged social media data and traditional methods are presented in a table format as in table 2.2.
Table 2.2: Matching GSMD to the current understanding of tourist flows
(Chua et al., 2016)

<table>
<thead>
<tr>
<th>Features</th>
<th>Insights</th>
<th>Current Understanding of Tourist Flows</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographic</td>
<td>Largely patronized by domestic tourists (67%).</td>
<td>More domestic (63%) than foreign (37%) tourists.</td>
</tr>
<tr>
<td></td>
<td>Foreign tourists (33%) originate from many locations but Greek (9%), Dutch (8%), North American (4%) and Danish (4%) are most prominent. The remaining (8%) originate from other locations in Europe, East Asia, South America, the Middle East and Africa.</td>
<td>Anecdotes suggest that majority of the tourist are Dutch and Greek.</td>
</tr>
<tr>
<td>Circulation</td>
<td>Limited mobility inland.</td>
<td>No Information</td>
</tr>
<tr>
<td></td>
<td>Tourists transit along the coastline to travel long distances.</td>
<td></td>
</tr>
<tr>
<td>Spatial</td>
<td>Tourists primarily travel in a southerly direction passing through the coastal settlements on the road or rail network.</td>
<td>No formal studies of tourist movements till date.</td>
</tr>
<tr>
<td>Directionality</td>
<td>Southward flow due to the configuration of transport where transit hubs are located in the north of the region while the tourist attractions are located in the south.</td>
<td>Anecdotal evidence regarding mode of transport suggest that domestic tourists drive to their destinations while foreign tourist journey to Cilento by train, where they alight at either Capaccio or Valle di Lucania. Subsequent trips towards various locations in the region are then made by bus.</td>
</tr>
<tr>
<td>Centrality</td>
<td>Popular tourist attractions are located along the coast and have immediate access to the transport infrastructure. As a result, these locations are better connected than those situated inland.</td>
<td>Currently no consensus on any form of ranking.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Anecdotes suggest that individual municipalities claim to be more important than others.</td>
</tr>
<tr>
<td>Temporal</td>
<td>Analysis reveals a bimodal trend where foreign and domestic tourist activities occur over different durations and peak at separate moments in time.</td>
<td>Official tourists season begins on the 2nd week of May till the end of August.</td>
</tr>
</tbody>
</table>

There are numerous studies that explore spatio-temporal patterns of aggregate and individual mobility patterns in city-like regions using VGI. There are various sources for the data however: Instagram, Twitter (Chua et al., 2016; Xia et al., 2014), Foursquare (Noulas et al., 2011) and their combination (Hasan, et al., 2013; Cheng et al., 2011) are seen as most common.

The next application field that to be examined is the concept of event detection. Social media platforms such as Instagram and Twitter have users that generate vast amounts of geo-tagged content and shares it publicly in real-time (Xia et al., 2014). This enables people to connect relevant information with time and location elements. The real-time available data is so precise regarding time and location that the concept of hyper-local event detection emerged (Hu et al., 2013). CityBeat by Xia et al. is such a project that enables users to detect ambient social media activity in real-time. Hyper-local here means the detections of local events like music events, exhibitions and emergencies such as fires and car accidents at a street or venue level precision. This application is particularly aimed at the use of journalists and city officials with
the creation of visualizations and a user interface. The paper demonstrates an example where a NYC fire in 2013 was detected before any other local news outlet has reported it. While CityBeat is made possible with the use of Instagram data, Whoo.ly (Hu et al., 2013) is another example that uses Twitter data to detects and summarizes information hyper-local events.

The concept of hyper-local event detection bears utility for multiple purposes. At the most basic level, having information about their immediate local community is convenient if not useful for participation and awareness purposes for the individuals. Social media data in form of blogs, microblogs and various social network posts is particularly useful for the aim of hyper-local event detection via having advantages over classical media channels. These advantages are mainly due to the ubiquity and the immediacy of the social media data and are demonstrated on countless occasions where individuals get informed about certain events via social media before the traditional news media (Hu et al., 2013). Another advantage of social media is it remains available even when the traditional media outlets are censored by governments or criminal organizations, which helps attracting global attention to the events. (Parkinson et al., 2014).

Crowdsense is a project where TU Delft is one of the collaborators, that does social media mining for the times of crises. The main goal is to improve public safety during disaster times. The project has implications on multiple fronts such as semantic understanding of tweets, real time filtering and browsing of social media posts in order to be able to separate meaningful data and noise ("Crowdsense.co - researching technology for analyzing social media in times of crises," n.d.). Twitcident is their application that allows the user to monitor incidents via tweets in real-time. The interface has parameters such as tweet property, topic, location, tweeters and time slot. This application also has advanced functions where it is possible to search for the answer to the questions “What is the damage?”, “Are there casualties?” or “What is the risk?” etc. Functionalities like these are possible with advanced analysis of the text part of a tweet. The functionalities of Crowdsense and Twitcident are built by extracting great value from the text body of the tweets from twitter, there are also statistics, plots and maps available. The project demonstrates what is available with VGI’s in an inspiring way, however, the functionality is only limited to crisis situations and disasters.

As a host to the general discussion of how various data sources can improve the understanding of cities, the paper “Accidental, open and everywhere: emerging data sources for the understanding of cities” from 2014 discusses such data sources that have emerged in the last decade. These data sources enable researchers and also the public to access massive amounts of streaming data that is gathered via the internet and the mobile devices. The emergence of mobile devices and their universally popular use makes people leave digital traces of their location, emotions, ideas and memories (Noulas et al. 2011). If one conceptualizes these mobile services as an extension of the people which empowers them, then the citizens become sensors of the population (Goodchild, 2007). Meaning that we produce streams of structured and unstructured data that can be analyzed by the researchers to reveal different aspects of their and the society’s nature.

Naturally, with so many sources accentuating the potential of VGI, it is important to know where the limits lie regarding the potential of these data sources. Hochman, N., & Manovich, L. (2013) combines the perspectives of social computing, digital humanities and software studies with the aim of analyzing VGI with a multi-scale technique to answer these questions.
Their technique is unique in the way that it allows for exploration of both the metadata (upload dates, filters used, spatial coordinates) and the patterns of the content while still allowing the examination of individual photographs. Results are presented in a format of multi-scale visualizations which is a demonstration of possibilities by creating a spatiotemporal visualization of various cities which shows how the data can offer social, cultural and political insights about the activities of individuals and groups in particular locations and time periods. Finally, this extensive study also contains an in-depth analysis of the affordances of the Instagram interface and API which is often necessary if one desires to make in-depth analysis via a social media platform.

Undeniably, Foursquare is a popular gathering point for researchers who want to utilize VGI. Since many prior studies have used Foursquare data, therefore, it is crucial to understand the user behavior and other attributes of the social network. The study “An Empirical Study of Geographic User Activity Patterns in Foursquare” (Noulas et al., 2011) inspects Foursquare user behavior and check-in patterns in a quantitative way. Aside from being useful for generating hypotheses by revealing crowd patterns, this information is valuable for identifying threats for validity and various biases. As an example from this study, figure 2.4 shows check-in numbers by venue type and time of the day in a weekday.

Unavoidably, there will be challenges a researcher will face if they wish to utilize social network data for the purpose of urban planning and policy making. Arribas-Bel (2014) has a short section about such challenges where he addresses three major ones. Related to this study are data quality and skewed population representation and high-level skill requirement to perform state of the art analysis. The challenges that are mentioned in this paper are also considered when performing the selection of data source later on in this project.

Since the context of analysis matters dearly for data science, this project’s scope is defined from the beginning. The City of Rotterdam is chosen as the subject in this case. The most prominent studies such as the livehood project (Cranshaw et al., 2012) and Hoodsquare (Zhang et al., 2013) focuses on cities in the United States and the others great metropolitans.
around the world. The collected and publicly available Foursquare datasets are outdated and focus on mega metropolitans such as New York and Tokyo ("Foursquare - NYC and Tokyo Check-ins | Kaggle; Yang et al., 2016). Consequently, there are not many similar studies who focus on European cities. One study is found (García-Palomares et al., 2015) where the potential of data from photo-sharing services are demonstrated for identifying and analyzing the main tourist attraction in major European cities. While Rotterdam is one of the eight cities in the study, the study focuses on tourist behavior and not the behavior of the local population as this thesis intends to. Therefore, I wanted to give some desired attention to the City of Rotterdam and selected it as the focus of my project.

Consequently, with the aim of demonstrating how VGI can be utilized for generating insights for the City of Rotterdam, the goal of this thesis project formulated: Performing a case study on Rotterdam where social media data with geolocational components (VGI) is brought under the spotlight to assess its usefulness and actionability for the benefit of urban scientists and city planners. A model-based analysis of the collected VGI is to be performed to generate an actionable statistical summary of the data at hand. Therefore, a set of generative models will be created depending on the characteristics of the data and the aim of the analysis. This approach aims to face the challenges first hand and generate recommendations on how to deal with them. Data collection, data cleaning and performing analysis will be explained in detail while elaborating on the challenges that are faced during the process. The results are aimed at serving as a guide for city planners, urban scientists, municipalities and policymakers who seek to use publicly available VGI for the purpose of improving their designs & decision-making processes. On top of this contribution, the collected dataset will be made publicly available for other researchers to perform various analysis. The specifics of the datasets will be presented with the final report.

As the literature review is concluded here, the next chapter, explains my approach to the problem at hand. The train of thoughts that ended up forming the actions and decisions as to how to handle the progress of this thesis project is explained in the upcoming chapter.
3  Methodology

This section explains the overall process of the project and the line of thoughts that ended up creating the approach. First, the approach to the thesis task is explained, followed by the general flow of the project which is visualized in a flow diagram. At the end of the first part, the contribution of the thesis project and why it is a typical CoSEM graduation project of the Technology, Policy Analysis and Management faculty of TU Delft is explained. This is followed by a discussion on the societal relevance of the rising importance of data in the context of urban and societal management. Existing and potential use cases of effective use of data for the said context is also mentioned here. Finally, the third part of this section contains a user scenario that is conceptualized in order to develop an in-depth understanding over a specific use case and derive design recommendations for later steps, which corresponds to the business understanding of the CRISP DM methodology mentioned below.

3.1  Approach

This thesis project mainly aims to create results that are actionable for decision makers, however, there is a vast pool of possible applications to choose from. Therefore, in order to create a style of approach that fits the goal, inspiration is drawn from the CRISP DM methodology (CRoss Industry Standard Process for Data Mining). This methodology is a widely accepted data mining model that covers the core components of a data mining-based analysis: problem description, data analysis and final actionable result (Caprace, J., & Liege, A. (2017).

![Figure 3.1: CRISP DM methodology](Caprace, J., & Liege, A., 2017)

The first step of the CRISP DM methodology is business understanding, which focuses on understanding the objectives and the requirements of the project. Then, this understanding is used to convert the problem into a data mining problem. This means that I focus on a specific user scenario to begin with and expand from there in the future steps. From this
conceptualization, requirements for the users and the technical aspect of the application are derived. Throughout the project, these derived requirements are used as choice guidelines regarding the data source and the models in the analysis as shown in the diagram in figure 3.1 above. Drawing inspiration from the CRISP DM methodology, a flow diagram of the thesis approach is generated in the next paragraph.

As mentioned, my next step after this chapter is to look at possible data sources that comply best with the requirements derived earlier. There, I identify the three main social media APIs and evaluate their feasibility regarding my requirements. I decide that the Twitter API is the best source to collect VGI and I describe the process of doing so. Following that, I inspect the data that I have collected; showcasing the content and the relationship between variables within to the reader. Later on, I look at the model requirements and determine which model types are fit for performing the analysis required for the user-case application. I decide that simple Markov Models and mixture models are good fits for the requirements at hand. Following this, I pre-process my data and build the models. With each model, I focus on one type of application and evaluate how useful that model is for that specific application. After performing modeling & analysis, the initial design requirements are re-visited and discussed. Finally, the results of the thesis project are summarized in a way that is actionable for the decision makers and urban scientists. The approach flow described above is visualized in figure 3.2 below.

![Figure 3.2: Flow diagram of the approach](image)

As a disclaimer, while the thesis project has the basic characteristics of a design-based project, this is not a thesis with a full design approach. It does not contain the prototyping and revising elements of a design approach simply because they are not fit for the scope and the aim of this project as the main goal is not creating a data product or service. Therefore, I refer to the process as a semi-design approach.
The goal and the approach of this thesis make it a typical CoSEM (Complex System Engineering and Management) graduation project: the work has design and engineering components as mentioned. Moreover, there is a clear technological component in it as the data-driven approach and the technical issues are addressed along the way. A complex issue that is understanding the societal and civic impact of an emerging technological change is put under the spotlight. Finally, the designed technical solution may have an impact on the society as well as public and private sector. These elements put together form the core of the CoSEM curriculum. The following sub-chapter further discusses the societal relevance of this study by inspecting how the domain of data science can be used to help improving cities which eventually pose a change on us as individuals, and society as a whole.

3.2 Societal relevance

As a CoSEM student I am particularly interested in the societal and managerial perspective which are the focus points of the latter two categories of Thakuriah et al. Even though the public sector is lagging behind the private sector when it comes to making good use of data to improve the systems (Mullich, 2013), the governments realize the potential that lies in making good use of data and a transformation is happening as a result of it (Romijn, 2014). The potential benefits lie in improving decision making by making it data driven and supported. Understanding the physical and the political environment as well as citizen demands and trends in a better way and with more speed is possible with effective collection and analysis of data. It is also possible to increase the reaction and execution speed to current events with more insightful and data driven decisions (Bertot & Choi, 2013).

As our cities get more complex by getting bigger more crowded, the systems that serve these people also get more complex in order to satisfy the ever-increasing demands and needs. The more complex these systems these gets the more difficult they get to manage. The good news is that this increase in complexity also increases the amount of data that can be collected from these systems, which is the key to effectively managing these ever-growing systems.

One of the first questions that come to mind is “How can data help improving cities?”. In order to be able to answer these, we should first know what the data can be applied onto in an urban system context to improve it or maintain in better. The answer is quite fruitful, as any topic within an urban system is open to improvement with smart use of data such as: healthcare, economy, housing, environment, transportation, mobility, political environment, material flow, energy flow, use of land, use of green space, pollution, waste and so on (Cunningham & Verbraeck, 2018). To take this thought process a step further, some tangible examples of civic problems that are tackled with a data-driven approach are presented below, under four categories ("Catalog of Civic Data Use Cases," n.d.).
Table 3.1: Tangible use cases of data science in a city context

<table>
<thead>
<tr>
<th>Meeting points of Data Science &amp; Civic problems</th>
<th>Example use cases where data-driven approach is being used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Health &amp; human services:</td>
<td>&quot;How can cities deliver social services more efficiently?&quot;</td>
</tr>
<tr>
<td></td>
<td>&quot;What are the causes of infant mortality that could be targets for intervention?&quot;</td>
</tr>
<tr>
<td></td>
<td>&quot;How can we identify causes of poor air quality causing health problems for residents?&quot;</td>
</tr>
<tr>
<td>Infrastructure:</td>
<td>&quot;How can we predict the effects of service interruptions and other disruptions on transit systems?&quot;</td>
</tr>
<tr>
<td></td>
<td>&quot;How can we reduce the number of traffic accidents?&quot;</td>
</tr>
<tr>
<td></td>
<td>&quot;How can we better predict where will the next major pavement failure will be?&quot;</td>
</tr>
<tr>
<td>Public Safety:</td>
<td>&quot;How can we preempt youth violent crime?&quot;</td>
</tr>
<tr>
<td></td>
<td>&quot;Which offenders are most at risk of recidivism?&quot;</td>
</tr>
<tr>
<td></td>
<td>&quot;How can social media data help identify public safety issues?&quot;</td>
</tr>
<tr>
<td>Regulations:</td>
<td>&quot;Which properties cause the most problems?&quot;</td>
</tr>
<tr>
<td></td>
<td>&quot;Which restaurants are most likely to have code violations?&quot;</td>
</tr>
<tr>
<td></td>
<td>&quot;Which businesses are likely to be underpaying their taxes? How can we increase the productivity of auditors?&quot;</td>
</tr>
</tbody>
</table>

An extensive list of such civic and urban problems being solved by data-driven approach can be found on Harvard Kennedy School’s Data-Smart City solutions catalogue (“Catalog of Civic Data Use Cases”). This division of use cases into four categories is quite useful for attempting to conceptualize how the domain of data science is useful in an urban science context. Surely, this is not the only way to categorize the use cases. For example, Thakuriah et al. calls the effort of using data to improve cities and urban systems as urban informatics and categorizes its use under four purposes. First is urban resource management which is concerned about creating and improving existing plans and strategies that are aimed towards more efficient use of resources within a city. Second is knowledge discovery about urban patterns and processes which is focused on attaining insights about patterns and relationships that are within and between urban processes. The third, urban engagement and civic participation which is focused on keeping citizens informed and involved by developing various incentives and strategies. Finally, the fourth category is urban management, planning and policy analysis which is focused on delivery of services, generation and application and policies, management and maintenance of infrastructure by developing innovative solutions and improvements.

Most likely, the most promising tool for the governments for tracking the trends and citizen needs is the social media data that is being constantly generated by the population. The fact that individuals are displaying their appreciations and complaints for everyone to see on platforms like Facebook, Twitter and Instagram is quite valuable for the decision makers. In the hands of the governments with good intentions, social media data can be used to improve the current systems and policies, decide on where to focus the infrastructural spending and improving the lives of the citizens overall. However, there is also the other side of the coin where the authoritative governments are able to track down the opposition with great accuracy. Effective manipulation of the public opinion is another controversial side which is not limited to authoritative governments only (Media outlets of USA is a prime example.) The
ubiquitous use of social media also makes it easier to inject ideas to the population and enables the campaign holders to immediately get feedback on the effectivity of their moves, which enables them to refine their strategy much quicker and effectively. Tufekci (2017) mentions various examples of ethically questionable use of data on populations in the context of politics and voices out concerns about topics such as behavior manipulation, nudging and surveillance.

As discussed, effective use of data has great potential to benefit our cities. However, in order to achieve that level of effectiveness, tools to collect, organize and make use of data should be designed. Therefore, the following part is based a conceptual design of a city app that is based on social media data and this conceptual design is used for generating design requirements regarding the users and the technicalities of such data applications.

3.3 User scenario

In order to be better conceptualize the potential use cases and benefits of VGI, an example application is conceptualized here. First, the approach and the idea are introduced, followed by the discussion on potential benefits. Finally, requirements necessary for achieving the conceptualized service are discussed. These requirements are divided into three categories as user, data and model requirements. Over the rest of the report, this list of requirements dictates what is being demanded of potential data sources and the analysis & modeling sections build on top of these requirements.

**Real-time event detection app: EventTeller**

Rotterdam is one of the liveliest cities in the Netherlands. The city is home to countless social, scientific, cultural, technological and entertainment events every year. While most of them are in big dedicated halls such as museums, opera halls, music venues etc., there are also many events that are organized and performed in public areas. Aside from the innately social nature of the Dutch culture, the municipalities are also interested in creating active communities within the city. Having local communities increases the quality of life for citizens by creating activities and creating a sense of belonging while helping the municipality portraying Rotterdam as a socially active, attractive and lively city. This agenda of setting a certain frame for the City of Rotterdam will have results that will make Rotterdam attractive for most desired individuals and companies alike.

One of the elements that make a location more livable for their residents is the sense of community that it provides. Online communities such as on Twitter have the potential to provide the basis for a sense of community (Gruzd et al., 2011). In order to create a sense of community, it should be limited to a set group of people that have a common point. For the sake of this example, let us put a neighborhood under the scope. If a community manager wants to create a sense of community in this neighborhood, first they would need to hear about people’s concerns and needs. These will reveal the common points between the residents of the neighborhood to bond over and will act as a reason to gather. Next step is to organize a meeting -or an equivalent online event in this case- where people can let their voices be heard while getting to know their peers who are in similar needs with them. It is important here to clearly describe whom the meeting is for as this sets the boundaries for the sense of belonging.
(McMillan & Chavis, 1986). People then start getting to get to know each other and care for each other’s problems which most people share. Once people start feeling for each other, they start helping each other to solve their problems.

According to McMillan & Chavis (1986), there are 3 key elements for community building in this example:

- Calling in the residents for a meeting: creates an opportunity for membership.
- Members start accepting other people’s needs as an influencing factor for their behavior: leads to acting as a group while caring about others’ wellbeing.
- This sense of community built in the neighborhood serves as a catalyst for participation.

The question is how can VGI help with these steps. Mainly, VGI data that represents the population is useful for understanding what groups are likely to form within a population. This is may be done by looking at the individuals’ location and comparing this information to their demographic or social info which will reveal how people prefer to form groups. On the other hand, if no sense of community exists whatsoever, this information may be used to identify how to initiate in the creation of a sense of community by developing a hypothesis about the attributes such as member characteristics and size of a community. The clear benefit here is to develop an understanding of what activities and causes people are likely to bond over most.

While community building is often a viable way to increase the quality of life in urban areas, there are others way to do so as well. Naturally, there are also events with negative connotations happening when a big human population is co-existing in a dense urban area. Traffic and biking accidents, emergencies such as fires and protests are just to name a few. It is always beneficial to be informed about these events as quick as possible. In order for the information to be transferred, it has to be collected and sensed first. Luckily, in the 21st century every crowded city is densely packed with sensors that are almost always active: the citizens with their mobile devices.

The current technology of mobile devices lets individuals to instantly record various data types such as text, sound and video and broadcast it instantly. It is in the nature of social mammals to inform others when an attention-grabbing event is occurring nearby. Humans and their mobile devices set a perfect example for this. Imagine the last time you have witnessed an unusual happening: fireworks, traffic accident or a street concert, it is almost certain that there were multiple people recording, commenting about it and sharing the event with their social media network.

To make the rather obvious connection, a group of data scientists working with the City of Rotterdam can collect real-time data from social media services in order to detect various types of events happening in the city. What could be done with this information in itself is a big discussion topic. At first glance, this application could be provided as a public service for the locals and the tourists who could use the information to move towards or further from the crowds depending on the occasion.

There are numerous use cases for such an application and both the public and private sector is likely to benefit from the service. For example, intervention time for emergency services could be reduced. One may say that these systems work quite well in developed countries,
however it should be remembered that every second matters in such a context. For example, in 2013, a fire in NYC was detected via a similar tool called CityBeat before any other local news outlet has reported it (Xia et al. 2014). Crowdsense and Twitcident mentioned in the literature review are similar examples in this case and pretty much cover the possibilities of what value can be extracted from VGI’s in this concept.

By inspecting the patterns in repeating incidents and identifying the meaningful and useful ones, it is possible to understand the criminal behavior within an urban system. These patterns can identify vulnerable areas which can be visualized on a map for various crime types. Santos (2017) explains this process extensively within his book. These visualizations, aside from being good communication tools, may help the decision-makers identify the roots of the crime problem in certain areas. For example, Klinenberg (2018) argues (and supports with other sources) that areas that have damaged properties attract crime. This is because properties with broken windows, damaged walls and weeded yards give the impression that there is low, or no attention given to that area by authorities. This situation attracts the criminal behavior which seeks similar circumstances. Imagine overlaying the two maps, one showing the physical conditions of the structures and their maintenance, and other one showing the crime rates. A common pattern could be recognized here and eventually, solving the problem of damaged properties by generating policies that require the owners to maintain the facet of their properties could help solve this problem.

Another use case is to improve policy making. Not too long ago, decision making process and policy cycles were subjectively influenced by the personal beliefs and dogmas of the decision makers (Esty & Rushing, 2007). However, with such tools and applications to present empirical evidence on the changes that are caused by the policies, result and impact of each decision can be better understood. This feedback step will help solidifying the design cycles for policy-making and help improving designing more effective ones. One of such initiatives is called Internet-enabled cities across Europe (IES Cities), which aims to design and generate user-centric urban apps which enables various platforms for four cities in Europe. These services will focus strategic topics such as health, entertainment, mobility, culture etc. Moreover, the users will be able to continuously improve the provided services with their own data while having a voice in the proposal and development of new services to be added later on (European Commission projects: IES Cities project).

As mentioned in the societal relevance discussion in the previous sub-chapter, there are numerous ways that a data-driven approach can prove useful for problem solving and improvements on the context of urban science and civic management. For the case of this example application called EventTeller, examples are given such as community building, civic participation, improvement of emergency services, crime management and policy-making. These are examples out of the public safety, regulation, urban management, civic participation and policy making related use-case categories in sub-chapter 3.2.

In order to be able to provide the services that are conceptualized above, some requirements for the system are derived. The first set is called the user requirements, which are thought for the users and mainly concerned about functionality. From this first set of requirements, technical requirements are generated as counterparts of these user requirements. They indicate what technical qualities are required in order to be able to satisfy requirements of
user. While these technical requirements are based on the user requirements, they are not limited to them as some requirements are purely technical by their nature.

### 3.3.1 User requirements

In a way, table 3.1 serves as a summary of the user scenario preceding it by presenting the main requirements for achieving the mentioned services in it.

**Table 3.2: User requirements**

<table>
<thead>
<tr>
<th>User Requirements</th>
</tr>
</thead>
<tbody>
<tr>
<td>The users should be able to find people who share interests and ideas with them.</td>
</tr>
<tr>
<td>This functionality is mainly necessary for the community engagement and event management aspects.</td>
</tr>
<tr>
<td>The users should be informed about various types of events.</td>
</tr>
<tr>
<td>Event types should be identified by the system by looking at previous similar activity patterns.</td>
</tr>
<tr>
<td>The users should be able to access the information regarding the location of various events in the area of interest.</td>
</tr>
<tr>
<td>As a part of the core functionality, the system should inform the users about the location of the crowds and events in their area of interest.</td>
</tr>
<tr>
<td>The information should be reflecting the contents of an event in real-time.</td>
</tr>
<tr>
<td>For the case of community activities such as concerts and gatherings, there is a tolerance room. However for the case of emergency events, the system should be quick to inform the users.</td>
</tr>
</tbody>
</table>

### 3.3.2 Technical requirements

To achieve the functionalities above, the data collected from the social media services would need to have certain characteristics. Similarly, in order to create value from the data, some use of modeling techniques will be required. These characteristics are transformed into the form of requirements here. As hinted, the technical requirements are handled in two categories: data requirements and the model requirements.

**Table 3.3: Data requirements**

<table>
<thead>
<tr>
<th>Data Requirements</th>
</tr>
</thead>
<tbody>
<tr>
<td>The data collected should be representative of the behavior of the target population.</td>
</tr>
<tr>
<td>As an example, in the context of social media services, the service should be popular and its user base should consist of a broad range of age, gender and other demographics. Otherwise, the data will reflect the behavior of said user group instead of the entire population.</td>
</tr>
<tr>
<td>The location of the user should be available in the data.</td>
</tr>
<tr>
<td>For purposes of this application, location of the event is a crucial part of the data. Therefore, our interest is directed towards social media services which may include the location of the post or activity.</td>
</tr>
<tr>
<td>The data should contain information about the user itself, in addition to the their location.</td>
</tr>
<tr>
<td>This is necessary for the identification of crowd types and to be able to generate recommendations regarding communities and activities.</td>
</tr>
<tr>
<td>The data should be accessible in real-time.</td>
</tr>
<tr>
<td>The data collected should be reflecting the situation in real life. Therefore the data should be real-time generated and accessible so it is relevant for immediate action.</td>
</tr>
</tbody>
</table>
First, I need to find a source of data that will satisfy the corresponding requirements. Looking at the requirements above: the data should contain the user’s location along with the user’s general info, should be accessible as it is being generated and be representative of the population. At first though, popular social media services come to mind as the perfect candidates. Especially, in order to fulfill the requirement of data being representative of the population’s behavior, only the social media services that are most prominent with vast user populations will be inspected as potential VGI data sources.

The next step after the data requirements is the model requirements. These requirements indicate what type of information can be expected as a result of various modeling types. The models that will be considered for the analysis do not need to satisfy all of the requirements below. Instead, satisfying at least one of the requirements is enough to be considered in the model selection discussion. This is because not all model needs to have the same scope and contribution, it is possible that some models may have specific but deep contributions to the understanding of the underlying principles.

### Table 3.4: Model requirements

<table>
<thead>
<tr>
<th>Model Requirements</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dynamic element:</strong> Need to understand the mobility of the individuals within the city geography and this requirement focuses on the individuals, which are the dynamic component of the city. Ideally, each individual has a movement trajectory. Individually this may not mean much, however, once aggregated and analyzed in the grand scale, these trajectories may help for understanding the movements of crowds with respect to time of the day, day of the week, time of the year, and by specific events such as public holidays.</td>
</tr>
<tr>
<td><strong>Spatial element:</strong> Need to classify the locations and hotspots of a city by the purpose they serve and on the other side of the coin, the interactions and the flow generated by the individuals have various collection points which corresponds to locations in a city. Developing an understanding of the spatial element of the system complements the understanding of the individuals as their movements may load meaning to locations. For example, the central train station may contain a mix of various groups but a school district is mostly specific to students in its contents. In summary, spatial element is about being able to classify locations and hotspots by what type of activity patterns they host and what types of individuals they attract.</td>
</tr>
<tr>
<td><strong>Latent element:</strong> Need to identify the underlying factors that determine the classification of the users and the locations. Some attributes that are conceptually existing within a population are not directly observable. For example, individual types such as tourist, local, student, worker etc. While not being directly observable, it is possible to estimate these types by looking at the combination of movement and attributes of the individuals that are observed in the dataset. For the user, in order to be able to inform people about the ones that share attributes &amp; interests with them, the system should first be able to classify people accordingly.</td>
</tr>
</tbody>
</table>

Within this chapter, I’ve conceptualized a user case and utilized it to generate technical requirements that I use for the tasks that need to be performed through the rest of the project. In the following chapters, first, the data source selection is made with respect to the data requirements above. Next, the initial data exploration is performed. Following that, the candidate models are briefly introduced and compared against each other via the requirements above. A decision of which models to use in the analysis is then made with respect to the constraints such as time & complexity. This model selection process is explained in detail in chapter 6.1.
4 Evaluation of Data Sources & Data Collection

The first step after the generation of the requirements is to select the data source that will be used in the analysis. This section explains the source selection for the data collection and the collection methods of the data. For this process, social network services are selected as the VGI source. Three data sources will be inspected in this chapter, these are Foursquare, Instagram and Twitter. Their attributes and how they correspond to the previously defined data requirements are also discussed.

To briefly introduce the social media platforms that are candidate data sources for this project: Twitter is a microblogging platform which allows the user to succinctly share their current status within a limit of 140 characters. Foursquare is a location-sharing service which allows the users to broadcast their location. This action is called a "check-in". The strong suit of these services is that they pack a social functionality that connects the users and transforms the experience from an individual one into a community-based one (Noulas et al. 2011). Finally, Instagram is a social media application with an emphasis on photo and video sharing which is owned by Facebook. As of 2018, Instagram is one of the most popular social media applications in the western world.

The evaluation of these social media services will be made on 5 criteria, which are based on the data requirements of the previous chapter. These criteria are the popularity of the social media platform, whether if the API provides user’s location data, whether if the API provides user’s personal data, ease of accessing the data via the API and the API limits. Figure 4.1 shows how the evaluation criteria are derived from the requirements.
To briefly introduce what an API is, application programming interfaces were initially designed to assist the building of third-party applications, however, their functionality also opens the door for the researchers to freely access the data gathered by these services (Noulas et al. 2011). Their documentation websites* contains example uses and guides the users through API functionalities. The information contained in these sites in combination with GitHub and other aiding sources is enough to utilize the API’s to collect data from these services that focus on the City of Rotterdam (Or any country, city, county or location for the matter.).

4.1 API and data collection

4.1.1 Foursquare

For the large group of people who were active smartphone users during the years 2011-2013, Foursquare was one of the first apps that came to mind for collecting user-focused geolocational data. However, its population has declined ever since. In Foursquare API, a ‘venue’ is a place or an establishment that a user can check in. The ideal data collection approach here would be to collect a list of venues that are in Rotterdam and get the users checked in at those venues over a period of time such as 2 weeks. However, this functionality has been depreciated in the past years, however, there is a workaround. The venue data and
locations may be collected from Foursquare and these may be combined with the
geolocational user data from Twitter. It needs to be addressed that this method has a
significant downside from the original API functionality: it will not be possible to collect the
geolocation data shared from the households of the users. As visualized below by Noulas et
al. (2011), households were one of the most popular check-in locations for the Foursquare
users. The API also allows access to the basic user information such as user id, name, gender,
home city, friends and check-ins. Due to the functionality of the service, user location is always
available with check-ins.

Figure 4.2: Weekday check-in behavior (Noulas et al., 2011)

Considering the decreased popularity of the Foursquare, I assumed that Foursquare would
not provide much data regarding users. On the other hand, the Foursquare API has the unique
capability of providing thorough information about the venues of Rotterdam. For the collecting
data of the venues in Rotterdam, explore the functionality of the API is utilized. The
documentation can be found here. This functionality allows the developer to ask for place
recommendations with limitations of 50 per page(called limit) and 50 pages(called offset),
adding up to 2500 check-ins per venue types from the list of ['food', 'drinks', 'coffee', 'shops',
'arts', 'outdoors', 'sights']. For the code, see Jupyter notebook named “Foursquare API data
collector” at Dropbox.

4.1.2 Instagram

Instagram API’s generosity regarding user data has diminished quite a bit in the last years due
to privacy concerns, similar to Foursquare. It is no longer possible to collect posts from a
location such as a City directly. There is a workaround method that was discovered while
searching various data related forums. The subreddit r/dataisbeautiful has an example where
the data of Instagram posts are collected and plotted over the city of Istanbul (“Instagram
Posts of Istanbul [OC] r/dataisbeautiful,” 2018). Since the API does not directly offer this
functionality, I have messaged the content creator and asked for their methods of collecting
Instagram data over a city and they were kind enough to share. Their method relies on a
specific function of the API where the API provides data of the posts within a maximum of a
750-meter radius of a specified point. The creator of the Istanbul post has stated that they have scanned the entire City of Istanbul with a large number of said circles, running numerous parallel operations to collect data from all the circles and then merging them after the collection stage. This method requires the researcher to write a code script that will “scan” a city with 750-meter-radius circles. Writing and using such a script is deemed too time-consuming and thus out of the scope of this project. However, the possibility and the method to do so is revealed to researchers who are willing to go the extra steps to collect such data.

Regardless of having the least generous API of the three, Instagram has risen to be the most popular social media service in last years. In fact, it is in top five of most popular apps in the western world ("Global social media ranking 2018 | Statistic," n.d.). While it is still possible to collect meaningful data from Instagram, it seems that at least an intermediate coding knowledge is necessary to be able to collect meaningful data on a city scale. On top of this, working with image-based data is more complicated that the text-based data that is offered by other social media platforms. These two aspects may be a learning barrier that keeps researchers from using Instagram as their data source.

Another thing to note is that, researchers may need to hurry to collect data from Instagram as the documentation page states that even the basic functionality of reading user’s profile info and media will be depreciated in early 2020. This change is likely due to the fact that Instagram is owned by Facebook and the data privacy debates that has turned heads towards Facebook in 2018. At April, Facebook CEO Mark Zuckerberg had to give a testimony to the US senate about Facebook’s commitment to the privacy of their users following the revelation that the political consultancy firm Cambridge Analytica used data collected from Facebook users to psychologically profile voters; which has likely had a remarkable impact on the close call of political race that was the 2016 US elections (The New York Times, 2018).

4.1.3 Twitter

Twitter is one of the best social media platforms for researchers who are looking to mine publicly available social network data and VGI. There are numerous studies in the literature that use Twitter data and the Twitter API is the most generous of the 3 platforms that are being evaluated. Moreover, Twitter is still relatively popular at the time of this project, therefore, it is the best option for live data collection. Similar to Instagram, user data and user location are available depending on the privacy setting of the users.

For the purpose of data collection on the scale of a city, using the Streaming API of Twitter is the best and the recommended way. The Streaming API has tolerant rate limits (not specified by Twitter on purpose) that are appropriate for long-time connections ("Connecting to a streaming endpoint," n.d.). The Streaming API allows the collection of Tweets as they are being posted real-time instead of allowing access to past Tweet archives.
4.1.4 Decision for the data source

After evaluating the three social media services, Twitter is chosen to be the main data source for this thesis project. Instagram data is not preferred mainly due to the recently tightened rate limits and the necessities for workarounds to collect data on a city scale. Foursquare has fallen from popularity and thus, there is not enough user activity. Nevertheless, I still collected some data from the Foursquare API since it seemed to have some potential value regarding collecting venue data. However, after collecting the data via the API and inspecting the contents, I have decided that the value it was providing was being overshadowed by the data collected from Twitter. The results of the evaluation are summarized in table 4.1 below and the notebook containing the data collection and initial exploration of Foursquare data is accessible via Dropbox.

Table 4.1: Summary of data source evaluation

<table>
<thead>
<tr>
<th></th>
<th>Foursquare</th>
<th>Instagram</th>
<th>Twitter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large active user population</td>
<td>Not so active</td>
<td>Most active</td>
<td>Somewhat active</td>
</tr>
<tr>
<td>User location data may be collected</td>
<td>Always</td>
<td>If allowed by user</td>
<td>If allowed by user</td>
</tr>
<tr>
<td>User information can be collected</td>
<td>Depreciating due to privacy concerns</td>
<td>Depreciating due to privacy concerns</td>
<td>Mostly available</td>
</tr>
<tr>
<td>Easy to access data via API</td>
<td>User data is accessible, venue data collection requires workarounds</td>
<td>Requires workarounds, functions get depreciated due to privacy concerns</td>
<td>User data is accessible with known structure</td>
</tr>
<tr>
<td>API limits</td>
<td>Generous</td>
<td>Got drastically limited recently</td>
<td>Very generous</td>
</tr>
</tbody>
</table>

To summarize, Twitter is selected and recommended as the data source because it has a moderately active user base, does allow the collection of user location data and still has generous API limits despite the data privacy concerns that caused limit tightening on other APIs.

As for the real-time data collection from Twitter, a total of 50940 tweets are collected in the timeframe between 06-04-2018 - 17:01 and 21-04-2018 - 15:40. The results were aggregated in a folder which had the output files, each containing 100 tweets. 100 is selected as the checkpoint number for minimizing the loss of data during the collection of the data in case of connectivity problems. Later on, these files are aggregated together inspected as demonstrated in chapter 5, initial data exploration. For the code, see Jupyter notebook named ‘Twitter streaming API data collector’ at Dropbox.
4.2 Concerns regarding the data

The collective of observations that form the data may not always reflect the behavior & situation of the real-world population. This gap between the data and the reality causes a threat to the validity of a scientific study (Cook & Campbell, 1980). This chapter remarks some of the foreseen threats to the validity regarding social media and Twitter data. To put more clearly, how various attributes of the data may affect the validity of the study in later steps will be discussed here.

As stated by Arribas-Bel (2014), one of the most prominent concerns when trying to understand a phenomenon through data is the question of how representative is the sample of the population of interest. It is most likely that the Twitter users do not form a perfectly balanced sample of the population. According to the analysis of Omnicore agency, 37% of Twitter users are between the ages of 18 and 29 and 25% of users are 30-49 years old (Aslam, 2018). Another analysis done on the UK user base done by Sloan et al. (2015) shows that the youth is likely to be over-represented in the Twitter data. The same paper also discusses the representation issue by occupation and social class. Since the study is done on the UK population, it is not directly representative of the Rotterdam population. However, the under and over-representation of different socioeconomic classes and demographic groups should be kept in mind when performing the upcoming analysis.

Another issue to address is the political and social events at the time of data collection. Such events may increase the user activity attributed to certain demographic groups more than others in comparison, leading to further skew in their representation. One major social event that occurred during the data collection time window is the Rotterdam marathon that was done on April 8th. The marathon’s impact on tweet popularity can be seen in figure 4.4 which shows the number of tweets by day. Considering that the geolocational point of where the tweets were sent is important for this study, this may prove to have a rather strong impact on the results. Tweets about the marathon are expected to cluster on certain points such as the
finishing points and the spectator collection areas. These clustering points may end up creating artificial clusters depending on how populated they were.

Another potential skew source for the data is that not everyone who uses the social network services is local population. Zhang et al., (2013) addresses this issue by looking into the location history of individual users and marking the individual as a tourist if more than 50% of their check-in activity is outside of the city. While this is an acceptable way to roughly address this issue, the data at hand must allow such an analysis (The dataset used in the study of Zhang et al. (2013) “includes check-ins from Foursquare accounts forwarded to public Twitter profiles and is from 2010.). A fine example of differentiating locals and tourists is found in the Flickr page of Eric Fischer (2010), where he visualizes the classification of photos regarding whether they are taken by the local or the tourists (Figure 4.5) by analyzing individuals’ following photo locations and activity. Fischer has used Flickr and Picasa data to generate the visualizations. The functionality of the services may not allow such analysis in the year 2018 due to shrinking API generosity associated with increased awareness in data privacy issues.
Figure 4.5: Locals and Tourists #27 (GTWA #61): Rotterdam

Eric Fischer (2010). Blue indicates the pictures taken by the locals. Red pictures are by tourists. Yellow pictures are unidentifiable and could be both. Attention points from top-left to bottom-right: Den Haag, Delft, Rotterdam.

Noulas et al. (2011) also show in their study that people have different check-in behaviors for different places. Another aspect to consider is that people perform check-ins when they are performing certain activities which may have perceived social value instead of checking-in in their homes. The same behavioral effect can also be expected from Twitter, surely people are more likely to tweet during certain events like on special occasions, during lunch or while commuting. Once aggregated, these behavioral quirks will have an impact on the dataset and should be kept in mind while evaluating the results. This discussion about data concerns is further expanded later at the end of the analysis section with the lessons acquired during the analysis.
5 Initial Data Exploration

This section aims to familiarize the reader with the tweet dataset by exploring the contents of the dataset and basic relationships between the variables in a visual manner. Jupyter notebooks are used since they allow the demonstration the step-by-step approach with their report-like structures. Overall, Jupyter notebooks are quite useful for creating human-readable documents such as the ones that are prepared for this thesis project.

For the structure of data exploration, some popular data exploration examples from Kaggle (Marcelino, 2017; "Mercari - Simple Data Exploration with Python | Kaggle," 2018; Neviadomski, 2017) are used as a source of inspiration. To start with, I inspect the main properties of the dataset such as the size of the dataset and the data types. Following that, I perform exploratory data analysis by asking simple questions that can be answered by simple grouping functions and visualizations. Asking and answering these questions reveal basic insights about the data at hand and lead to further questions which move the exploration forward. The chart in figure 5.1 below shows the steps taken to explore the data and serves as a summary of the Jupyter notebook used for this chapter. (The Jupyter notebook named “Analysis of collected tweets” at https://github.com/Dozagac/Master-thesis contains the exploration done in this chapter.)

![Figure 5.1: Structure of data exploration](image-url)

- Column names
- Column Structures
- Size
- Data types

- NA values
- Empty Columns

- Locations

- Users

- Language use

- Time series analysis
5.1 Understanding the dataset

The first thing to do when exploring a new dataset is to take a general look at its size and the data types that are contained within. If there are data type mismatches, I will need to fix them before moving forward. The collected dataset consists of 50940 rows (tweets) of which 10158 have geolocational data available as the latitude and longitude of the tweet. There are 34 columns available, from these the most interesting and relevant ones are “coordinates”, “created_at”, “entities”, “lang”, “text” and finally “user” which is in a dictionary format that contains the user data. The data types of these columns are interpreted as an object by python, which is a flexible value type of python that enables the user to convert to any other type such as integer, string, datetime etc. Therefore, I do not need to worry about data types yet since I use python language. The summary table of the dataset which include all of the mentioned characteristics in detail can be found in appendix 1. The full structure and information content of a tweet is explained visually by Pierrot Péladeau. The document goes into full detail regarding what is in a tweet. While viewing it may trigger inspiration, not every detail of it is relevant to this report right now. The full image can be found in appendix 2 (Péladeau, 2012).

5.2 Inspecting missing data

Upon briefly inspecting the dataset, I see that the dataset has all null values for “contributors”, and always has zero for “reply” and “retweet”. The expectation is that there should be replies and retweets within 50940 tweets, which is not the case in the dataset. I theorize that this is due to the API’s privacy concerns. However, the main values that are necessary for the following analysis such as tweet’s creation time, user id, user language, place, text and date are always present in the dataset. As mentioned earlier, 40782 of the tweets do not have location enabled, which means they are not useful for an analysis that is interested in the locations. More detailed information about missing data can be found in the table at appendix 1.

5.3 Exploratory analysis

Now that I have inspected the basics such as the size, variable types and missing values, I can now start performing exploratory analysis by asking questions and answering them with data manipulation and visualizations. By doing this, I can find answers to my questions and reveal insights which lead to new questions. First, I will start by visualizing the locations of the tweets. Therefore, the first question I ask is “Where are the tweets in the dataset are located at?”.

5.3.1 Location of the tweets

One curious problem that was encountered during data collection stage is that the twitter API provided data that was outside the provided bounding box of (4.379512, 51.861634, 4.601300, 51.994311) which is as in figure 5.2 below.
Once the collected data was visualized, it was found that there were tweets located as far as to the German border and even Dunkirk in France as seen in figure 5.3 below. My theory is that the API collects the users that are registered inside the bounding box provided in the code, but still, collect the tweets of these users that were from outside of the bounding box.
This theory is further supported by the fact that around 20% of the collected tweets (10158 out of 50940 did) have coordinates enabled, even though the collection filter was already limiting the tweets to a bounding box. This means that the API is judging whether a tweet is belonging to the bounding box with something other than the tweet coordinates, which is most likely the user’s registration location. Nevertheless, this is not a big issue since the API ended up providing more data than it was asked for by expanding outside the bounding box. Therefore, the excess data is thrown out.

Once the outlier points are removed, a clearer visualization of tweet locations from Rotterdam is acquired. Figure 5.4 shows the plotted latitude and longitude values of the geo-enabled tweets. As expected, there is a dense cluster in Rotterdam city center area. Smaller clusters on Charlois, Delfshaven and Schiedam are also visible.

![Figure 5.4: Geolocations of Tweets from Rotterdam](image)

5.3.2 User data

The dataset contains tweets as rows which are not necessarily sent by unique users. With this in mind, I ask the question “How many unique users are there in the dataset?”. Upon a quick look, I see that there are 4997 unique accounts that tweets were sent from. Next, I ask “Who are the most popular tweeters in the dataset”. An inspection of the 10 most popular Twitter accounts shows that comedian Jochem Myjer has the most Twitter followers with roughly 1.3 million followers.
The most popular users are not necessarily the most active ones. Next, I ask “Who are the users with most tweets sent?”. The answer can be seen in table 5.1. This immediately reveals an interesting result, “BrugOpen” which is a public service that tweets information about which bridges are being opened for passing boats. Similarly, I see “Intercity Defect” which shares information about the defects of intercity trains. This public service functionality can be expanded to social city events and crisis situations such as fires and extreme weather conditions.

I see that not all the tweeters are people, there are also institutions and businesses that are active tweeters. This realization makes me curious about the language use within the tweets. My next question is “Which languages are the most popular among Twitter users in Rotterdam?”.
5.3.3 Language

Table 5.3 shows the top 5 most used languages within the dataset. As expected, Dutch is the most popular language as it is the detected language of 60% of the tweets within the dataset. This is followed by English, which corresponds to 23%. 7 percent of the tweets have unidentified language as they could simply contain only a link and no text, only mentions of others, non-grammatical use of words such as “Yessss!” or have no language specific text such as “???” . 2% of the Tweets are in Turkish while official reports show that 8% of the population is Turkish in Rotterdam (Centrum voor Onderzoek en Statistiek (COS), 2017). This is most likely because the Turkish users also use the Dutch language, nevertheless, this difference in percentages of representation hints at the challenges for accurately identifying the demographics of the tweeters.

<table>
<thead>
<tr>
<th>lang</th>
<th>% of all</th>
</tr>
</thead>
<tbody>
<tr>
<td>nl</td>
<td>0.602886</td>
</tr>
<tr>
<td>en</td>
<td>0.231115</td>
</tr>
<tr>
<td>und</td>
<td>0.070534</td>
</tr>
<tr>
<td>tr</td>
<td>0.019945</td>
</tr>
<tr>
<td>es</td>
<td>0.018021</td>
</tr>
</tbody>
</table>

Up until this point, I have explored location data and various user data. While focusing on the attributes of the columns reveal insights, it is also beneficial to take a step back and look at the aggregate data. For example, a time series analysis often reveals valuable insights.

5.3.4 Time series analysis

As for the exploration question, I ask “How does the user activity change by day?”. Figure 5.5 below contains the answer displaying number of tweets per day of the month in April when the data was collected.
The first thing that catches the eye is the highest activity point on the Sunday 8th, which was the day of the Rotterdam marathon. The high activity around the time is most likely due to the marathon since it is quite big of an event where people interact a lot and like sharing their successes with others online. Regarding the least active day that is 21st, the data collection was stopped mid-day on that day which explains the seemingly low user activity. Other than these, the user activity seems relatively close-leveled.

In a similar way, I ask “How is the user activity by the day of the week?”. The problem I anticipate is that the data collection window does not cover the same number of days of the week. From **06-04-2018 - 17:01 to 21-04-2018 - 15:40**, there are 2 of Monday, Tuesday, Wednesday, Thursday and Sunday but 2.5 of Friday and Saturday (half days of start and finish). Even while keeping this in mind when viewing figure 5.5, Thursdays still have the least activity which corresponds to the 12th and 19th in the graphic. Other than this, the weekends seem slightly more active than the weekdays, but I suspect that this does not reflect the general behavior due to the marathon that was on a Sunday and the fact that Saturdays have more representation. Figure 5.6 below shows the activity by day of the week.
Another time-related question to ask is “What time of the day do people tweet the most?” In order to answer this, I simply need to group the tweets by the hour. The results are as seen in figure 5.7.

The tweets are grouped by hourly bins. For example, the label 18 here contains all the tweets sent between 18:00 and 18:59. I see that there is a small peak around lunchtime, most likely since people have more time to send tweets around their lunch breaks. The peak is at 18:00 and a sharp drop around 21:00 can also be seen. Another style to visualize for the same question is presented in figure 5.8 below. This visualization is made using Tableau.
It is possible to ask more detailed and difficult-to-answer questions that would involve combining and manipulating multiple numbers of columns, however, I believe the exploration at hand is sufficient for the goal of getting a general understanding of the Twitter behavior in the dataset. With that being said, I stop asking questions for data exploration and point my attention towards exploring how VGI at hand as seen above can be made useful for the urban scientists with a more advanced analysis. The following chapter takes a step back from the data and starts discussing what various model types can offer for the analysis. Later on, these different model types are compared with respect to the model requirements a selection is made regarding which models will be used in the analysis section ahead.
6 Model Selection

This chapter explains the decision-making process regarding selecting models to perform the analysis with. First, the candidate model types are introduced, which includes information about their basic mathematical principles, notations and the parameters. After this introduction, the models are compared and two of the candidate models are selected to perform the analysis. The thought process stems from the model requirements that were determined in chapter 3. Upon reviewing the emergence papers of few model types, four model types are identified as potentially useful for working with VGI data that was collected from the twitter. The candidate models are simple Markov models, mixture models, hidden Markov models and probabilistic PCA.

6.1 Introduction to simple Markov model

Initially introduced in the late 1960’s, statistical method of Markov models has been rapidly gaining popularity recently. Markov models may be used for few goals such as learning the statistics of sequential data, doing prediction or estimation and recognizing patterns (Rabiner, 1989). For the readers that want to acquire a practical and introductory level of knowledge of the Markov models, Fosler-Lussier's paper (1998) is an ideal source. On the other hand, Rabiner (1989) is a more in-depth source for curious readers.

The Markov model is based on the first-order Markovian assumption as shown below. It indicates that the probability of an observation at time $n$ only depends on the adjacent observation at $n-1$.

$$P(w_n | w_{n-1}, w_{n-2}, \ldots, w_1) \approx P(w_n | w_{n-1})$$

(1)

Where,

$w_n$: Probability of going to hotspot i, given that the user is already at hotspot j ($w_{n-1}$)

This main assumption of the model offers a tradeoff where computational ease is acquired in return for a simplified representation of reality. Looking at the real numbers from a part of a matrix in table 6.1, If the user is already at hotspot 0, their probability of staying at the same hotspot is 21.1% and their probability of moving to hotspot 1 is 15.29%.

Table 6.1: Upper-left piece of the transition matrix

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.211033</td>
<td>0.152916</td>
<td>0.000000</td>
<td>0.000000</td>
</tr>
<tr>
<td>1</td>
<td>0.046348</td>
<td>0.129384</td>
<td>0.000000</td>
<td>0.000000</td>
</tr>
<tr>
<td>2</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.640121</td>
<td>0.000000</td>
</tr>
<tr>
<td>3</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.066495</td>
</tr>
</tbody>
</table>
The second part of the value that can be attained by applying a Markov model comes from the analysis of categorical random variable. The term categorical random variable here corresponds to a row of a Markov matrix which indicates the aggregated behavior pattern of the tweet dataset for a movement that starts from the point that the matrix row corresponds to.

There are \( k = 1, 2, \ldots, k \) rows each corresponding to a different probability distribution.

With two constraints (Teodorescu, 2009):

- First is that the individual probabilities should be bigger than 0:
  \[
P_{ij} > 0
  \]

- Second is that the summary of the probabilities in a row must be equal to 1:
  \[
  \sum_{i=1}^{j} p_{ij} = 1
  \]

(2)

Where the probability values are individually calculated as

\[
p_{ij} = \frac{M_{ij}}{M_i}
\]

The main use case for the categorical random variable analysis is that it allows the calculation of likelihood values for new data points. For example, for the trajectory \( t \),

\[
t = [1, 12, 7, 8]
\]

where points are taken from the list of hotspots, the likelihood of trajectory \( t \) is,

\[
l_l = P_{1,12} \times P_{12,7} \times P_{7,8}
\]

The calculated likelihood value is useful for assessing how likely is the new trajectory to belong to the population that generated the data. This is useful for purposes such as anomaly detection where a new trajectory with unusually low likelihood value may be flagged for unusual behavior.

6.2 Introduction to mixture models

The main purpose of cluster analysis such as the mixture models is the determine the inner structure of the clusters within the data when no other information is available other than the observed values (Picard, 2007). The approach of mixture models assumes that the collective data at hand is a mixture of a specified number of populations in different proportions. In our
approach, our assumption is that the population of users has sub-populations that could be students, tourists, commuters and so on. I assume that the trajectory data is a combination of these populations and the mixture model will help me separate the trajectories of these populations from each other.

\[ Y = \{Y_1, \ldots, Y_n\} \] notates a random sample of size n, where \( Y_t \) is a q-dimensional random vector with probability density function \( f_p(Y_t) \) on \( \mathbb{R}^q \). The assumptions it that \( Y_t \) can be written as:

\[ f(y_t) = \sum_{p=1}^{P} \pi_p f_p(y_t), \]

Where,

\( f_p(y_t) \) is a component density of the mixture, and \( \pi_p \) the weight of population \( p \).

With constraints:

- \( 0 \leq \pi_p \leq 1 \)
- \( \sum_p \pi_p = 1 \)

### 6.3 Introduction to hidden Markov models

The key idea is that an HMM is a finite model that describes a probability distribution over an infinite number of possible sequences.

The rough description of a hidden Markov model (HMM) is that it is a finite model that describes a probability distribution over possible sequences that are infinite in the count (Eddy, 1996). In a more practical sense, Blunsom (Blunsom, 2004) describes HMMs as “HMM is a tool for modeling generative sequences that can be characterized by an underlying process generating an observable sequence.”

As mentioned, HMMs are useful for when an analyst wants to understand variables and constructs - states, as in a state of the system - that are not directly measurable or observable. These states are “hidden” as they cannot be observable directly, hence the name. Instead, these hidden variables are inferred from the related data that is directly observable and measurable, also called the evidence in the model context.

A popular example is measuring IQ (intelligence quotient) points. The IQ points of a person is not a number that can be directly measured, instead, it is estimated from the test results that generate observable results such as success in solving puzzles and how fast one can answer various questions.

Mathematical representation of an HMM is as follows: (Blunsom, 2004)

\[ \lambda = (A, B, \pi) \]
Here, \( S \) is the state set and \( V \) is the observation set,

\[
S = (s_1, s_2, ..., s_N) \\
V = (v_1, v_2, ..., v_M)
\]

\( Q \) is defined to be a fixed state sequence with length \( T \), and corresponding observations \( O \),

\[
Q = q_1, q_2, ..., q_T \\
O = o_1, o_2, ..., o_t
\]

\( A \) is a transition array which stores the probability of getting to state \( j \) from state \( i \).

\[
A = [a_{i:j}], a_{i:j} = P(q_t = s_j | q_{t-1} = s_i)
\]  \hspace{1cm} (4)

\( B \) is the observation array which stores the probability of state \( j \) producing the observation \( k \). Both \( A \) and \( B \) are independent of time.

\[
B = [b_i(k)], b_i(k) = P(x_t = v_k | q_t = s_i)
\]  \hspace{1cm} (5)

Initial probability array \( \pi \) is then defined as such,

\[
\pi = [\pi_i], \pi_i = P(q_1 = s_i)
\]

The model has a key assumption that the current state is only dependent on the previous state. Similarly, the next state is only dependent on the current state. In the context of the Rotterdam tweets, the observation at hand is where individuals go and from where.

### 6.4 Introduction to probabilistic PCA

Principal component analysis (PCA) is a well-established dimensionality reduction technique, however, does not originally have a probabilistic component. Fundamentally, probabilistic models have signal and noise components; classical PCA is not a probability model because it does not address the noise. This means that while the analysis will yield meaningful results and even an R-squared score, it will not tell how robust the analysis actually is. Therefore, it will not be possible to comment on the likelihoods of rows and columns via this analysis.

The first probabilistic approach to the PCA technique is mentioned by Tipping, M., & Christopher, B. (1999). In this paper, probabilistic principal component analysis (PPCA) is introduced as a dimension reduction technique that analyzes the data via lower dimensional hidden space. It is useful for the cases where there are missing values in the data ("Edward - Probabilistic PCA).

They define a fully generative probabilistic model for PCA. Which means the model lets the analyst know what the probability is of seeing the observed data is, given a set of parameters.

Defining the generative model (Bishop, 1991; ZarlabUCLA, 2017):
Hidden variable $z_n$ is the projection of the $n$th data point, which lies in a $K$ dimensional space. Here, $z$ comes from $K$ dimensional multivariate normal distribution.

$$z_n \sim N(0, I_K)$$

The parameters of the model are $z_n$. The matrix $W$ and a new parameter $\sigma^2$ which is a variance parameter. Given these, it is possible to generate data $x_n$ according to the given distribution as shown below.

$$p(x_n|z_n, W, \sigma^2) = N(Wz_n, \sigma^2 I_M) \quad (6)$$

What this distribution is telling is to take $z_n$ which is a $K$ dimensional vector, transform it to $W$ which is an $N$ by $K$ matrix, which results in an $N$ dimensional vector. The $x_n$ is essentially centered around this $N$ dimensional vector, $Wz_n$, with some noise $\sigma^2 I_M$ added to it.

This process is done individually by each of the $n$ data points. Then we can write the joint probability density for $X$ and $Z$ (the $n$ data points) as below, which is simply the product of each individual step.

$$p(X, Z|W, \sigma^2) = \prod_{n=1}^{N} p(x_n, z_n|W, \sigma^2) \quad (7)$$

Following this, the goal is to set the parameters $W, \sigma^2 I_M$ to their maximum likelihood values.

### 6.5 Model comparison and selection

Each of the models that were discussed above has the potential to reveal valuable insights from the Rotterdam tweet dataset that was formed earlier. However, performing an analysis containing all of the model types would be impossible due to the time constraint and the scope of this thesis project. Therefore, a decision has to be made regarding which models will be used in the analysis. Table 6.2 below compares the models with respect to the model requirements that were generated earlier at the end of chapter 3. Each of the columns indicates a model’s potential of fulfilling the corresponding requirement.

<table>
<thead>
<tr>
<th>Model Name</th>
<th>Type</th>
<th>Dynamic</th>
<th>Spatial</th>
<th>Latent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple Markov Model</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mixture model</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hidden Markov Model</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Probabilistic PCA</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Mixture of Markov Models</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Mixture of Hidden Markov M.</td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>

Table 6.2: Model comparison with respect to model requirements
The **dynamic** column indicates whether the model has the potential to reveal insights regarding the movement within the city or not. This data for movement is mainly expressed as trajectories of individuals. Potential findings include understanding the movements of the crowd with respect to time of the day, the day of the week and time of the year.

**Spatial** indicates whether the model has the potential to reveal insights regarding the places and locations within the city or not. An example is understanding what purpose a location/hotspot is used for, which can be derived from what type of individuals they attract the most. Another example is understanding and developing an expectation of when a location will be host to a large number of individuals.

**Latent** variables have multiple definitions, in this context, however, I use the term for a construct, attribute or variable that cannot be directly observed from the data. Instead, it should be inferred from the other related data that is observable (Bollen, 2002). Therefore, the **latent** column on the table indicates if the model has the capability to reveal attributes that are not directly observable. For example, the type of an individual looking at their trajectories and how much time they spend at each hotspot. Here, I expect student trajectories to pass through schools and students to spend time at hotspots that correspond to schools.

The table is also organized so that the models are ordered in progressive complexity. That is, the models get more complex and difficult to apply as they are placed lower in the table. On top of this, as their names indicate, Mixture models and Mixture of Hidden Markov models are conceptually advanced versions of mixture models and simple Markov models. Therefore, there is a prerequisite for understanding the more basic models first.

In addition, the aim of this thesis project is not to build and demonstrate complex models but rather to serve as a bridge between the analysts and the city planner and the decision makers. It is most beneficial to establish an understanding of the simple Markov and mixture models first, which will also pave the way for future research to focus on more advanced models. Therefore, **simple Markov models and mixture models are selected to be used in the upcoming analysis in this order.**

Now that the model selection justified, the next chapter first explains how the cleaned data is wrangled to be used as a model input. Following this process, the analysis is performed by building and testing a simple Markov model and a general mixture model. The model sections are finalized by model evaluations and a general evaluation of the performed analysis.
7 Data Preprocessing and Analysis

On top of the data cleaning that was performed after collecting the data, in order to be able to use my data with the upcoming models, I also need to format the data properly. Section 7.1 is dedicated to performing this preprocessing of data. Within this chapter, the boxes with black background contain real code from the Jupyter notebooks while the boxes with gray background contain pseudocode that explains the process in a simplified and high-level way.

7.1 Preprocessing the data

1) Create a dataframe structure that contains the columns by transforming the tweet object with a the json format to a pandas dataframe.

At the minimum, I will need the time and location of the tweet. I also included tweet id and user id in the dataframe.

Structure of columns: [tweet_id, user_id, time, long, lat]

2) Limit tweets to Rotterdam area only

As mentioned in chapter 5.3, the initial collection of tweets was not strictly originating from the city of Rotterdam. Here, I manually filter the tweets to a bounding box.

From the dataframe structure previously generated:

Mask for 4.3895 < longitude < 4.5826 and
Mask for 51.868 < latitude < 51.988

The result corresponds to the bounding box in figure 7.1 below.
3) Identify the Hotspots

Identification of hotspots is done with a Gaussian mixture model. A Gaussian mixture model is a probabilistic model. The underlying assumption of the model is that all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters (“2.1. Gaussian mixture models — scikit-learn 0.19.1 documentation,” n.d.).

It is advantageous with its speed, according to the scikit-learn documentation it is the fastest algorithm for learning mixture models. The documentation also states that the model is agnostic which means that it will not bias the centers towards zero or bias the cluster sizes in certain ways.

The downside of the model is that it does not perform well with fewer data points. If there are an insufficient number of points per mixture, the algorithm diverges and finds solutions with infinite likelihood, however, this does not apply to our case. The code below generates the hotspots and figure 7.2 shows the hotspots in the City of Rotterdam.

```python
import sklearn.mixture as GMM

NHspt = 50
points = X[:,3:5]
model = GMM.GaussianMixture(n_components=NHspt, covariance_type='spherical').fit(points)
```
Naturally, some hotspots are more popular than others. These hotspots have higher weight attributes within the model. These weights are visualized in figure 7.3 below where the color of the hotspot indicates the quartiles of the weights of the hotspots.
As intuitively expected, the central train station and its very near hotspots are the most central ones.

4) Assign tweets to the generated hotspots

\[
\text{Labels} = \text{model.predict(points)}
\]

Now, all the individual tweets are assigned to one of the 50 hotspots generated above.

5) Detect user Trajectories

The pseudocode below explains the method used for generating trajectories.
Build a structure that contains the times of every tweet of a user, sorted from earliest to latest.

Within this structure:

For every user:
   Initiate a trajectory with their earliest tweet
   Compare the time passed until the next tweet in the structure (delta)
      If delta < 3:
          Add the tweet to the trajectory as the next step
      If delta > 3:
          Start a new trajectory with the second tweet.

6) Map trajectory steps to hotspots

Create an empty list to be filled.

For every trajectory in user:
   For every step in the trajectory:
      Find the corresponding hotspot to the location of the step.
      Append the hotspot number to the empty list

With the steps above completed, I can now initialize the Markov and mixture models. In the following sections, I will initiate and perform the models. To finalize each of the model analysis, I will look into the results of the models and discuss points of improvement.

7.2 Analysis

This section explains the process of building the models and the decisions made while doing so. Within the chapter, first the generative models are built and then the additional functionalities are explained. While doing so, each model's use cases and constraints are included into the discussion. The results are discussed in the next chapter.

7.2.1 Simple Markov model

With the size of the data at hand, I expect to have about two hundred trajectories. However, I will need to keep the model simple. Previous models were working with 50 hotspots but this time I will reduce it to 9 to attain more visually comprehensible results. Evaluation section ahead further discusses the effect of the parameter selection. The pseudocode below explains the steps taken in the preparation of the main matrix that will, later on, be used for identifying the mixture chains. Appendix 3 and the Jupyter notebook named “Model analysis of tweets” contains the exact code used, which accessible via Dropbox.
Detect hotspots with \( n=9 \) and covariance type = 'spherical'

Assign each tweet to a hotspot.

Count how many tweets are in each hotspot.

Create a hotspot lookup dictionary which matches tweet id to hotspot id (This dictionary will be used to map trajectory steps to hotspots.)

Map trajectories to hotspots

Calculate prior probabilities from the data at hand.

Create the matrix

Fill in the matrix (skip trajectories of length 1)

Add 1 to the diagonal so I do not divide by 0 later on.

Again, add 1 to the matrix so the visualization has smooth color comparison.

Figure 8.8: Combination of mixtures with \( n=9 \) hotspots

It is visible that hotspot 0 is the most popular one, which is the central train station as expected. Hotspot 1 only connects to itself and to hotspot 5 as seen figure 7.6 above, which is surprising since they are far apart and there are multiple hotspots that are closer to hotspot 1.
Now that I have this main matrix that includes the entirety of the tweet populations, I can start looking for the sub-populations that form this mixture as the mixture models assume.

7.2.1.1 Expectation Maximization Algorithm

In order to ensure that I am using the model with the highest quality of results, I am defining a set of functions which I use to build an expectation maximization (EM) algorithm (Dempster et al., 1977). EM starts by taking a random initial guess at the parameters then tweaks the parameters of the model to fit the guesses and the observed data ("EM Algorithm (Expectation-maximization): Simple Definition," 2015). The basic steps of the algorithm are:

1) Start the model with a random guess of initial parameters and probability distribution.
2) Include the new observed data into the model
3) The input in the first step are modified to include the new data
4) Repeat step 2-4 until stability of distribution is acquired.
5) Save the parameters that have the greatest likelihood value.

Appendix 4 and the Jupyter notebook named “Model analysis of tweets” contains the exact code used, which is accessible via Dropbox.

Now the required functions are built, I will start with a random guess and iteratively improve the model and capture the parameters of the model with the highest likelihood score. This
application with 200 iterations is shown in figure 7.8 below. After 200 steps, the parameters of the best acquired are saved in order to be able to recreate the same results in the future.

![Figure 8.10: Log-likelihood scores starting from a random guess](image)

Looking at the y-axis of figure x, we see that the trials start negative and keep decreasing the likelihood which is not how I was expecting it. Log-likelihood scores are negative as expected, however, the likelihood value was supposed to increase toward 0 throughout the iterations. My theory is that the model assumes trajectories to be a uniform length where in reality most trajectory lengths differ with longer ones being rarer. When multiplying probabilities ($0<P<1$), the longer the multiplication chain, the less likely a chain will appear. Therefore, even though the model is most likely improving through the iterations, the likelihood graph does not show it. Model evaluation section below offers this problem a solution.

Once the parameters of the best model are saved and then used to run the model to get visualizations, the results including the mixture chains below are acquired.
This main matrix above is separated into 3 chains that form it. Upon visual inspection, it can also be seen that the top matrix is a visual overlay of them 3 chains with different weights. The contribution weight of each chain is indicated by their pi values. The results section further discusses the results of the model.

7.2.1.2 Assigning new trajectories/users to the chains

As previously mentioned in the simple Markov model section, likelihood calculations are also useful for mixture models. For the case of mixture models, if a trajectory is known, I will be able to estimate which chain the trajectory most likely belongs to.

Categorical random distribution approach that was previously used for simple Markov model is useful again. I need to ensure that the values of the matrix rows of each chain sums to 1 since I want to treat the row as a probability distribution. Code block below shows the python code that achieves this.

```python
chain1_normalized = pd.DataFrame(chain1)
chain1_normalized = chain1_normalized.div(chain1_normalized.sum(axis=1), axis=0)
```

The operation is almost identical to the simple Markov application performed earlier in chapter 7.2.1.2. The difference is that now I will calculate the likelihood of the trajectory for all the...
chains, and then select the chain that has the highest likelihood value as the chain most likely owns the trajectory as shown in the pseudocode below. Instead of performing a hard assignment as in simply selecting the chain with highest likelihood, a probability of assignment is calculated. This probability of assignment is simply the likelihood value of the selected chain divided over the sum of all likelihood values.

Function: Chain_selector
Function input: Matrices of chains, a trajectory

If: Trajectory length <= 1

    Print “Error: Comparison cannot be made”

If: Trajectory length > 1

    Calculate likelihood of the trajectory in all 3 chains
    Select and return the highest likelihood value and the pairing chain.
    Calculate probability of assignment = likelihood of selected chain / sum of all likelihood values of chains

This analysis section has demonstrated how the models were built and what important functions were used in their generation. In the next chapter, discussion, I inspect the results of the models and try to estimate what the resulting Markov matrices and the chains may mean in sense of human behavior and preferences.

7.2.2 Mixture model

A brief scan of the literature did not reveal any papers that have demonstrated the applicability of Markov models to the case of geolocational social media data. There seems to be a gap in the literature regarding the combination of this method and approach, therefore it is exciting to move forward with Markov models and assess their applicability for this project. Markov models could be useful for the activities of crowd management where the transition between attraction points can be notated as a transition matrix. They also have potential to be useful in anomaly detection.

This model maps the co-occurrences of hotspots through the movements of the users regardless of time. Assumedly, there is a pattern to the movement of users, but this model is blind to the movement and only “sees” the snapshots of users being on certain hotspots. This is due to the Markov property which indicates that the next state of the system is only dependent on the current state and not on the sequence of preceding states (Washha et al., 2017).

The visits by each user are divided by the square root of the total number of visits made by that user. This is done so that the matrix acquired as the result can be used as a transition matrix between hotspots. If this normalization is not made, then the diagonal of the matrix does not behave as a count of total visits to each hotspot. Next, the matrix is dot-produced by
its transpose to achieve the transition matrix. Finally, the values within the matrix are logarithmized for better visualization and figure 7.4 below is the result.

```python
# all_users is a list of trajectories with minimum length of 2 (as we need an edge between 2 nodes to create a matrix)

U = np.array(all_users)
ss = np.sum(U,axis=1) # Total number of visits made by that user
w = np.power(ss,0.5) # Take square root
U = U/w[:,None]

# Multiply the matrix by its transpose to get the transition matrix.
UT = U.T
UU = UT.dot(U)

pd.DataFrame(data = UU) # This better for viewing

# transform the Markov plot for easier display
markov_log2 = np.log(UU+1) # add 1 to not divide by 0
```

![Transition Matrix](image)

**Figure 7.4: The transition matrix between hotspots**

The values of hotspot 6, 12 and 27 (yellow) catch the eye as the highest values, these points are the points where the people are likely to retweet the most (Retweet in this sense means tweeting from the same geolocation by the same user, not the platform functionality called retweeting).

### 7.2.2.1 Random walks and Crowd management

The transition matrix which was previously obtained from the Gaussian model shows the probabilities of going from a certain hotspot to each other hotspots. With this information, I can
generate a random walk where artificial users would go from their initial hotspots with a series of random steps. Where the step probabilities are notated in the row of the matrix that corresponds to the current hotspot, a random number picking function such as the one from the "random" library of python is sufficient to perform this. Figure 7.8 shows an example random walk of a single user on the city map.

![Random walk starting from hotspot 1 with 5 steps](image)

Figure 7.5: Example random walk with 5 steps
(Starting from central train station)

Selection of how many steps to take is an important part of the model and defines the core assumptions. Understanding the human behavior plays a big role in the selection of this parameters. The purpose of the users will also matter. While tourists may have several spots to visit within the city, regulars of the city are likely to have a single destination spot which is their place of residence or work. One aspect to keep in mind is that these steps are randomly picked. Which inherently means that some generated trajectories will be more likely to occur than other ones. The next chapter is about this issue.

### 7.2.2.2 Categorical random distributions and likelihood estimation

Since the process above includes a random selection from a probability matrix, some very unlikely trajectories are bound to be generated. I would like to be able to assess how likely it is for a trajectory to occur in real life (As we understand from the dataset). In order to achieve this functionality, categorical random distribution with likelihood estimation is utilized.
To simply put, each row of the Markov matrix that has been generated is a categorical random
distribution. When an individual is starting their movement from the hotspot row that
 corresponds to the said matrix row, probabilities of their next steps are demonstrated in the
said row. In order for this to work, the sum of the probabilities in the row should be equal to 1
since it is a probability distribution. This is achieved by dividing each element in the row to the
sum of the row and replacing the original value with the result. I call the resulting matrix as
transition matrix as it shows the transition probabilities between the hotspots.

\[
\frac{\text{Divide the individual cell to the summation of the row.}}{
P(i \to j) = \frac{M_{ij}}{\text{sum}(M, \text{by axis 0})}}
\]

Next, I create a function to generate likelihood scores. Likelihood scores are simply the
multiplication of every step’s probabilities in a trajectory. It is convenient to log these values
as I only care about the comparison between them rather than their actual values. A simple
function that calculates a comparative likelihood score is generated as in the pseudocode
below.

\begin{verbatim}
Function input: (a trajectory of hotspots, a matrix to get the probabilities)

For every step starting in step i in the trajectory,
    Get the corresponding probability from the transition matrix that shows the movement
    of step i to j. (M_{ij})
    Multiply the probabilities of each step.
    Return the log of the multiplication of all step probabilities.
\end{verbatim}

With the use of a likelihood calculation such as this one, I will be able question how likely a
trajectory at hand is to occur. If taken a step further by setting a threshold of normal likelihood,
some of the actions can be labeled as anomalies. Doing so would result in a basic anomaly
detection functionality. This approach of likelihood calculation is also used later in the mixture
models for assigning a trajectory to one of the derived chains.
8 Discussion on Model Results

Here, I initiate the discussion about the models by looking at the resulting matrices of the models and inspecting the interesting hotspots on the map of Rotterdam. By looking at the nearby landmarks, I try to understand the purpose of various hotspots and then I try to utilize this understanding to develop an idea about population’s behavior and preferences.

8.1 Simple Markov model

Here, I look at the resulting Markov matrix of the simple Markov model in figure 8.1 and discuss what type of human behavior or preferences it may hint at.

![Figure 8.1: The transition matrix between hotspots](image)

The diagonal values being high (indicated by light colors) can be interpreted as users being more likely to tweet at the same spots. This is most likely due to the relatively small size of the data; however, this also makes sense in a behavioral way. Individuals could be more likely to send posts during routine activities such as waiting for transport, commuting, during breaks or at entertainment venues. The values of hotspot 6,12 and 27 (yellow) catch the eye as the highest values, these points are the points where the people are likely to retweet the most (Retweet in this sense means tweeting from the same geolocation by the same user, not the platform functionality called re-tweeting). Upon inspection, it is also visible that vertical and horizontal lines of some hotspots have lighter colors(higher values) than the other hotspots around them on the graph, an example is the cross shape that can be seen at horizontal & vertical of hotspots 27 and 28. This means that the people from other hotspots are more likely to come to these hotspots.
As expected, these spots are rather easy to recognize. Hotspot 27 (left) is at a block that consists of a casino, a movie theater, a cinema and a concert hall right across the central station. Hotspot 12 (right) is right where the Beurs metro/tram station and the Maritime Museum is at. Hotspot 6 (top) is exactly where the Wolfert Lansing High School is. This could be a slight indication of the fact that tourists are more active than the residents when it comes to Tweeting in the city of Rotterdam. With this hint, I hypothesize that tourists have a high representation within the collected dataset. Also, the appearance of high school confirms that the youth(students) is rather active on social media.
By calculating the column sums of the matrix, I can rank the relative attractiveness of the hotspots. Starting from the most attractive with the rank order, hotspots 27, 12, 28, 8, 38 can be seen above in Figure 8.3. These are the top 5 most attractive hotspots of Rotterdam.

Hotspot 28 is on top of Rotterdam Vakcollege de Hef, similar to hotspot 6, the activity here could be associated with the rather high social media activity of youth and the students. Hotspot 8 is where the “Cube houses” (Kijk Kubus) is located at, one of the main tourist attractions of the city. Finally, the Hotspot 38 is right on top of Casino Roman Palace, further supporting the hypothesis about the dominant demographic of the tweet data being tourists and the youth.

Overall, I believe that approaching Twitter data with the Markovian modeling techniques has potential value. However, the simple Markov model results in a very high level of abstraction of the actual human behavior. The Markovian principle that indicates that the next state of the system is only dependent on the current state of the system and not the ones before is a good example to discuss the high abstraction. While human beings have patterns in their daily lives that may be captured by this mode, our behavior is often responsive to the series of events that led to our current state.

Moreover, the model simply shows an aggregation of the entire population’s behavior over a certain time. Without a doubt, this basic approach has room for improvement. In the case of mixture models, for example, I assume that each hotspot is a specific type of place that attracts a certain type of users. This attraction type and strength are changing over time, but the model
still can provide insights into the time period in which the data is spanning over. Therefore, in my opinion, the mixture model is superior to the simple Markov model in this analysis. Another example of an improvement over the simple Markov model is the Hidden Markov model. HMM has the base assumption that the population that generated the data at hand is made of a limited number of subpopulations.

8.2 Mixture model

Here, I will take a closer look at all 3 chains that are the result of the mixture model. I will inspect the interesting patterns on the matrices which are indicated by brighter colors and try to come up with behavioral explanations for these patterns. Figure 8.4 below shows the hotspots that were identified by the mixture model on the map of Rotterdam. In order to understand the nature of these hotspots and their nearby environment, I have manually inspected these locations with the use of Google maps. The following table 8.1 shows the contents of the immediate region of the hotspot points. Based on the attraction points near each hotspot, a classification type is estimated. Table 8.2 shows these assigned types, which are merely hypotheses generated by looking at the google satellite map.
Table 8.1: Potential attraction points and nearby buildings of hotspots

<table>
<thead>
<tr>
<th>Hotspot #</th>
<th>Nearby:</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Rotterdam Centraal</td>
</tr>
<tr>
<td>1</td>
<td>Restaurant Aan de Zweth (Michelin Guide mentioned)</td>
</tr>
<tr>
<td>2</td>
<td>Special Education School: Instituut Mr. Schats Loc. zuid, Rotterdam Stadion</td>
</tr>
<tr>
<td>3</td>
<td>Paintball Rotterdam, Community garden de Venhoeve, PostNL ScB Rotterdam</td>
</tr>
<tr>
<td>4</td>
<td>Bakkersland Rotterdam, Schuttevaerweg bus stop</td>
</tr>
<tr>
<td>5</td>
<td>Residencies</td>
</tr>
<tr>
<td>6</td>
<td>Wartburg College and International Christian Fellowship of Rotterdam, Zuidplein Shopping Center</td>
</tr>
<tr>
<td>7</td>
<td>Consumer attractions such as restaurants, supermarkets, massage salons, winery. Male modelling agency</td>
</tr>
<tr>
<td>8</td>
<td>Markthal and the Kijk Kubus</td>
</tr>
</tbody>
</table>

Table 8.2: Classification of Hotspots according to satellite view

<table>
<thead>
<tr>
<th>Hotspot #</th>
<th>Location:</th>
<th>Type/Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Rotterdam Centraal train station</td>
<td>Public transport center</td>
</tr>
<tr>
<td>1</td>
<td>Kandelaarbrug, on Delftse Schie canal</td>
<td>Consumer</td>
</tr>
<tr>
<td>2</td>
<td>By Rotterdam Stadion</td>
<td>Sports</td>
</tr>
<tr>
<td>3</td>
<td>Highway E19 and A20 intersection network</td>
<td>Entertainment</td>
</tr>
<tr>
<td>4</td>
<td>West of city, Car repair and dealership district</td>
<td>Residency / industrial</td>
</tr>
<tr>
<td>5</td>
<td>Rotterdam South-east</td>
<td>Residency</td>
</tr>
<tr>
<td>6</td>
<td>The Zuiderparkplein and Zuiderpark</td>
<td>Education / consumer</td>
</tr>
<tr>
<td>7</td>
<td>Rotterdam Vasteland-Leuvehaven</td>
<td>Consumer</td>
</tr>
<tr>
<td>8</td>
<td>Markthal and the Kijk Kubus</td>
<td>Touristic</td>
</tr>
</tbody>
</table>

Now that I have knowledge about the hotspots and have generated these tables above, I can use this knowledge to inspect their collective impact in a behavioral sense. I now will now inspect the chains and attempt to assign behavioral meanings to the each of the 3 chains.
With the weight of 0.629, this chain is roughly representing the movement of ⅔ of the tweeter trajectories. The first element that catches the eye is that the chain looks like a diagonal matrix. The diagonal elements on the chains indicate that the users are not moving between hotspots, instead, they are tweeting often from the same spot.

Other than that, the hotspot 6 and 8 seem to attract the most users from other hotspots. Hotspot 6 is at where the Zuiderparkplein and Zuiderpark are at, the closest building being Wartburg College, Zuidplein Shopping Center and International Christian Fellowship of Rotterdam. Hotspot 8 is at where the Markthal and the Kijk Kubus are at. I hypothesize that the locals who often tweet from the same spots may prefer these places at hotspot 6 and 8 for their leisure time activities. Adding on top of this hypothesis, I can assume that these stationary locals correspond to roughly ⅔ of the population as the chain weight (0.629) indicates.
Chain 2:

[Image: Map showing hotspot locations]

Figure 8.6: Chain 2 accompanied by a map that shows hotspot locations

With the weight of 0.134, this chain is the one that explains the smallest number of trajectories. The first thing that catches the eye is that hotspot 2 has no transitions other than itself and hotspot 5, which is the closest hotspot to 2 on the map. This can be interpreted as the people who tweet from hotspot 2 stay there and tweet at the same spot. At the location of Hotspot 2, there are three establishments that may be the source of such behavior: Breederode Hogeschool, a church named H. Kruisvinding and special education institute named Institute Mr. Schatz. My theory is that the hotspot 2’s self-containing tweet behavior is best explained by the Breederode Hogeschool since I expect the Hogeschool students to be more active tweeters.

The bright colored horizontal lines on the matrix can be interpreted as the hotspot of this line is sending out the most users to other hotspots. If I look at visible horizontal lines, I see that hotspot 0, 3, and 8 are the most visible ones. Number 0 is the central train station, so this behavior is expected. Number 3 is located on top of a paintball center and a big Post NL office. The first correspondence explains why people would be tweeting from that spot since it is possible that the paintball is a tweet-worthy experience for the most. However, it doesn’t explain why it would be a point that is often transitioned from. It is also worth mentioning that the second most bright colored point on the matrix is from hotspot 0 (central station) to hotspot 3 (paintball field). Finally, as previously mentioned hotspot 8 is the tourist attraction of Kijk Kubus and Markthal. The behavior around this spot may often be explained by the preferences of tourists.
The three brightest spots on the matrix are the connection of 0-0, 2-2 and 8-8. I have already mentioned how the points on the matrix diagonals can be interpreted while discussing chain 1. This chain has an interesting rectangular bright spot in the middle that was not observed on the other two chains. This bright spot corresponds to the connections from hotspot 3 to 3, 4, 6 and 4 to 3, 4, 5, 6. My initial hypothesis that these spots may be public transport points since they act as trajectory starting or propagating points. Upon inspection, I see that hotspot 4 is by the Schuttevaerweg bus stop. However as mentioned previously, hotspot 3 is by a paintball field and a large PostNL office, which does not support my hypothesis. However, if the PostNL office is a distribution center, then the tweeting behavior of the drivers may support this hypothesis. This last chain the weight of 0.236 which corresponds to about the quarter of the total trajectories.

Overall, I believe the chain 1 has the most correspondence behavior of a segment of the population. Chain 2 and 3 seems like they could be further broken down or the mixture model’s separation of trajectories is not very interpretable with an intuitive approach. After inspecting the results and having discussed over them, it is time to question what these models lack and what can be made different in the future to improve these models. The next chapter answers these questions.
9 Evaluation and Validation

Within the chapter, the model analysis and the choice of models are evaluated and discussed upon with respects to my assumptions and their shortcomings. Using the models as a starting point, I shift my focus on to validating my decisions regarding the various aspects of the thesis project. I start by explaining how I validated my code as I was handling the data and the model work. Next, I question the relationship between the collected data and to what degree it reflects the reality that I am interested in understanding. Finally, I recommend improvements for further attempts.

9.1 Regarding the Markov models

The number of hotspots were selected to be 50 for this model. The number of 50 was selected to be an intuitively sensible number to represent the number of trajectories which was around 750. One point to consider when working with Dutch cities is that there is a Dutch way of dividing a city into sections. As explained by Wassenberg (2006), there are larger neighborhood areas which are called “Wijk” and these consists of smaller ones which are called “Buurt”. In Rotterdam there are 22 wijks and 92 buurts (“Alle wijken en buurten in Rotterdam,” 2018). An interesting approach is to choose the number of hotspots to be equal to the number of wijks (n=22) or buurts (n=92). Figure x and x below displays the hotspots when n = 27 and n = 92 respectively.

![Figure 9.1: Hotspots of Rotterdam (n=22)](image-url)
While figure 9.2 seems a bit too crowded, this approach to decide the number of hotspots may have interesting results for future research.

The model's main weakness is its lack of memory that is a result of the Markov property. On the most basic level, this means that three consecutive steps could be a loop in the form of $A > B > A$ in the random step generation. While this certainly is within the possibilities, it will not fit every situation. Expanding on this idea, in reality, some hotspot pairs could be less likely to be on the same trajectory or more likely to be observed together. Realistically, human behavior is dependent on the previous steps and locations which the model is not capturing due to its high level of abstraction that comes from the Markov property.

As a big improvement, adding a memory function to the model can try to address this. This can be done by looking at triplets and quadruplets of steps instead of only pairs which have a single step in between.

### 9.2 Regarding the mixture models

There are three main parameters that can be modified in the current application at hand. First is regarding the generation of trajectories. The delta value which indicates how much can time pass in between tweets from the same user so that the tweets belong to the same trajectory is selected as 3 in the current approach. If this delta is decreased to 2, for example, it is expected that the number of trajectories will drastically drop. Doing so may not be beneficial for understanding the movement of individuals as people are likely to not tweet every two hours. There are also issues with increasing this delta. I assume that the tweeting locations are the only places that the individuals go to, however, if I put tweets that have 5 hours in between them into the same trajectory, it would be a naive assumption to expect that the user...
has not been moving in between. Therefore, the selection of this delta value is a crucial task as it may undermine the main assumptions and create a threat to the validity of the model.

The second parameter is the number of hotspots which was selected to be 9 for the mixture models. It was lowered from the 50 in the previous Markov model analysis as fewer hotspots would ensure more tweets per hotspot which results in a better model. It also makes the visual interpretation of the results easier. Therefore, the selection of this parameter depends on the aim of the analysis. If one aims to get an in-depth explanation with complex results that uses a data source that has a big number (and frequent) entries (that may be provided by a major cell phone service provider for example) a high number of hotspots would be more fitting.

The last parameter is the number of chains that the main trajectory matrix is assumed to contain. Again, the number of mixture chains selected to be a low number such as 3 in order to ensure that there are enough data points per mixture chains. In theory, each of these chains is expected to reflect the movement of a major user group such as tourists, commuters and students. Trying a different number of chains and visually inspecting each chain to see if their movement is corresponding to a certain user group could be a feasible way to select this parameter. Of course, this will depend on how many user groups are expected by the researchers. Initial assumptions might be generated with a combination of initial data exploration and context knowledge.

One problem of the model that I discovered while trying to improve the model parameters is that the current version of the model does not take into the account that there is also a distribution in the lengths of the trajectories. As seen in table 9.1, almost all of the trajectories consist of 6 or fewer steps. Each of the transition steps within a trajectory corresponds to a probability value of going from a hotspot to the next. When multiplying probability values (0<P<1), the longer the multiplication chain, the less likely a chain will appear. This property makes it tricky to compare the likelihoods of trajectories with different lengths. In order to address this, a future version of the model should also consider the length of the trajectory when calculating the log-likelihood scores. For determining the probability of choosing which trajectory length to choose, the distribution in the “Weight” column in table 9.1 can be used.
Another thing to consider is that the main assumption of the model, which is the assumption that there are subpopulations within the main Rotterdam Twitter user population and the mix of these populations in different proportions form the total population. However, this may not be the case, identification of individuals’ trajectories may require more than one tag in reality. For example, while the current version of the model is trying to identify hotspots as a hub for students, there are likely groups such as commuting students, local students, students who work part-time etc. within this interest group. The current state of the model is too simplistic to address this level of complexity.

Increasing the number of chains is not the answer as it is also important to have enough elements within a mixture chain for the reliability of the model. On the other hand, too little number of mixture chains may cause a high level of abstraction from the reality of the population. For future attempts, the use of hidden Markov models and the mixture of hidden Markov models have the potential to identify multiple overlapping populations such as the previously mentioned student sub-groups.
9.3 Internal validity test for code

For the purpose of checking the internal validity of the code, white box and black box testing methods were used. The concept of white box testing indicates that the interior of the function is tested in regard to outputs, broken or poorly structured functions, the functionality of conditional loops, and testing of each variable and statement. Black box tests, on the other hand, are tests to see if the functions generate the expected outputs for the given inputs (Nidhra & Dondeti, 2012). Figure 9.3 below shows the logic of these test techniques.

![Figure 9.3: White and black box testing](image)

As the figure indicates, a test is done by checking a set of predefined desired inputs against the generated outputs. Whenever a specific input does not result in the desired output, the test catches an error in the code which could be a bug or a logic error. White box testing is practical when the interior of the functions is known, and black box testing is practical when it is not.

Therefore, I used this testing technique for the functions that I have defined myself. On the other hand, for the functions which I did not write the code myself, black box testing was a better fit to test the internal validity of the code. Python libraries have predefined functions that always expect a certain format of input to generate the expected output. These predefined functions have often complex interior structures which are not practical to individually test. Instead, the documentations of the used libraries ARE used to understand which inputs should be put in and which outputs should be expected. As mentioned, I have tested these predefined library functions through black box testing. As the name indicates, the library functions are black boxes to the users since they can only see the input and the output of the functions.
### 9.4 Regarding the data

The first problem to address is the relatively small size of the data used for this project. Some projects found in the literature (Cranshaw et al. 2012; Cheng et al., 2011) collect data for months and end up with millions of data entries. However, for this project data was collected for around a month, which ended up gathering roughly 51000 tweets. From this starting pool, I ended up with 4500 tweets which had geolocations enabled and checked in at least once in Rotterdam. Finally, only around 750 users had geolocation enabled and tweeted from multiple spots within the assumed same trajectory window of 3 hours.

One other phenomenon to be aware of regarding the source data is the overrepresentation of certain groups within the dataset. For example, findings of chapter 8.1 suggest that the youth and the tourists are most likely over-represented within this Twitter dataset. This problem is inevitable while working with social media platforms as every platform has certain demographics under and over-represented. Yang et al. (2016) and Chua et al. (2016) also acknowledges a similar phenomenon of over-representation of youth and tech-savvy individuals while working with social media data.

Another behavioral phenomenon that raises questions about the validity of using tweets for analysis is that people are simply more likely to tweet regarding certain types of events and thoughts. Simply put, people are likely to tweet while performing more exciting activities and from unique occasions than their routine lives. This causes a mismatch between the user behavior and the researcher’s goal if the aim of the research is understanding people’s daily movement patterns for example. Being aware of behavioral effects like these are crucial before performing the analysis as the results may be misleading to the researcher. Intuitively, this phenomenon is less disrupting if the goal of the study is to understand tourist movements for example.

Tweets do not reveal mobility, they are intrinsically in the forms of time snapshots. I assume that the information acquired by looking at snapshots is still useful for understanding movement behavior. However, in the time that is in between two snapshots that I observe, the users may be doing various activities which the model would miss. This is most relevant to the generations of trajectories, which assumes two tweets to be consequent steps of movement if they are created with less than 3 hours of time in between. Missing the fact that within the time window of 3 hours between tweets people could be moving significantly but the approach would miss it.

Finally, similar to the phenomenon related to tweets not revealing mobility, the frequency of tweets may not reflect the intensity of the crowd. If the crowd of subject consists of a demographically underrepresented group such as the elderly, I would most likely fail to observe the real size of the crowd solely looking at the number of tweets.
One interesting finding regarding the tweet data from Rotterdam was that the great majority of most active Twitter accounts belongs to non-human users. In fact, only 6 accounts of the top 30 most active tweeters seem to belong to human users judging by the account names as shown in figure 9.4. Treating these services in the same way with the human accounts is a mistake and it causes noise in the results and may misdirect the interpretation of the analysis. For future work, I should decide from the data cleaning step whether if I want to keep these accounts or not depending on my analysis goals. For the goal of understanding human behavior within a city, for example, they should be excluded from the dataset.

Public services such as “Intercity Defect”, “Brug Open”, various 112 emergency services and that serve a purpose of informing the citizens about the anomalies happening within the city. The account “Intercity Defect” has 600 tweets that are sent from the same stationary spot. This and other stationery but active public services create a bias on the Markov matrices and cause the diagonal of the matrices have higher values that the population’s behavior would cause. On the other hand, tweet accounts such as “TomTom Jamnet NL” and other 112 services such as “112Alarm.net RRM” are not stationary and still quite active as seen in figure 9.5 below. These services are harder to automatically detect compared to the stationary ones and may require a manual cleaning by looking at the account names. Cleaning the most active tweeters at the least would help the dataset to be a better representative of human behavior and mobility.
Now that I encounter these problems, I see a need for a defined data cleaning process. The basic idea is to add a screening process to the data cleaning activities which would separate non-human accounts from the human ones. The ideal way would be to have a binary categorical in the dataset that indicates the type of the account such as “business”, “organization”, “individual”, however tweets not have such a variable in their structure.

9.5 Future research

This section discusses how the approach at hand can be expanded and improved in the future by utilizing other various data sources.

For the goal of expanding the analysis with further models in the future, the use of hidden Markov models has the potential to reveal hidden states within the sum of trajectories similar to what is done with the mixture models in this paper. Using hidden Markov models are likely to also help the researchers label the hotspots by their intended uses. Moreover, the mixture of hidden Markov models which is a combination of both model as its name indicates, have the potential to identify hidden distributions within the sub-populations.

A way of data collection has not been addressed so far is data scraping. You might be familiar with the popularity graphics of venues if you’ve searched for the opening hours of a restaurant.
for example. These are usually displayed at the right side of the screen after your search and look like figure 9.6. If one manages to scrape these, their collection on a city-wide scale may help the researchers simulate the crowd movements within the city center by the hour. I have mentioned scraping because this data is not offered by Google API to be collected. However, one library called “populartimes” was found on GitHub which was specifically developed to collect this information (“m-wrzr/populartimes,” 2017). I encourage the interested readers to check out the GitHub page as there are also demonstrations of what visualization for the cities of Munich, Barcelona, Berlin and London which shows hour-by-hour crowdedness of these cities.

Another way of acquiring real-time generated geographical information is the use of positional tracking data acquired by having the users connected to a Wi-Fi source. When a mobile device is connected to the Wi-Fi signal, the provider of the Wi-Fi signal may access the exact location of the user. A company named Lone Rooftop (“Building Intelligence - Lone Rooftop,” n.d.) uses this functionality in a building scale to track where the users are and inside the building and minimize the waste of heat and electricity at the unused sections of the buildings. A similar approach may be used for cities, let it be for resource conservation or crowd movement tracking. A potential application is to explicitly observe how the regular movements of crowds get affected by temporary changes in their paths. An example question that may be answered could be: “How would the movement of pedestrians change according to the construction in this specific street?” or “How would the movement of commuters get impacted if this specific public transport stop was closed?”. Even though this technique requires entire sections of cities to be provided with “free” (Payment in form of personal data) Wi-Fi, I believe that the Netherlands is capable of being able to provide free Wi-Fi to the almost entire landscape of its major cities (if that is not the case already).

Earlier in the threats to validity section, I have mentioned that the frequency of tweets may not reflect the intensity of the crowd. With the use of a different data source such as the location or activity log from the service providers such as Verizon, Vodafone etc. this issue could be addressed. It is expected that these service providers have a continuous stream of
geolocational data which can be used to capture evenly spaced snapshots of crowds (Every 5 minutes for example). The data from the service providers would also provide a rather homogeneous representation of the population, unlike Twitter data. As expected, the use of such data may raise privacy concerns and access rights discussion to the data between different parties who are also interested in acquiring it.
10 Theoretical Reflection

There are numerous examples in the literature where social media data was used for improving the understanding we have about various systems in cities and their residents. In this context, social media data has been used for numerous purposes such as semantic discovery (Gao et al., 2014), understanding human behavior & activity (Coleman et al., 2009, Hochman & Manovich, 2013; Hasan, 2013; Liu et al., 2014), understanding community structures (Gruzd et al., 2011; Tang & Hui, 2010), hyper-local event detection (Xia et al., 2014) and understanding dynamics of a city (Cranshaw et al., 2012). For researchers, there are many sources for social media data to choose from such as Facebook, Instagram, Foursquare, Flickr, Twitter and the relevancy, functionality and popularity of these social platforms rapidly change.

Cunningham & Verbraeck provides a high-level overview of what data types and sources available to understand a city. While performing a similar analysis, I have focused on social media data sources and specifically focused on data collection opportunities and challenges of Twitter, Foursquare and Instagram. In this context, I have systematically compared these to each other within chapter 4. As a result of this comparison, I have found out that Foursquare’s user base has diminished quite a lot since its popular time around 2011-2013 time-window and is unlikely to be very useful for researchers who wish to understand human behavior via using it. On the other hand, Twitter is still reliable as a social media data source. While Twitter has its flaws in biased population representation (see 9.4) and not-so-high user activity, a better alternative does not exist in terms of social media services. Finally, it is apparent that Instagram has quite the potential to help us achieve a deep understanding of social and cultural activities and patterns of people (Hu et al., 2014). However, since it offers image-based data that is rather difficult to work with and the API is not making it any easier for researchers to collect data on a city-scale it does not see the interest it deserves from the academic community, which is also stated by Hu et al. (2014).

Earlier in the literature review, I have mentioned Milgram’s theory about how cities have different characteristics that result in unique “atmospheres” of cities. While I did not dedicate myself to understand the atmosphere of the City of Rotterdam, one thing I perceived from the tweet data that I collected is that the City of Rotterdam hosts an international community. Roughly one-third of the tweet’s languages are non-Dutch (table 5.3) and there is respectable touristic activity around casinos and attraction points such as the Erasmus bridge, Grotemarkt and Kijk Kubus. Looking at these attributes, I can state that certain areas within the city also have different atmospheres and specializations attributed to them.

Milgram also states that the individual’s perception of a city is directly affected by the nature of their stay. A tourist, a resident and a newcomer will have different ideas about the city. While agreeing to his statement, I have also seen during this project that Twitter data has a lot of potential to help confirm or reject this idea or to understand to what degree it has an impact. Simply the location, language, and the text data that is contained in a tweet can provide great insights into how different demographics experience the “atmosphere” of the city in Milgram’s terms. Without a doubt, tourists of Rotterdam would have a different cognitive map of the city than the ones living in the less affluent areas at the south. These cognitive maps are also functionally similar to the “chains” that are generated as the result of the mixture models. Each
of the chains corresponds to a collection of behaviors, meaning that it is likely the case that one chain corresponds to the group that can be labeled as tourists while another one corresponds to the "locals" label. Milgram suggests that these cognitive maps could help measure the cognitive significance of various areas within the city. These perceptions are expected to vary between demographic groups. In my opinion, understanding the nature of these variations is sure to have value for public and private organizations.

Overall, the main realization that I had while being in the process of creating this thesis project is that cities are great areas for knowledge discovery. There are numerous data sources available for doing so such as built-in sensor systems of the city, administrative data (census data, neighborhood records), data from the private sector (phone network providers for example) and of course social media data (Thakuriah et al., 2016). With many of these data sources, it is also possible to scope down to various levels of society by grouping people by their location, situation, behavior and preferences. This focused look at the fabric of society can help researchers answer specific questions and attempt to solve related problems. However, as I have experienced first-hand with my analysis of my selected social media data sources; some data sources are often more fit for the challenge at hand than the others. Therefore, it is indeed important for urban scientists and decision-makers to have an understanding about which problems can or cannot be solved with which data sources (Cunningham & Verbraeck, 2018).
11 Conclusion

The main goal of this thesis project is to form a bridge between the data science and urban science domains, extending from the data science side of the gap. To achieve this, I took a semi design-oriented, model-based approach using real data with the aim of investigating and communicating the potential benefits that lie in utilizing volunteered geographic information (VGI) sources for the city planners and urban scientists. VGI in this sense refers to the publicly available social media data with a geolocational component that has been created by the users in a voluntary fashion. While the conceptualized event detection app EventTeller is not realized as a part of this thesis, many other applications already exist (Hoodsquare, CityBeat, Whoo.ly, Plenario, Crowdsense, Event Alert, CrowdTrack for example) and function for the benefit of the cities and their residents. These applications are mainly built by collecting and analyzing social media data, which often comes with a geolocational component. The early chapters of the thesis show that these data applications have multiple ways to benefit city planner, urban scientist and decision makers.

Social media data is popular amongst urban science researchers as well. There are numerous studies in the past that have used one or more of Twitter, Foursquare and Instagram as their data source (García-Palomares et al. 2015; Noulas et al. 2011; Zhang et al. 2013, Cheng et al. 2011; Li et al. 2014; Cranshaw et al. 2012). As years pass by, applications and their users change; meaning that the reliability of these social media platforms are bound to change over time. Especially since 2018 has been a year full of data privacy discussions, end of 2018 is a good time to understand where these platforms stand from an availability of data perspective. Therefore, this thesis project evaluates the usefulness and accessibility of the API’s of these social media platforms for researchers and brings them up to date for late 2018, which was a year where a lot of attention was paid to data privacy. As the result of the comparison between selected social media applications, Twitter is selected as the main data source for the analysis section due to providing the most granular data and being the alternative with the most extensive API access. The considerable downsides of the other two were that Foursquare lacked the user activity for the case of Rotterdam and collecting Instagram data required rather advanced code and working with image-based data.

For the model analysis part of this thesis, simple Markov and mixture models were used. A unique element of this thesis project is that to my knowledge, it is the only academic paper that uses simple Markov and mixture models to analyze Twitter data. In this regard, it serves as an example that demonstrates the usability of the said models and may serve as an introductory guide to researchers who desire to perform a similar analysis. For now, I can state that the simple use of these models may be too abstract to reflect the complex human behavior. However, there is potential in their use and trying similar more advanced methods such as hidden Markov models may yield promising results. Another way to improve these results is using other promising data sources such as Wi-fi tracking data, cellphone data, governmental census data and Google API to answer similar questions to the ones in this thesis project may result in better models and better understanding of which data sources reflect human behavior better.
REFERENCES


73. Tufekci, Z. (2017, September). We’re building a dystopia just to make people click on ads [Video file]. Retrieved from https://www.ted.com/talks/zeynep_tufekci_we_re_building_a_dystopia_just_to_make_people_click_on_ads#t-1360611
APPENDICES

APPENDIX 1 - Summary tables of the dataset

```
df.info()
# 10158 Tweets have geo or coordinates enabled
<class 'pandas.core.frame.DataFrame'>
Int64Index: 50940 entries, 0 to 50939
Data columns (total 34 columns):
contributors       0 non-null float64
coordinates        10158 non-null object
created_at         50940 non-null object
display_text_range 31371 non-null object
entities           50940 non-null object
extended_entities  5981 non-null object
extended_tweet     10255 non-null object
favorite_count     50940 non-null int64
favorited           50940 non-null bool
filter_level       50940 non-null object
geo                10158 non-null object
id                 50940 non-null int64
id_str             50940 non-null int64
in_reply_to_screen_name 23335 non-null object
in_reply_to_status_id 21873 non-null float64
in_reply_to_status_id_str 21873 non-null float64
in_reply_to_user_id  23335 non-null float64
in_reply_to_user_id_str 23335 non-null float64
is_quote_status     50940 non-null bool
lang               50940 non-null object
place              50940 non-null object
possibly_sensitive 24051 non-null object
quote_count        50940 non-null int64
quoted_status      4124 non-null object
quoted_status_id   4124 non-null float64
quoted_status_id_str 4124 non-null float64
reply_count        50940 non-null int64
retweet_count      50940 non-null int64
retweeted          50940 non-null bool
source             50940 non-null object
text               50940 non-null object
timestamp_ms       50940 non-null int64
truncated          50940 non-null bool
user               50940 non-null object
dtypes: bool(4), float64(7), int64(7), object(16)
memory usage: 12.2+ MB
```
APPENDIX 2 - Structure of a tweet

A Tweet message is less than 140 characters. However, information items of code generated by Tweeter represents dozens of lines.

**Text**: A unique identification number of this message.

**Created at**: The date and time of this message in universal time (UTC).

**In reply to user id**: Identification number of the user who sent the reply.

**In reply to screen name**: The name of the user who sent the reply.

**In reply to status id**: Identification number of the status that was replied to.

**Favorited**: Indicates that the tweet has been favorited.

**Truncated**: Indicates that the message was truncated after the 140th character.

**Entities**: Information about context or constraints on the message's delivery.

**Possibly sensitive**: Indicates that the tweet includes an URL that links to sensitive content.

**Scope**: If applicable, the message's components that are clickable or screenable: hashtags, URLs, mentions of other users, and media.

**Retweet count**: The number of times that this message has been retweeted.

**Text**: The account name is withheld in: Greece, Hong Kong, Malaysia.

**Withheld in countries**: "GR", "HK", "MY".

**User**: A unique identification number of the user.

**Screen name**: The name of the user.

**Description**: Social assessment of interpersonal information systems.

**Entities**: If applicable, URL (internet link) that user included in the description.

**URL**: A URL (internet link) to the user's profile.
"created_at":"Fri Feb 27 21:00:14 +0000 2009",  * Date of creation of Twitter account by user
"profile_background_image_url":"https://api.twitter.com/profile_background_images/4b62863/wood_characters.jpg","profile_background_tile":false,  * Background image chosen by user for own Twitter page
"profile_banner_url":"https://api.twitter.com/profile_banners/819797/1348102824","profile_background_color":"1babf1", "profile_link_color":"2c6ef1",  * Default or user chosen colors for page's characters and bars
"profile_sidebar_border_color":"87bec4", "profile_sidebar_fill_color":"00f9f2","profile_text_color":"000000",  * Number of messages picked by user as favorite across whole account's life
"favourites_count":112,  * Number of messages produced by user up to now
"statuses_count":1735,  * Number of other users' accounts to which this user subscribes
"friends_count":135,  * Number of other users subscribing to this user's account
"followers_count":270,  * Number of public lists including this user's Twitter feed
"listed_count":18,  * Number of public lists including this user's Twitter feed
"time_zone":"Eastern Time (US & Canada)";  * Local time zone that user associated with this account
"utc_offset":-18000,"lang":"fr",  * Interface language chosen by user
"protected":false,  * If applicable, fact that user opted to restrict messages' access only to own followers
"geo_enabled":true,  * If user permitted or not geotagging of own messages
"verified":true,  * If user has a verified account certifying identity (against hoaxes and identity thieves)
"contributors_enabled":false,  * If this user permitted another user to author messages
"contributors":null,
{"id_str":"819797","screen_name":"prete-plume"},  * Identification, if applicable, of user who wrote this message for the official author
"coordinates":null,
{"coordinates":null,"type":null},  * Identification, if user permitted it, of the longitude and latitude of location the user associated to this message
"place":null,
"bounding_box":null,
{"coordinates":null,"type":null},  * Longitudes et latitudes of intersections between sides of location's perimeter
"source":"web";  * Digital utility used to post this message

* Minimal screenable content  Other accessible content
Requested by user  Requested by another
Partial list of information items linked to a Twitter status
Pierrot-Peladensu.net  Dec 15, 2012
# Predict centroids only based on longitude and latitude.
model = GMM.GaussianMixture(n_components=9,
covariance_type='spherical').fit(X[:,3:5])

# Assign each tweet to a hotspot.
labels = model.predict(X[:,3:5])

# Count how many tweets are in each hotspot.
label_count = np.array([0]*9)
for alabel in labels:
    label_count[alabel]+=1

# This dictionary will be used to map trajectory steps to hotspots.
hotspot_lookup = {}
for n,row in enumerate(X):
    print(n)
    tweet_id = row[0]
    hotspot_id=labels[n]
    hotspot_lookup[tweet_id]=hotspot_id
    print(tweet_id,hotspot_id)

# Map trajectories to hotspots
all_traject = []
for traject in all_tweets:
    print(traject)
    atraject = []
    for an_id in traject:
        print(an_id)
        hotspot_id = hotspot_lookup[an_id]
        print(hotspot_id)
        atraject.append(hotspot_id)
    all_traject.append(atraject)

# Calculate prior probabilities from the data at hand.
priors = np.array([0]*9)
for traject in all_traject:
    priors[traject[0]]+=1
priors = priors/np.sum(priors)
# Create the matrix
markov=np.zeros((9,9))
# Fill in the matrix
for traject in all_traject:
    traject_length = len(traject)
    n_last = None
    if (traject_length > 1):
        # skip trajectories of length 1
        # n is the sequence number of the trajectory
        # i is the hotspot number
        for n,i in enumerate(traject):

            if (n==0):
                next
            else:
                markov[n_last,i]+=1
        n_last = traject[n]

# Add 1 to the diagonal so I do not divide by 0 later on.
markov = markov+np.identity((9))

# Again adding 1 to the matrix so the visualization has smooth color comparison
plt.imshow(np.log(markov+1)) visibility of relations.
# Initialization
# Randomly assign all trajectories to one of three mixtures

def initial(all_trajec):
    k = len(all_trajec)
    h = []
    for i in range(k):
        v = np.random.dirichlet((-1, -1, -1))
        v = v.tolist()
        h.append(v)
    h = np.array(h)
    return h

# We are going to need to soft-weight the maximization

def expectation(all_trajec, params, pi):
    h = calc_ll(all_trajec, params, pi)  # Calculate the likelihood of each trajectory being generated by each chain
    l = calc_ll_total(h)  # Calculate log likelihood total
    h = norm_ll(h)  # Normalize
    h = np.array(h)  # Convert to array
    return l, h

def calc_ll_total(h):
    # Calculate log likelihood total
    l = np.sum(h)
    return l

def calc_pi(h):
    # Calculate pi
    pi = np.sum(h, axis=0)
    pi = pi / np.sum(pi)
    return pi
def norm_ll(h):
    # Normalize the likelihood
    h1 = []
    for row in h:
        row_min = np.min(row)
        row = np.exp(row - row_min)
        row = row/np.sum(row)
        row = row.tolist()
        h1.append(row)
    return h1

def calc_ll(all_traj, params, pi, num_mixture=3):
    # Calculate likelihood
    k = len(all_traj)
    h = np.empty((k, num_mixture))

    # Calculate the likelihood of each trajectory being generated by each chain
    # Do this for all trajectories given the chains one by one
    for i in range(num_mixture):
        (prior, chain) = params[i]
        ll = 0
        for m, traj in enumerate(all_traj):
            for n, visit in enumerate(traj):
                if (n == 0):
                    ll = np.log(prior[visit]) + np.log(pi[i])
                    last = visit
                else:
                    ll += np.log(chain[last, visit])
                    last = visit
        h[m][i] = ll
    return h
# Given a sorted list of prospective trajectories, calculate the best fitting markov model
# I have to do this one for each chain

def maximize(traject_sort, h, num_hs=9, num_mixture=3):  # Finds the max likelihood.
    
    params = {}
    priors = None
    chain = None
    for i in range(num_mixture):
        hh = h[:, i]
        priors = calc_priors(all_traject, hh)
        chain = calc_chain(all_traject, hh)
        params[i] = (priors, chain)
    return params


def calc_priors(all_traject, hh, num_hs=9):  # Calculate prior probabilities
    priors = np.zeros(num_hs)
    for n, traject in enumerate(all_traject):
        entry = traject[0]
        priors[entry] += hh[n]
    priors = add_prior_noise(priors)
    priors = priors / np.sum(priors)
    return priors


def calc_chain(all_traject, hh, num_hs=9):  # Calculate chain
    chain = np.zeros((num_hs, num_hs))
    for n, atraject in enumerate(all_traject):
        for m, val in enumerate(atraject):
            if (m == 0):
                last = val
            else:
                chain[last, val] += hh[n]
                last = val
    chain = add_chain_noise(chain)
    chain = chain / np.sum(chain, axis=1)
    return chain
def add_prior_noise(prior, num_hs=9):
    N = np.sum(prior) * .02 / num_hs
    for i in range(num_hs):
        prior[i] = prior[i] + N * np.random.random()
    return prior

def add_chain_noise(chain, num_hs=9, num_mixture=3):
    for k in range(num_mixture):
        total = np.sum(chain) * .02 / (num_hs * num_hs)
        for i in range(num_hs):
            for j in range(num_hs):
                chain[i][j] += chain[i][j] + np.random.random() * total
    return chain

h = initial(all_traject)
pi = calc_pi(h)
ll=[]
for i in range(10000):
    # Improve the model parameters with new data
    params = maximize(all_traject, h)
    # Calculate log-likelihood
    l, h = expectation(all_traject, params, pi)
    pi = calc_pi(h)
    ll.append(l)
Urban Knowledge Discovery: How and Why, with a Focus on Social Media Data

Academic Paper as supplement to the thesis document

COSEM Scientific Paper

Master of Science in Complex System Engineering and Management

at Delft University of Technology

Deniz Özağaç

26 November 2018
Urban Knowledge Discovery: How and Why with a Focus on Social Media Data

Abstract

This paper focuses on the topic of knowledge discovery in an urban context. Complex systems such as cities generate data in numerous ways. In so many ways in fact that it may be problematic to keep track of these various data sources and the attributes of the data that the output. Which leads to the first main question of this paper. “What are all the data sources that are available to understand a city?”. In order to answer it, a literature review is conducted. Multiple divisions and descriptions are found regarding what data types are available and how can they be of help. Then, the paper leaves the generalist approach behind and focuses on a specific type of data for urban context: social media data. Here, three major social media data sources are inspected regarding their extent of the user base, researcher-friendliness, API functionalities and limitations. Finally, a discussion is had on the impact of privacy concerns and data quality.

Key words: Urban discovery, knowledge discovery, Urban informatics

Preface

The knowledge generated within this paper is shared with and master thesis project at the Delft University of Technology. As required by the Complex System Engineering and Management program, students are expected to prove academic skills and capability of performing research that is up to academic standards. This requirement of graduation also entails formulating a scientific paper that demonstrates the mentioned feats of the graduation candidate. Therefore, this scientific paper aims to demonstrate the academic skills that were acquired during the study and contribute to the scientific collection of knowledge in urban knowledge discovery.
1. Introduction

As of 2018, approximately %55 of the world’s population lives in cities. This high percentage has not always been the case, in 1950’s this number was around %30. Considering the world population also grew dramatically in the meantime, our cities are host to the most crowded populations in their history. Moreover, this percentage has been increasing and the United Nations reports estimate this ratio to reach %68 by 2050 (World urbanization prospects, 2018). As the cities grow in size and complexity, in order to be able to attend to the needs of increasing number of inhabitants, the more robust and complex systems need to be implemented in urban environments which gets progressively more difficult to manage. This is where the data become critical to sense, gather information about, understand, discover and solve problems within these complex systems. As the number of inhabitants rises and new technologies and services emerge, the amount of data generated outpaced even the growth of the number of inhabitants. Aside from the increasing amount of data, new novel data types and sources also got introduced to the problem solving and research cycle over time. With these changes in the last decades, it is now not too ambitious to claim that cities are knowledge veins that are available to various mining operations. This discipline and organized effort of discovering interesting and previously unknown knowledge from various sources of information sources with the goal of improving efficiency has been named as knowledge discovery, where data mining is a step in the process (Martin et al., 1999).

With that said, this paper first focuses on the earlier steps of this process where the question of “Which data sources are available and valuable for the purpose of gaining understanding about urban areas and cities?” (RQ1) and “What is possible to do with this data?” (RQ 2?) exists. First, these questions are answered by a literature review. Moving forward, a specific type of data -geolocational data acquired from social media services- is brought under the spotlight and three major social media platforms are compared with each other in the context of usefulness for urban research.

2. Literature review

This literature review is conducted with the goal of (1) acquiring an understanding of what data sources are available and (2) what is possible to do with this data with the purpose of generating value for improving our understanding of urban areas, cities and their inhabitants. The initial expectation is that there will be numerous data sources and use cases that have been addressed or used by the papers in the literature and this literature review will serve as an aggregation of them. It should be noted that the literature review does not go into the smart cities topic due to the noise that exists as a result of the topic’s popularity and how broad the return is for the key word “smart city”. Therefore, the following search terms are used in conjunction with the Google Scholar search engine: urban data discovery, urban knowledge discovery, urban informatics.
2.1. A review of available data sources

Le-Phuoc et al. (2011) initiate the discussion by stating that the vast amount of data that has been produced by various sensors in our daily lives need to be utilized to help smart cities make quick and smart decisions. These various sensors are further discussed by the authors and they have been categorized into three groups as physical sensors, mobile and wearable sensors and virtual sensors. They are described as following.

**Physical Sensors:** Sensors that are installed by the city administrations with the aim of observing the status of the infrastructures within the city. Few examples are traffic congestion level, public transportation status, air quality, trash level, temperature, water, waste, energy, weather etc.

**Mobile and Wearable Sensors:** Almost every modern mobile device and car act as sensors that provide information about their users. Other wearable sensors also exist, mainly for healthcare purposes.

**Virtual Sensors (Social Media Data Streams):** Social media data provides up-to-date information about the events within a city and also may provide insights into the infrastructure of the city. These include disturbance situations such as fires, traffic jams, protests etc. Twitter feeds are a prime example. Reliability of these data sources is the main concern in their use.

While covering a respectable segment of the data that is used, these data sources are also few of the most well-known and utilized data sources. It is expected that there should be numerous novel sources of data other than the mentioned ones above.

A paper that has great value for the purpose of this literature review is from Thakuriah et al. Related to the last statement, their discussion starts by underlining an observation that states the discussion regarding big data sources related to cities is often limited to IoT(internet of things) or social media data. However, there are other potentially useful data sources such as transactions data, governmental administrative data, data from arts and humanities collections and so on. With that said, Thakuriah et al. (2016) also provide an extensive table (table 1) where various data sources -including some novel sources- for urban research is categorized with respect to their associated user groups.
Table 1: Types of Urban big data and their associated users

<table>
<thead>
<tr>
<th>Urban Big Data</th>
<th>Examples</th>
<th>Illustrative user communities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensor systems (infrastructure-based or moving object</td>
<td>Environmental, water, transportation, building management sensory systems; connected systems;</td>
<td>Public and private urban operations and management organizations, independent ICT developers,</td>
</tr>
<tr>
<td>sensors)</td>
<td>Internet of Things</td>
<td>researchers in the engineering sciences</td>
</tr>
<tr>
<td>User-Generated Content (“social” or “human” sensors)</td>
<td>Participatory sensing systems, citizen science projects, social media, web use, GPS, online social</td>
<td>Private businesses, customer/client-focused public organizations, independent developers,</td>
</tr>
<tr>
<td></td>
<td>networks and other socially generated data</td>
<td>researchers in data sciences and urban social sciences</td>
</tr>
<tr>
<td>Administrative (governmental data) (open and confidential</td>
<td>Open administrative data on transactions, taxes and revenue, payments and registrations;</td>
<td>Open data: innovators, civic hackers, researchers</td>
</tr>
<tr>
<td>micro-data)</td>
<td>confidential person-level micro-data on employment, health, welfare payments, education records</td>
<td>Confidential data: government data agencies, urban social scientists involved in economic and</td>
</tr>
<tr>
<td></td>
<td></td>
<td>social policy research, public health and medical researchers</td>
</tr>
<tr>
<td>Private Sector Data (customer and transactions records)</td>
<td>Customer transactions data from store cards and business records; fleet management systems;</td>
<td>Private businesses, public agencies, independent developers, researchers in data sciences and</td>
</tr>
<tr>
<td></td>
<td>customer profile data from application forms; usage data from utilities and financial institutions;</td>
<td>urban social sciences</td>
</tr>
<tr>
<td></td>
<td>product purchases and terms of service agreements</td>
<td>Urban design community, historical, art, architecture and digital humanities organizations,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>community organizations, data scientists and developers, private organizations</td>
</tr>
<tr>
<td>Arts and Humanities Data</td>
<td>Repositories of text, images, sound recordings, linguistic data, film, art and material culture,</td>
<td>Urban planning and social policy community, government data organizations, private businesses</td>
</tr>
<tr>
<td></td>
<td>and digital objects, and other media</td>
<td>and consultants</td>
</tr>
<tr>
<td>Hybrid data (linked and synthetic data)</td>
<td>Linked data including survey-sensor, census-administrative records</td>
<td>Urban planning and social policy community, government data organizations, private businesses</td>
</tr>
</tbody>
</table>

This table by Thakuriah et al. (2016) could potentially serve as a solid base to build on as new data sources are explored. A data type that has not been explicitly covered by the table, however, is the geographical top-down view of cities or other various map type visuals. Gil et al. (2009) focus on this data type, covering the approach, application and evaluation of it for the use of typomorphology efforts by presenting a method. The paper ultimately focuses on this method which is useful for creating urban form typologies that are derived from the local characteristics of a selected place. By doing so, they demonstrate the potential that the data type has for improving our understanding of urban areas.
2.2. A review of use cases

With the assumption that the first part of the literature review provides a base understanding of what various data types are available for urban research, the next thing to get curious about is what to do with this data. If we take one more step back, a good place to start is to ask, “What is there to understand?”. The general answer is: anything that is a part of what is creating the complex urban environment such as: transportation, urban mobility, health, economy, environment, housing, social fabric, political fabric, semantics of the populations, flow of materials and energy, land use, green space, location and nature of pollution (Cunningham & Verbraeck, 2018) etc. The effort of improving these systems goes back probably as far as the first city. However, nowadays there is a name for this collective effort. The emerging discipline of Urban Informatics focuses on exploring and understanding urban systems by utilizing existing and more importantly, unusual sources of data. As refined by Thakuriah et al. (2016) the aim of this new discipline is to help urban systems with issues such as resource management, urban engagement and civic participation, knowledge discovery and policy analysis. Therefore, urban informatics has the potential to benefit urban science and decision makers in mainly four areas as defined by Thakuriah et al.:

1) Urban resource management: Developing plans and strategies for making more efficient use of resources and managing them smarter or improving said existing strategies.

2) Knowledge discovery about urban patterns and processes: Generating theoretical insights about the patterns and relationships between urban processes.

3) Urban engagement and civic participation: Developing strategies, technologies with the aim of keeping the citizens informed the best and incentivizing their involvement in social and civic matters.

4) Urban management, planning and policy analysis: Developing innovations, solutions and improvements for practices such as service delivery, generation of new policies and for the management and maintenance of infrastructure.

Other approaches also exist for categorizing the potential benefits of various data types. A paper from Naaman (2011) categorizes the relevant uses of VGI (Volunteered geographic information) and SAS (Social Awareness Streams) in 4 points. These points are fundamentally different answers to the question “What value can the analysis of VGI and SAS data bring to the domain of city planning and urban science?”.

The four elements of a city in this context are defined to be districts, landmarks, paths and activities as inspired by Naarman (2011). With these 4 elements of the city at hand, the four main application fields that focus on each element respectively are 1) Boundary definition and detections (District identification) 2) Computation of attractions (Landmarks identification), 3) Derivation and recommendation of paths (Mobility flow detection), 4) Evaluation of activities, interests and temporal trends.

Boundary definition and detection: The knowledge of delineation of neighborhoods by various functionalities has multiple potential ways to be valuable. One of the first ideas that
come to mind is that these delineations can be used to understand the demographic mapping of the city area. Demographic mappings may be according to gender, age, education level, economic status etc. Once combined, this information could be used for understanding more complex constructs within the city such as property values by areas, fluctuations in these property values over time and tracking gentrification. A related is study from Cranshaw et al. (2012), where they call the neighborhood clusters they generate via social media data as “livehoods” and compare them to official neighborhood delineations to question their functional appropriateness.

**Computation of attractions:** Milgram’s work in “Psychological Maps of Paris” (Milgram, S., 1977) aimed to understand the image of the city of Paris from the perspective of its residents by individually asking them to draw a map of Paris from on top of their head. As expected, the locals drew their maps based on the attraction points that were most relevant to their lifestyle. This study helped in understanding the image of Paris from the perspective of its locals. With the available technologies nowadays, researchers do not need to individually ask people what the most prominent locations of the city are. Instead, they can track their movements and view their preferences on their social media accounts. Similar to Milgram’s work, the benefits lie in understanding how the locations in cities are used and perceived by the locals. This understanding is useful for many purposes such as acquiring feedback in existing infrastructure and creating requirements for future designs within city space.

**Derivation and recommendation of paths:** Derivation and recommendation of paths have various purposes such as understanding tourist flows, running paths and other mobility preferences. Especially, the cities that are having problems with tourist populations to the degree that it interferes with the comfort of the locals (Such as Barcelona and Amsterdam) would be interested in understanding the paths that tourists and the locals tend to traverse. If these paths are understood, crowds can be managed to steer away from each other. Moreover, via the data provided from mobile apps such as the ones for running & biking, researchers can develop a better grasp of how the paths & parks of the cities are used by the residents.

**Evaluation of activities, interests and temporal trends:** There are numerous applications available that focus on detecting and reacting to different types of events. Certain social media platforms such as Twitter are especially useful for this purpose due to how they function in spreading information between users. Some application examples are “Crowdsense” for natural disaster and crisis situations (“Crowdsense.co - researching technology for analyzing social media in times of crises,” n.d.), “Event Alert” for school shootings and similar threats in USA (“Event Alert – ELERTS Corp,” n.d.) and “CrowdTrack” for tracking when and where urban crowds tend to gather (“CrowdTrack,” n.d.). These and similar applications usually focus on certain regions and event types. While some event detection notification services are syntax based such as SAS (Le-Phuoc et al., 2011) and WSN (Li & Jacobsen, 2005) where some are more advanced, semantic-based such as ACEIS (Gao et al., 2014).

Main obvious benefits regarding the evaluation of events & activities and temporal trends include improved crisis management, community forming, event tracking, anomaly detection, crowd management and city service optimization (Ex: for commuters). Aside from the direct benefit provided to the individuals and the organizations, tracking of such events in data logs
would create valuable datasets that may be analyzed to understand the dynamics and rhythm of a city and its population.

It is clear that numerous data types are available for multiple purposes when it comes to performing research on urban areas to improve and understand them. There even is an extensive list of existing data solutions for urban systems created by Harvard University called “Catalog of Civic Data Use Cases” (“Catalog of Civic Data Use Cases,” n.d.). The data solutions here are divided into four categories as the Health & humanity services, infrastructure, public safety and regulations. On top of existing examples, plenty of future research questions are also collected here which is quite valuable as a resource since demonstrates what kind of questions are answerable with data to researchers (data scientist or not). After all, asking the right question to the data at hand is just as important as having clean and complete data(Mennis & Guo, 2009).

Within these possibilities, I find a certain data type particularly interesting due to its merging use cases and accessibility while being generated by the inhabitants themselves. The next chapter focuses on this data type: geotagged social media data.

3. An evaluation of SAS data sources

Mainly within the last two decades, we have witnessed a new classed of communication and information platforms called social awareness streams (SAS) emerge (Naaman 2011). The main characteristics of these platforms such as Facebook, Twitter, Instagram, Foursquare, Flicker and others is that they allow the users to post relatively small and uniformized content in the form of pictures, videos or text. Thanks to the devices that these services are being used on, most of these posts or activities come up paired up with the exact location of the user. This creation of “geotagged” data is likely to increase as cellphones keep becoming more accessible and these services become a fundamental part of society. With such great volumes, SAS data offers a unique opportunity to understand citizens with regard to their behaviors, opinions and interests (Naaman 2011). Goodchild (2006) even calls the citizens as sensors of their environments.

Social media data provides information about the fabric of the cities from the perspective of their residents, at least the ones who are the users of the chosen platform. Therefore, they are valuable for the researchers, especially when the data is “geotagged”. This explains the plethora of research that has been performed with and about Foursquare and Twitter data. A third platform, which is just as, and even more popular than these two in 2018 is Instagram which is not used nearly as much as other platforms in the context of academic research. This is mainly due to the fact that images are more difficult to store, categorize and analyze than the data in text form. However, due to the tremendous increase in popularity, it now deserves the same attention that Twitter and Foursquare got from researchers (Hu et al., 2014). Improvements in readily available image processing and analysis libraries also contribute to the fact that Instagram should get more attention from the researchers more than ever. Facebook and LinkedIn also deserve a mention here as they have great user populations and latter offers insights to a niche area. However, they are not included in this comparison since they do not provide “geotagged” data due to their structure and functionality.
To summarize, this section focuses on assessing of three social media data sources that have geotagged components - Twitter, Foursquare and Instagram- with regards to how useful they are to researchers in the context of urban science. Individual assessment is followed by a comparison of three by their overall usefulness for urban science research.

### 3.1. Foursquare

Foursquare has been one of the favorite social media platforms of the researchers in this decade. There are numerous many studies who have used Foursquare data in some way to understand urban fabric and the inhabitants (Noulas et al., 2013; Silva et al., 2013; Widmer, 2015; Quercia & Saez, 2014; Aubrecht et al., 2011). However, most of the research is in the time window of 2011-2013 where Foursquare had its prime with regards to its popularity and which has been diminished drastically since.

While the user activity is quite low recently, Foursquare is a good source to gain insights about the venues in various urban areas. A “venue” in this context is any sort of establishment that a user can check-in to via their Foursquare account, giving away their current location to their network. This location data is paired with the data that is associated with the said user such as user’s unique id, name, gender, hometown, friends, and other check-ins and so on, depending on their privacy settings. One advantage for the researchers is that since Foursquare’s main functionality is based on the location element of the check-in, geolocational data is always available whenever collecting data from Foursquare API unlike Instagram and Twitter where it is optional for the users to share their location with their network.

Currently, the best use of the Foursquare API would be to use it to acquire comprehensive data about the venues of an urban area (A role that may be challenged by another platform: TripAdvisor). In order to collect the said data, Özağaç (2018) recommends using Foursquare API’s “explore” functionality. The functionality allows the researcher to collect a set of venue recommendations for a city with limitations of 50 per page(limit in API) and for 50 pages(offset in API). This results in a maximum of 2500 venues to be recommended, which contains the venue’s name, id, location in latitude and longitude and type which one of food, drinks, coffee, shops, arts, outdoors, sights. 2500 venues are expected to be enough to address all the venues in most cities. For reference, once this technique was applied for the city of Rotterdam, the return was only 487 unique venues (Ozagac,2018). It is also worth mentioning here that not all venues have Foursquare accounts for them or have users checking in at them, especially now that the popularity of the app has diminished.

The Foursquare API does not give away all the user’s check ins within a city. However, there is a workaround to get the check-in data of all users from a selected city. This is due to the fact that the API allows the collection of check-in data that belongs to a venue when a venue id is given. Therefore, it is possible to take the user ids collected from the method in the last paragraph and iteratively collect check-ins for all of the 2500 venues which the ids are known. This will result in all the check-in data that has been associated with the 2500 recommended venues within a city.
Taking a step back from this workaround, the previous collection of venue data may prove useful if the researcher utilizes the collected data in a way to generate insights from multiple datasets. If used in a complementary way to other data sources such as Twitter or TripAdvisor, Foursquare venue data may the potential to prove useful still.

### 3.2 Instagram

Despite being the most popular photo sharing app ("Global social media ranking 2018 | Statistic," 2018), Instagram does not receive as much attention as it deserves from the researchers (Hu et al., 2014). While there are some papers that are pointed towards the meeting point between Instagram and urban science (Silva et al., 2013; Guerrero et al., 2016; Hochman & Manovich, 2013), their numbers are very limited compared to the other social media platforms.

One reason to why this is the case that may be the act the image data -which Instagram is mainly based on- is more difficult to store and analyze than text data. This may result in a knowledge barrier to the researchers who want to study them since a greater understanding of data operations and a knowledge of image-based data analysis is required. However, if one can manage to move past these barriers, analysis of Instagram data has the potential to reveal valuable insights about social and cultural fabric of urban areas that text-based data is not able to reveal. As Hu et al., also accentuates, “A picture is worth thousand words”.

Unfortunately, the challenges of using Instagram data are not only limited to knowledge barriers, the API's functionalities have also been diminishing lately. The front page of the API's website has the note.

“To continuously improve Instagram users' privacy and security, we are accelerating the deprecation of Instagram API Platform, making the following changes effective immediately. We understand that this may affect your business or services, and we appreciate your support in keeping our platform secure."

Followed by a link to certain capabilities that will be disabled permanently. An example is that it is no longer possible to collect all of the Instagram posts that originate from a specific area such as a City. However, there is a workaround to achieve this functionality as mentioned by Özağaç (2018). There, he mentions an example of use of such data from a reddit post where Instagram posts of Istanbul are plotted over the map of the city by their latitude and longitudes ("Instagram Posts of Istanbul [OC] r/dataisbeautiful," 2018). The visualization is achieved by utilizing a dataset that does not seem to be available with the current functionalities with the API, yet it exists.

In order to collect such data, a specific functionality of the Instagram API is used which allows the collection of data of all activity(posts) that originate from within 750-meter radius of a certain given point. The difficult step is to cover the entire area that the researcher is interested in with these 750-meter radius circles, which is what the creators of the Istanbul visualization has done. This is no task for a researcher without an advanced level of coding skills. This further increases the knowledge barrier that is in front of the researchers who want to utilize Instagram data for the context of urban science.
While data acquired Instagram may be of great value to urban science, researchers should definitely hurry to extract that value or at least to collect the data. As the front page of the Instagram API indicates, the functionalities will keep diminishing due to the data privacy concerns. The current API change schedule states that fundamental functionalities such as accessing the public content and the basic profile info and media of the user will be disabled as of early 2020 and December of 2018 in this order ("Instagram Developer Documentation," n.d.). With that said, Instagram API seems to be on a track that aims to satisfy the benefits of the businesses more and researchers less. However, once the changes are implemented and a semi-stable set of functions are complying to the data privacy concerns are achieved, this debate will need to be reinitiated.

### 3.3 Twitter

Twitter is likely the single most broadly used social media platform for that is used by researchers. No other social media platform has attracted as much attention as Twitter from the academia. There are countless studies in with various backgrounds in literature that uses Twitter data for analysis that are mainly based on network theory, natural language processing and social behavior theories. While Twitter is not as popular as it used to be, it is still one of the top 10 most popular social media platforms. There is also a vast collection of web-based tools that specialize in working with Twitter data. A collection of these tools can be found on Ahmed’s blog of 2017. ("Using Twitter as a data source: an overview of social media research tools (updated for 2017)," 2017)

The main reason to why Twitter is preferred so much for academic purposes could be given to the fact that tweets are limited to 140 characters which makes them easy to work with various data tools. Aside from this design choice, Twitter is also a knowledge sharing platform by purpose that also notates the network of connections between people by the unidirectional follower system. Adding the fact that activity can also be geotagged (Location of the activity is available if the user has their settings enabling it.) is the final icing on the cake for the researchers who are interested in studying the static and dynamic behavior of information and knowledge within communities.

For the purpose of collecting data on a big scale such as for an entire City, the Streaming API of Twitter is the most popular way. The Streaming API provides tweets in real-time as they are being posted and is filtered on search terms that could be keywords or location. Filtering tweets by location most likely the preferred choice if the aim is to collects tweets from a city. The location filter is defined via a bounding box, which is an area limited by four latitude and longitude lines. For cities which are difficult to cover with a rectangular shape, covering the desired area with multiple bounding boxes is a reliable option.

Now that the three social media platforms are covered with respect to their API generosity, popularity and other attributes, it is timely to aggregate these statements. The next section provides a comparison of these platforms in respect to their usefulness in the context of urban science and urban research.
3.4 Comparison of data sources

After evaluating the three social media services with respect to what they have to offer for the researchers, it is now possible to make a comparison between these platforms on even ground. The results of this comparison are summarized in table 2 below.

<table>
<thead>
<tr>
<th></th>
<th>Foursquare</th>
<th>Instagram</th>
<th>Twitter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Main Data type</td>
<td>Text + geolocational</td>
<td>Photos &amp; images</td>
<td>Text</td>
</tr>
<tr>
<td>API generosity</td>
<td>Medium, diminishing</td>
<td>Low, rapidly diminishing</td>
<td>High, diminishing</td>
</tr>
<tr>
<td>Popularity</td>
<td>Low, past its prime.</td>
<td>Very high, on the rise</td>
<td>Medium-high, stable</td>
</tr>
<tr>
<td>Knowledge barrier</td>
<td>API usage and low-intermediate coding skills.</td>
<td>API usage and high-intermediate coding skills for workarounds, working with image based data.</td>
<td>API usage and some coding. Many tools also exist.</td>
</tr>
</tbody>
</table>

When I asked my younger colleagues (early 20’s) what did they think about Foursquare, their reaction was often in the lines of “What is Foursquare?”, which I believe expresses the main problem of Foursquare quite clearly. There simply is not enough user activity on most places to gain an understanding of human behavior on a population scale. Foursquare had its prime time around 2011-2013 time-window and now has fallen from popularity. Nevertheless, Foursquare API may still prove useful if used for the purpose of collecting information about the number of, types and locations of venues within a city.

Due to its immense popularity, Instagram data may arguably have the most untapped value out of the three, however, it’s API is more difficult to work with because (1) there are major functionalities that has been removed from the API and (2) In order to get the equivalent data that is easily collectable from Twitter API, one has to come up with workaround that require respectable coding skills. The second point in combination with the fact that Instagram data is image based, poses a knowledge barrier for researchers who wish to start using Instagram data for their research.

Twitter seems to be deserving of being the go-to data source for working with cities using social media data. As seen in the table, Twitter offers text-based data with a generous API while being still relevant to the society with its respectable user base. On top of these attributes, because of its popularity in the academic communities for research purposes, there are multiple third-party tools, comprehensive documentation and numerous up-to-date tutorials are available for researchers who wish to start using Twitter data for their research.

During the process of accumulating knowledge about the user bases, API’s and recent changes around the social media services above, I have realized some phenomena that did not receive the attention that it may be deserving from the literature. In the following section,
these points are brought onto the table for discussion as they are relevant to the process of collecting social media data for urban research purposes.

5. Discussion

5.1 Data quality

As social media became a critical part of our lives, businesses and organizations also caught on and started actively using them. This situation is something that the researchers should be aware of when collecting and analyzing social media data with the purpose of understanding the population in an urban area. Social media accounts that belong to organizations and businesses should be treated differently from the personal accounts due to multiple reasons. First, organization accounts often have more than one individual behind them which results in them showing activity levels that is not possible for an individual to constantly output. This also means that organization accounts may end up having more representation depending on the analysis that is being performed. Özağaç (2018) mentions that only 6 of the top 30 active tweeters in their Rotterdam tweet dataset were likely to be personal accounts judging by their usernames. Second, because they are not accounts belonging to individuals, they will not display behavior and patterns that is representative of the behavior of an individual. However, they may be analyzed as if they were if the researchers fail to classify them or pick them apart. In this situation, they may act as if they were anomalies and increase the noise that is attributed with the data that is to be used for behavioral analysis.

Hu et al. (2014) handles this problem by classifying “Regular active” users. They define the regular active user as someone that has at least 30 followers, 30 friends and 60 pictures (instance of posts) and who are not organizations, businesses, spammers or bots. An adaptation of this approach to the social media platform at hand is highly recommended for the researchers if their main interest is understanding the human behavior in an urban area.

5.2 Diminishing API functionalities:

One phenomenon that is shared between all three of the social media platforms above is that all of their APIs have key functionalities that has been or are being depreciated. This is very likely due to the heated data privacy debates regarding Facebook following the USA’s presidential election of 2016, where a data analysis company accessed user data in controversial ways to provide advantage to the candidates (Harris, 2018). Performing perception manipulation on a respectable portion of the population is quite questionable in an ethical sense. However, the main problem that is relevant to the topic at hand is that private data of hundreds of thousands have been collected and used against them without their explicit consent.

Naturally, once discovered this situation triggered a chain of discussions and many decision and policy makers had their eyes open regarding the data privacy subject, which resulted in
appearance of laws and regulations and tightening of the existing ones in a way to protect personal data like never before. Alongside its many effects, (including encountering privacy policy updates and notifications on every website and service that you are using.) APIs of social media platforms were also forced to remove major functionalities that were available for the researchers and businesses to use alike.

The situation certainly makes it harder for the researchers. However, this shift that is caused by realization of importance of data privacy is still an ongoing process. Instagram and Twitter APIs for example are still in the process of removing and modifying functionalities. Once the dust is settled (in a few years possibly), it may be more possible to clearly evaluate the state of use of social media platforms for urban research purposes.

6. Conclusion

In this paper, two main questions are answered. First is the question of “What data types and sources are available for researchers who aim to understand and improve urban areas”. Second is concerned about how to make the said data useful and goes as “What is there to understand in urban areas”. The answers to these questions are collected under one roof via a literature review. As expected, the literature contained various answers to both of these questions, therefore, the literature review in this paper serves as an aggregation and organization of knowledge. Following that, the attention is shifted towards a specific data type: social media data with geolocational components. Here, three major social media data sources are analyzed with regards to their usefulness for the researchers who are interested in urban science. Individual analysis is assembled together as a resulting table which presents the comparison of these data sources for the purpose at hand. Finally, the discussion part pulls attention to two phenomenon that were discovered during the knowledge accumulation of this paper. Quality of data and privacy concerns are discussed in a way that is concerned about the usefulness of social media platforms for urban science researchers. Here, it is stated that researchers should be aware of the fact that businesses and organizations are also using social media platforms and generate data while doing so. Moreover, it seems to be the case that policies and law are also catching up with the data privacy concerns as of 2018, which is bound to have impact on the researchers as well. The following years will show how much social media users’ data will be available for researchers.

References

Center in Geography and Regional Planning of the Faculty of Social Sciences and Humanities of the Nova University of Lisbon.


