Ingenieurfakultät Bau Geo Umwelt
Technische Universität München

Mobility–related Information from Emerging Data sources: the case of Social Media

Emmanouil Chaniotakis

Vollständiger Abdruck der von der Ingenieurfakultät Bau Geo Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktor-Ingenieurs (Dr.-Ing.)

genehmigten Dissertation.

Vorsitzender:
Prof. Dr.-Ing. Klaus Bogenberger

Prüfende der Dissertation:
1. Prof. Dr. Constantinos Antoniou
2. Prof. Dr. Konstantinos Goulias,
   University of California Santa Barbara

Die Dissertation wurde am 19.09.2019 bei der Technischen Universität München eingereicht und durch die Ingenieurfakultät Bau Geo Umwelt am 12.03.2020 angenommen.

Abstract

The wide spread of Information and Communication Technologies is changing data availability with new data sources emerging. The size of datasets and their diversity has changed: these changes can be attributed to the evolution of pervasive systems and, especially, the Internet connectivity available to a growing number of individuals. The latest available data sources have been found to produce an immense volume of data that could be potentially used to improve transportation systems: first, in terms of identification and prediction and, second in terms of optimization. One of the most challenging to use, yet rich set of datasets lately explored in transportation, originates from Social Media platforms. The statistics of Social Media use are astonishing: social media websites, Facebook, and Twitter, are visited or used by millions on a daily basis. Since their emergence, Social Media have drawn attention from the scientific community, due to the unprecedented amount of information that can be extracted and the opportunities that Social Media platforms can provide regarding direct communication with users.

This dissertation provides empirical evidence of the potential of using Social Media in transportation and defines methods for the extraction of useful information. These methods include the exploration of the differences and similarities between Social Media and other transport related data collection methods, the investigation of people's activity spaces, data enrichment methods, and the people's activity identification . The analysis is based on various datasets including data collected from a selection of 10 cities in Europe and the USA. These datasets are analysed using various descriptive statistics to firstly examine aspects of transferability for Social Media data and secondly, extract the different users characteristics that have been observed. The analysis performed includes the classification of users posts, as well as the analysis of the potential fitting

of power-law distributions. The comparison is also extended to the investigation of the connection to conventional mobility data by empirically examining data from three popular Social Media and conventional travel-diary surveys data.

Both spatial and temporal aspects of the activities' representation are considered. Finally, an efficient framework for the inference of users activities from Social Media data following a user–centric approach is proposed. The framework is applied to data extracted from Twitter, combined with inferred data from Foursquare that contains information about the type of location visited. Users data are classified, allowing for the identification of commonly visited locations and data enrichment. Based on the known activities and the Social Media text, a set of classification algorithms is applied for the inference of activities of tweets. The findings of this research contribute towards the derivation of methodologies that can support the inference of meaningful travel information from Social Media, which could in turn improve transportation.

Zusammenfassung

Die weite Verbreitung der Informations- und Kommunikationstechnologie verändert die Datenverfügbarkeit und es entstehen neue Datenquellen. Die Größe der Datensätze und ihre Vielfalt haben sich verändert. Diese Veränderungen können auf die Entwicklung der allgegenwärtigen Systeme und insbesondere auf die Konnektivität zurückgeführt werden, die einer wachsenden Anzahl von Personen mit der Entwicklung des Internets zur Verfügung stand. Die in letzter Zeit verfügbaren Datenquellen haben zu einer immensen Datenmenge geführt, die potenziell zur Verbesserung der Verkehrssysteme genutzt werden könnte, erstens in Bezug auf die Identifizierung und Vorhersage und zweitens auf die Optimierung. Einer der anspruchsvollsten, aber dennoch reichhaltigsten Datensätze, die in letzter Zeit im Transportwesen erforscht wurden, stammt von Social Media Plattformen. Die Statistiken über die Nutzung von Social Media sind erstaunlich: Social Media Websites, Facebook und Twitter sind in den Besucherrankings mit Millionen von täglich aktiven Nutzern auf Platz eins. Seit ihrem Aufstieg haben Social Media aufgrund der beispiellosen Menge an Informationen, die extrahiert werden können, und der Möglichkeiten, die die Social Media-Plattform für die direkte Kommunikation mit den Nutzern bietet, die Aufmerksamkeit der wissenschaftlichen Gemeinschaft auf sich gezogen.

Diese Dissertation liefert empirische Belege für das Potenzial der Nutzung von Social Media im Transportwesen und definiert Methoden für die Extraktion nützlicher Informationen. Diese Methoden umfassen die Erforschung der Unterschiede und Gemeinsamkeiten zwischen Social Media und anderen verkehrsbezogenen Datenerhebungsmethoden, die Untersuchung von Aktivitätsräumen, die Datenanreicherung und die Aktivitätsidentifikation mit Social Media-Daten. Die Analyse basiert auf verschiedenen Datensätzen, darunter Daten aus einer Auswahl von 10 Städten in Europa und

den USA. Diese werden mit Hilfe verschiedener deskriptiver Statistiken analysiert, um zunächst Aspekte der Übertragbarkeit von Social Media-Daten zu untersuchen und dann die beobachteten Merkmale der verschiedenen Nutzer zu extrahieren. Die durchgeführte Analyse umfasst die Klassifizierung der Beiträge der Nutzer sowie die Analyse der möglichen Anpassung von kraftwerksrechtlichen Verteilungen. Der Vergleich wird auch auf die Untersuchung der Verbindung zu konventionellen Mobilitätsdaten ausgedehnt, indem empirisch Daten aus drei populären Social Media und konventionellen Reisetagebuchdaten untersucht werden.

Dabei werden sowohl räumliche als auch zeitliche Aspekte der Darstellung der Aktivitäten berücksichtigt. Schließlich wird ein effizienter Rahmen für die Ableitung der Aktivitäten der Nutzer aus Social Media-Daten nach einem benutzerzentrierten Ansatz vorgeschlagen. Das Framework wird auf Daten von Twitter angewendet, kombiniert mit abgeleiteten Daten von Foursquare, die Informationen über die Art des besuchten Ortes enthalten. Die Daten der Benutzer werden klassifiziert, so dass häufig besuchte Standorte identifiziert und Daten angereichert werden können. Basierend auf den bekannten Aktivitäten und dem Social Media Text wird ein Satz von Klassifikationsalgorithmen für die Ableitung von Aktivitäten von Tweets verwendet. Die Ergebnisse dieser Forschung tragen dazu bei, Methoden abzuleiten, die den Rückschluss von aussagekräftigen Reiseinformationen aus Social Media unterstützen können, was wiederum den Transport verbessern könnte.

Acknowledgments

This dissertation would not have been completed without the practical and intellectual help and support of my supervisor, Professor Constantinos Antoniou. His guidance and advises have been instrumental for the development of this work. I would also like to thank my supervisor on a personal level, as his attitude and his subtle way of advising have helped me evolve into being a better person and a better researcher.

I would also like to thank the professors I worked with, Prof. Konstantinos Goulias, Lecturer Loukas Dimitriou, Professor Bin Jiang, Professor Hani Mahmassani and Professor Francisco Camara Pereira, for their valuable guidance, feedback and stimulating discussions.

I am deeply grateful to all my former colleagues and friends that have helped me complete my dissertation. To begin with I would like to thank my former colleagues at the Hellenic Institute of Transport. Dr. Georgia Aifantopoulou has supported me since our first encounter, both by enabling my research with the provision of data and guidance, as well as inspiring me to pursue my dreams. I can never thank enough my dearest friends and colleagues, Iraklis Stamos and Dr. Josep Maria Salanova Grau, who have believed in me and have been there for me for more than 5 years. I am also grateful to all the fellow-researchers at TUM. Additionally, I would like to thank all my friends that have supported me in pursuing this thesis. With apologies to those that are not mentioned, I am gradeful to Dr. Georgios Birpoutsoukis, Christos Bafatakis, Evangelos Komninakidis and Dr. Hariton Efstathiades.

Last but not least, I would like to thank my wife Zoi, my daughter Irini, my parents Dionysis and Soula and my sister Maria for their support and understanding of all the endless days and nights that I was not with them.

Table of Contents

List of Figures

List of Tables

1 Introduction

Contents

1.1 Motivation

The widespread use of Information and Communication Technologies is changing data availability in a broad spectrum of applications creating new data sources to be explored. The size of datasets have changed, as has the diversity of the available datasets. These changes in data availability can be attributed to the evolution of pervasive systems (i.e. GPS handsets, cellular networks) and especially the connectivity that has been available to a growing number of individuals with the evolution of the Internet. The deployment of these systems has received wide attention from the transport scientific community towards the utilization of the increasingly available data (Big Data) [Buckley and Lightman, 2015, Reades et al., 2007]. The lately available data sources have been found to produce an immense volume of data that could potentially be used to improve mobility, first in terms of identification and prediction and second in terms of optimization.

Data sources can be either be actively generated (by sensors deployed in order to periodically measure a particular phenomenon, such as weather data) or passively collected (e.g. as social media data). Efforts in working with this growing volume of data have been directed towards all aspects of the Big Data Life Cycle [data acquisition, information extraction and cleaning, data integration, aggregation and representation, modelling analysis and interpretation; see Jagadish et al., 2014] constituting a rather multidisciplinary research topic.

In transportation, these efforts have been mainly focusing on the aspects of data acquisition –mostly in terms of data collection–, information extraction and cleaning and modelling analysis. The analyses most commonly performed are based –to name but a few– on Floating Car Data [Astarita et al., 2019], mobile phone data [Huang et al., 2018, Wang et al., 2018, Zhou et al., 2018], payment and transit card data [Sulis et al., 2018, Yap et al., 2018, Utsunomiya et al., 2006], GPS enabled mobile phone data [Bachir et al., 2019, Bwambale et al., 2017] and social media [Chaniotakis and Antoniou, 2015, Zheng et al., 2016]. Of particular interest in regard to the increased data availability is the evolution of pervasive systems (i.e., GPS handsets, cellular networks) and especially the connectivity that has been available to a growing number of individuals that allows the sharing of different information types such as spatial, temporal, and textual information.

One of the most challenging, yet rich set of data lately explored in mobility research comprise data originating from Social Media platforms. The statistics of Social Media use are astonishing: social media websites, Facebook, and Twitter, are ranked as third and eleventh most visited websites globally (www.Alexa.com). In 2018, there were 100 million daily active users on Twitter sending 500 million tweets every day (www.omnicoreagency.com); while there are around 1.47 billion daily active users on Facebook, 88% of them access the site via mobile phones (blog.hootsuite.com). A growing amount of related work has been published in the last few years, showcasing the potential of using Social Media in transportation. In short, the directions that the literature takes include the following uses of Social Media : modelling and forecasting purposes (e.g. OD Estimation, Attraction Models, activity modelling), extraction of

mobility–related and spatial characteristics, transportation-related sentiment analysis, prediction and event detection, and accessibility analysis with the complementary use of Twitter data. In addition, the use of Social Media, for direct communication of end-users, is another central research direction. Such usages are oriented towards public engagement and for information sharing, by transport providers.

The exploration of generally emerging data sources and particularly social media, for the extraction of transportation–relevant information is becoming timelier than ever. Research needs to focus on the exploration of the descriptive quantities that provide insights on the information that can be extracted, but also takes into account differences between different places around the world. Additionally, the exploration of the types of information that can be extracted should be examined based on the characteristics of the social media data. Such an endeavour should be well positioned in an information-based cycle that allows the further uptake of the findings from the various methods for transportation.

1.2 Information Extraction and Transportation

In an attempt to provide a generalized definition of the multi-source fusion problem Bloch [1996] define information fusion as "the study of efficient methods for automatically or semi-automatically transforming information from different sources and different points in time into a representation that provides effective support for human or automated decision making".

The multi-source transportation demand modelling can be defined on the premises of efficient use of information to increase the accuracy of transportation models. The inclusion of diverse data can be examined at one of the transportation modelling levels, based on its efficiency. Conventionally, use of information for mobility purposes has been based on the data that was or could be available. Data collection was an endeavour taking place on a need basis (e.g. travel surveys) and it was followed by data processing and data modelling on the basis of theories. The availability of other, in some cases seemingly unconnected datasets, is introduced as a newly emerging field of research focusing on the collection of other sources of data, which is then examined on the premises of information extraction, i.e. feature extraction, object refinement and data modelling. These two cycles should be brought together to firstly improve the conventional data collection and modelling and secondly provide insights that were impossible to get a few years ago.

In order to do so, several steps should be followed to allow for data of different typologies (non-commensurate data) to be used in transportation research. These include data collection, information extraction, object refinement, feature definition, and emerging data modelling, to reach the operationalization of the use of these sources of data 1.1.

Performing these steps, requires the adequate understanding of the data, and its mapping in existing mobility–related features. Thus, methods are needed that first

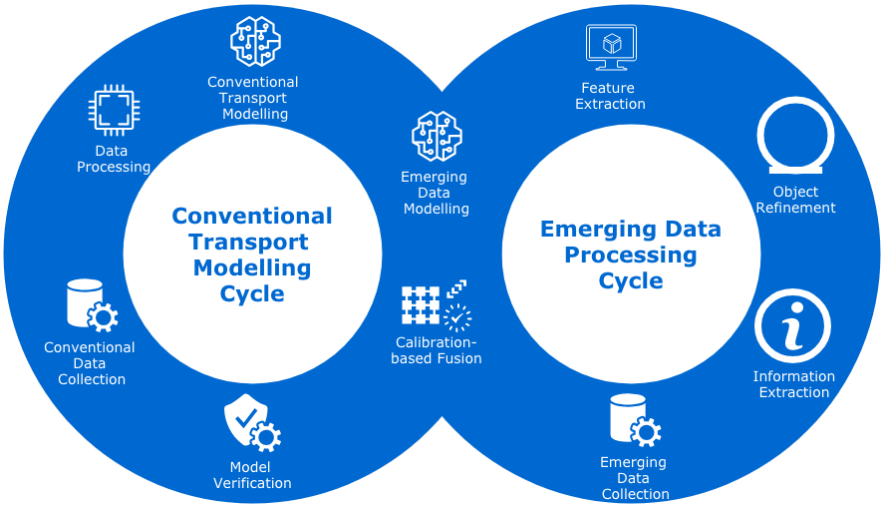


Figure 1.1: Incorporation of emerging data sources for modelling mobility

explore the data itself, its typology and the connection to the examined system and second enable the extraction of mobility–related information (and features).

1.3 Problem Definition and Thesis Objectives

In order to accommodate the foreseen and much needed improvements related to the transportation system, this new stream of data should be first transformed to information [Kitchin, 2014] in an efficient way. In this direction, one of the main issues identified within this research concerning the use of Big Data in transportation is the rather restrictive environment (in terms of the modelling scale, data used or methods employed) in which pertinent research revolves. Researchers in most cases focus on one data source using traditional or machine learning methods to extract partial information, thus using in a sub-optimal way the data available. Although these exploratory efforts aim at showcasing the merits of using Big Data in transportation and their limitations can be attributed to the challenges that data use imposes, it is believed that this could disorient the efforts on combining various data-sources in an efficient way, only allowing for limited use of the available data.

In this context, this dissertation focuses on the use of Big Data in transportation analytics by defining an efficient methodological framework for exploiting information that can be extracted from Big Data sources in regards to transportation modelling and prediction. The methodological framework is oriented towards the combined exploration of diverse data sources and the implementation of domain based feature selection methods. The main research question is:

What is the methodological framework including the pertinent methodologies and algorithms that could be used to enable the adequate and efficient use of diverse data sources in transportation systems analysis, modelling and prediction?

The above main research question elicits a set of objectives that are foreseen to guide the work required to be performed:

- Identify and review related work of use of diverse data for modelling
- Derive the requirements and specifications of related methodologies
- Design the methodologies to be implemented in transportation related problems
- Verify the methodologies applicability
- Validate the derived methodologies

1.4 Thesis Contributions

This thesis contributions are distinguished on theoretical and methodological, and practical levels.

1. **Theoretical and Methodological Contributions**
   a) *Synthesis on Transport Data and Social Media Research* In Chapter 2, a review of the pertinent work on the use of conventional end emerging data is presented focusing on the possible sources of biases and errors. Additionally, a comprehensive synthesis of the Social Media research is presented.
   b) *Evaluating Correlations between physical and online world* In Chapter 4, 5 and 6, the definition of methods of the exploration of the connections between the two worlds in terms of correlations and differences between them.
   c) *Evaluation of the use of Social Media in Transport* The exploration of the various aspects of using Social Media in transport is pursued with the adoption of a Social Media Taxonomy in Transport and the evaluation of different data sources as well as the creation of a Strengths, Weaknesses, Opportunities and Threats (SWOT) analysis.
   d) *Data enrichment for Social Media data* The use of classification techniques for the evaluation of recurrent patterns is introduced aiming at enriching social media data in terms of activities performed and locations visited.
   e) *Modelling Activities using Social Media* An extensive framework for the extraction of information from Social Media data with regards to activities conducted is presented. Additionally, a method for the extraction of activities spaces using clustering techniques is presented aiming at better defining the concept of areas visited by different individuals.

2. **Practical Contributions**
   a) *Practical Aspects of Using Social Media in Transport* Throughout this thesis, practical contributions in terms of the activities identified in different places around the world as well the actual use of social media.
   b) *Data Collection Aspects*
   c) *Machine Learning use in Transport* For the implementation of the methodological framework of this thesis, several Machine Learning techniques are utilized allowing for a discussion of the practical considerations for future research.

1.5 Thesis Outline

This remainder of this dissertation is structure as follows:

Chapter 2: State-of-the-Art This chapter presents an overview of the related work, including first a description of the conventional mobility–related data collection methods and the exploration of the emerging data sources. Special focus is given on the use of Social Media in Transportation studies, where the literature review is performed in terms of literature analytics, followed by a targeted presentation of topics pertinent to this thesis.

Chapter 3: Mapping Social Media for Transportation Studies A comprehensive discussion of the aspects to be considered when working with Social Media data is being developed, adapting the honeycomb Social Media framework of Kietzmann et al. [2011] for transportation. Various aspects of Social Media are discussed in terms of both data availability and Social Media functionalities offered, followed by a SWOT analysis and a discussion on privacy related implications of Social Media use.

Chapter 4: Methods The methods used for understanding Social Media use around the world and for extracting mobility–related features are being discussed. Starting with Data Collection, the methods available to harvest data from Social Media are discussed, with a detailed framework presented for Twitter. Following, details of the descriptive analysis followed are described, as well as data enrichment framework that allows for the increase of information that can be gathered from Social Media. Finally, the extraction of features is described for the exploration of activities from Social Media data. This includes the definition of areas of activities, the extraction of activities characterization based on text and the definition of activity spaces based on clustered Social Media data.

Chapter 5: Social Media and Transportation The exploration of the generic characteristics of Social Media use and their connection to the usually investigated mobility related features takes place in this chapter. This is performed on the basis of descriptive analysis 4.2, including comparisons to conventional transport–related

data. Specifically, the comparison of the data collection and the data using different data collection techniques is pursued for a collection of cities around the world. A similar exploration takes place for the historical data collection (user–based) to gain insights concerning the way that people post. Finally, the exploration of spatio–temporal of Social Media posts and the connection to Travel Survey data is included to better understand mobility patterns.

Chapter 6: Extracting Transport Features from Social Media Data This chapter focuses on the exploration of feature extraction for human activities. The main feature extracted is human activities using text data and by enriching the data with spatio–temporal characteristics. Additionally, Areas of Interest for different user groups are defined and the case of Athens is examined for the definition of the different areas from which individuals post on Social Media, for two user classes: residents and tourists. Finally, the exploration of activity space based on Social Media takes place herein.

Chapter 7: Conclusions and Future Work The last chapter of this dissertation provides an overview of the methods presented and the results achieved. Future work is outlined in general and for specific parts of the methods, for which the implementation raised additional questions. Finally, an outline of methods provisionally interesting with regards to the operationalization of the use of Social Media for demand extraction are presented; focusing on the deployment of data fusion.

2 State-of-the-Art

Contents

Components of this chapter are presented in:

E. Chaniotakis, C. Antoniou, and F. Pereira. Mapping social media for transportation studies. *IEEE Intelligent Systems*, 31(6):64–70, Nov 2016b. ISSN 1541-1672. doi: 10.1109/MIS.2016.98

Emmanouil Chaniotakis, Dimitrios Efthymiou, and Constantinos Antoniou. *Demand for Emerging Transportation Systems*, chapter Data Aspects of the Evaluation of demand for Transportation Systems, pages 0–0. Elsevier, Cham, 2010b. ISBN 9780128150184. URL `https://www.elsevier.com/books/demand-for-emerging-transportation-systems/antoniou/978-0-12-815018-4`

Emmanouil Chaniotakis, Constantinos Antoniou, and Loukas Dimitriou. *Digital Social Networks and Travel Behaviour in Urban Environments*, chapter Social Media and Travel Behaviour., pages 0–0. CRC Press, Cham, 2010a. ISBN 9781138594630. URL `https://www.crcpress.com/Digital-Social-Networks-and-Travel-Behaviour-in-Urban-Environments/Plaut-Pinsly/p/book/9781138594630`

2.1 Evolution of Data Sources for Transportation Studies

As the transportation system is rather complex, the diversity of the collected data is vast. A wide categorization of the various data sources can be viewed from the perspective of the examined component of the transportation system on which the data collection is oriented, for example user–oriented or network–oriented (Figure 2.1). Choosing the data to base the analysis required needs to be a cautious decision based on some key aspects that define the requirements of this analysis. The overall task of data collection is to provide responses to a research question.

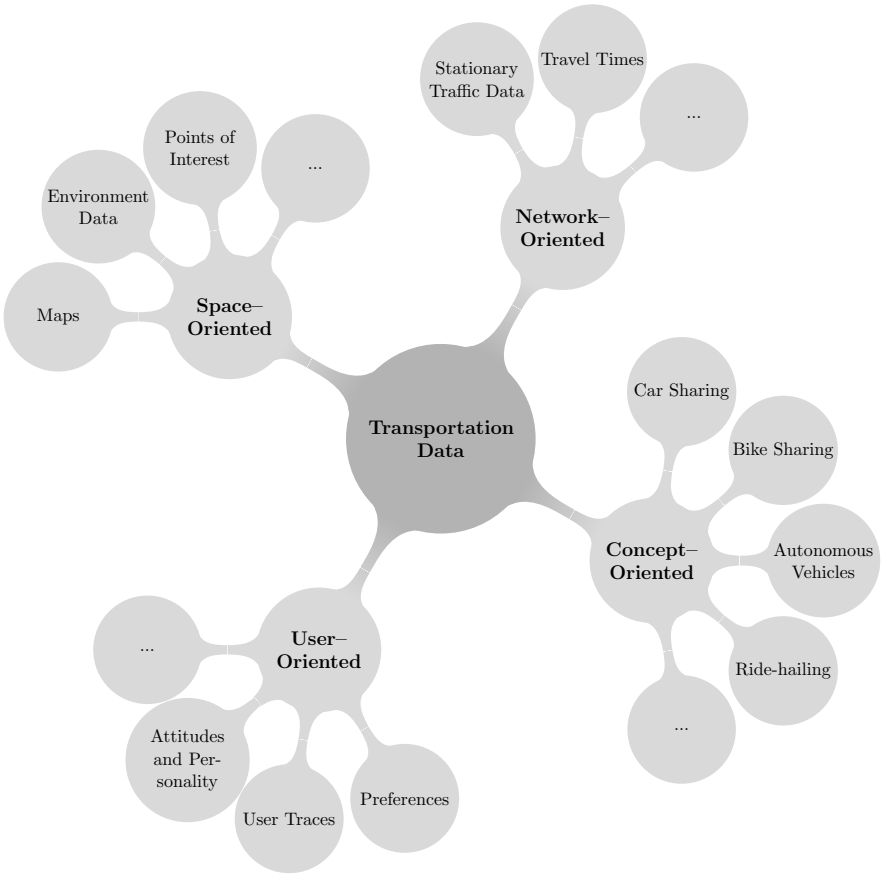


Figure 2.1: Classification of data sources types for Transportation

Historically, collecting data for transportation studies has been a rather difficult endeavor, with the most widely used method to be travel surveys, dating back in the 1950s. The first approach to be face-to-face interviews, followed by the exploration of mail-surveys and telephone surveys [Shen and Stopher, 2014]. In addition, the prevailing method to collect network–oriented data has been the use of observers or traffic counts. Concerns of data quality and high cost have fueled the exploration different data sources. Evidences from content analysis performed (Figure 2.2), for the evolution of data sources used in transportation research is rather informative: In April 2019

the Scopus query "TITLE (data AND transport) AND (LIMIT-TO (SUBJAREA , "ENGI")) " performed, resulting in 830 documents. After a manual relevance filtering (two observers) 158 were found to be relevant. The documents were reviewed and the data sources used where extracted. As it is observed in Figure 2.2 for the 5 most frequently explored data sources, exploration of different data sources emerged the last 2 decades with the exploration of smart-card data and mobile phone data. The last few years combination of different data source is also explored. Another interesting finding is the emergence of the discussion on open-source data.

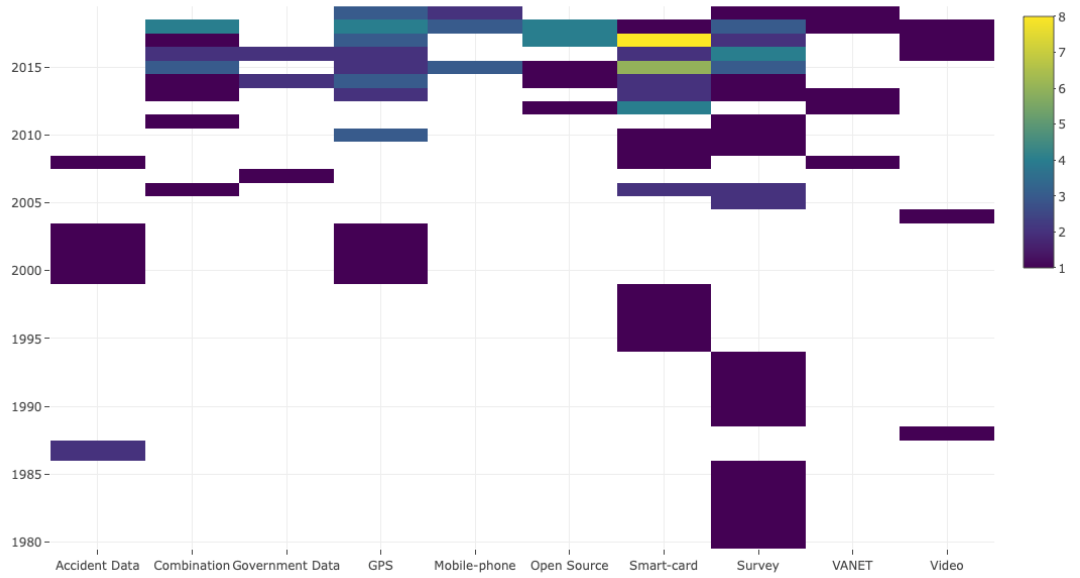


Figure 2.2: Historical Evolution of Data Sources

Apart from the top-5 data sources that are presented in Figure 2.2 researchers commonly study data related to connected and/or autonomous vehicles and others. Finally the combination of different data sources is found to emerge as an important topic of research.

2.2 Conventional Data Collection Methods

Conventional data is characterized the data collected following a hypothesis-based approach. A critical discussion on the characteristics of the conventional datasets used in transportation is considered fundamental to get a basic understanding of the benefits that emerging data sources might have. Given the prevalence of travel surveys (primarily for user–oriented data), their basic aspects are discussed in Section 2.2.1. Additionally, other datasets considered to be conventional (such as land use data or loop detectors data) are discussed briefly in Section 2.2.2.

2.2.1 Surveys

Travel surveys constitute experiments that are created upon factors that are relevant to the research question examined. The main categorization of experiments is on the basis of **stated** or **revealed** preference (SP or RP respectively) survey research [Louviere et al., 2000, Ben-Akiva and Lerman, 1985]. Revealed preference data is related to people's actual choices in real-world situations, while SP presents hypothetical choice situations [Train, 2009]. For both types of data collection a long debate took place about the ability to acquire valid data and on to what extent they can reproduce actual people's behaviour and be used to draw conclusions for further planning [Audirac, 1999, Hensher, 1994].

SP methods receives criticism due to their reference to a non-experienced situation and as a consequence possibly not well defined by the researcher. Since the information is not directly observed, they are susceptible to various sorts of biases. On the other hand there is the advantage that researchers control the situation presented including subjects that they want to be analysed. Also SP allow trading-off for attributes of different alternatives which allows for the estimation of important for transportation constructs, such as the willingness to pay and allow for multiple observations from each individual [Hess et al., 2010]. With all the advantages and disadvantages SP is widely used in transport research [Hensher, 1994]. Finally, for situations that happen rarely, it is doubtful if people act as they answer and it has been argued that choices taken, recorded on revealed preference survey do not imply the same behaviour to conditions other than the one exposed [Audirac, 1999, Train, 2009].

For both methods, the construction of the experiment requires the selection of the preferences to be recorded. For example, when interested in investigating the factors that affect mode choice using RP, the question would be "how many times did you use the indicated service", while for the case of SP the question would be structured as a choice experiment, where participant would be asked to select what they "would choose" given a described by the researcher situation. Given this flexibility in defining the experiment scenarios and the ability to ask more than one question in SP experiment, research focuses on the determination of the methods to create the experiment. However, the number of scenarios increases with the number of attributes/alternatives [Louviere et al., 2000], making the evaluation of all combinations (full factorial design) in most cases impossible. To overcome this issue several methods have been proposed (random selection of scenarios; orthogonal designs or efficient designs) [Walker et al., 2018, Rose et al., 2008].

Conventional surveys also introduce characteristics of the individuals, in the form of socio-demographic characteristics, personality traits as well as attitudinal characteristics with regards to the transportation system. This is commonly performed in the form of indirect questions which aim at evaluating the impact that they have, using usually latent class models and factor analysis. The definition of the settings of such experiments involve trade-offs between quantity of responses, quality of the results and information volumes [Richardson et al., 1995]. This might lead to situations where hypotheses and related work might dictate many more factors affecting demand than

what can be included in a survey. For example in Durán Rodas et al. [2019] there are on average around 17 built environment parameters were found to be significant when estimating models for demand for bike-sharing using a data driven approach. Such a high number is in many cases prohibiting to evaluate in a survey. Respondents can process only a limited survey load before fatigue kicks-in [Porter et al., 2004]. Additionally, if much effort is required to understand the questions respondents tend to either abandon the survey or answer in an inconsistent/non-attendance or other way, introducing biases in their responses [Hess et al., 2010].

2.2.1.1 Collection Instrument

The instruments to perform data collection for transportation has also been found to affect the data collection process. For many years the collection of data was performed on the premises of paper-based surveys and focus groups, which were later extended to telephone surveys. Lately, online surveys has been widely used, however not without issues; with the main to be the representativeness of the internet users, given the ease of using online social networks for distribution. As it is rather clear, each of these methods bear the risk of not being representative as access to the instruments might be limited.

All collection instruments deployed, impose constraints to the data collection process. For instance, telephone surveys are restrictive in terms of describing scenarios with different attributes for alternatives, while distributing surveys over post, does not allow for the explanation of the situation by an interviewer. They also bear the risk of introducing additional errors and biases in the data collection process. For example, an interviewer can contribute to biases related to acquaintance; presentation of scenarios misusing colors might create conditions for prompt choice; fatigue of an interview over the phone, could result in different reporting than paper survey.

A special case of data collection is that of the focus group [Morgan, 1997]. The idea is that a group of people is formed to discuss a topic, with the main goal is to have a guided (by a moderator) discussion, on the subject to be investigated. Focus groups are set to answer questions in a usually open-ended way. In focus groups the moderator is plays a sensitive role and should exhibit specific skills that would allow the extraction of proper information [Krueger and Casey, 2002]. In transportation the use of focus groups has been pursued for a wide variety of topics. In most of the reported cases of using focus groups, this has been done in conjunction to performing surveys and other data collection methods; while the focus has primarily been placed on the collection of qualitative aspects of the examined topic. To name but a few, Preston and Rajé [2007] studied social exclusion in terms of accessibility using a combination of surveys and focus groups; Akyelken et al. [2018] studied aspects of shared mobility for London; Daziano et al. [2017] studied willingness to pay for automation, using a combination of survey data and qualitative indicators from focus groups; while Politis et al. [2018] studied driving habits, in relation to handovers of control in autonomous cars.

2.2.1.2 Sampling

The sampling procedure plays a significant role in the data collection. Most of the models used, assume data from random respondents (strata are commonly specified to ensure representativeness, but within strata data collection is assumed to be random) [Ben-Akiva and Lerman, 1985]. However, this is a prerequisite which is difficult to comply with.

An ill–defined sample, could lead to what is commonly referred to in the literature as sampling bias. Sampling biases affect the quality of the data, and refer to situations where some categories of the examined population are less likely to be represented. One widely discussed bias is the non-response bias (or also discussed as the counterpart self-selection bias). Apart from the obvious issues related to spatial distribution, and representativeness, researchers have to deal with non-response bias, which refers to individuals who choose to not respond, or responds late to such surveys [Richardson, 2003]. Non-respondents, or late respondents might have a completely different behavior from those who respond (on time), as well as with people who deliberately choose to respond to a survey. Brog and Meyburg [1980] analyzed data from different cities in Germany and found that there is a strong correlation between trip-making characteristics reported and time taken to respond to the survey. The impact of sampling has become more apparent with the widespread of web-based data collection instruments and distribution methods, as they reach specific population groups, who have access to these instruments and can be reached from the distribution method. Some additional categories include undercoverage bias; survivorship bias; exclusion bias; overmatching bias; pre-screening, or advertising bias.

Aiming at correcting some of the potential sources of bias and being able to have enough data to have the essential degrees of freedom to perform model estimation and capture heterogeneity in responses, research focuses on the estimation of the sample sizes. Again here controversy arises. The proper statistical way of estimating sample sizes usually requires some prior knowledge of the system studied. This leads to the use of rule of thumps for coming up with an adequate sample size, which as it is understandable, suffers from the danger to introduce bias in the estimation [Washington et al., 2010, Johnson and Wichern, 2002].

2.2.1.3 Response

Measurement errors in surveys, refer generally to the situation where responses provide other than the true results [Biemer et al., 2011]. This could be attributed broadly to any type of error source which result in difference between the true response and the recorded one. For example, respondents could not understand the question or do not want to answer in a true way or a mistake takes place when the interviewee records the answers, or there is a GPS positioning error when using data-loggers. Biemer et al. [2011] identify measurement errors sources: a) Survey design, b) Data collection method; c) Interviewer and d) Respondent.

Measurement errors could be of random nature (non-sampling errors), or they could be systematic ones. When the later relates to the way that participants respond, it refers to a category of errors commonly known as response bias. Response bias is generally referring to the situation when respondents illustrate a tendency to answer questions in a different way that what the context defined suggests [Paulhus, 1991]. For example, a respondent would like to have a particular service offered to them, so they report over-use of this service (if it would have been available), or a respondent wants to be liked from the interviewer. Response biases can widely affect the validity of the data collected [Furnham, 1986, Orne, 1962]. Response biases are related to response styles, which describe the tendency of individuals to follow a specific response policy. Some widely discussed response biases pertain social desirability and its opposite, acquiescence or yea-saying and its opposite and extremity response or mid-point response. In their seminal work Furnham [1986], Nederhof [1985] describe the techniques used to try and control for these types of biases, by measures of prevention, measurement and correction.

Another family of systematic bias relates to halo effects and leniency or severity biases. In the former, respondents tend to be positive or negative with the evaluation subject (e.g. car sharing) by aspects not observed or included in the survey such as own expectations and preferences [Kahneman, 2011]. Leniency bias may occur when an observer has the tendency to be lenient in all of his/her assessments while severity bias may occur when the respondent has the opposite tendency to be harsh/severe.

Stated preference experiments bear also a significant amount of other biases. The main reason is that the scenarios collected from the same participant could provide more information on how individual choose. In that sense, the effect of non-trading, lexicographic and inconsistent choice has received attention [Hess et al., 2010]. In non-trading behavior, it is clearly observed that individuals do not change their choices even if other alternatives are more attractive, while in lexicographic behavior, individuals choose always the alternative that has a specific property (for example the cheapest in every case). Finally, inconsistent choices refer to the situation where choices between alternatives are not consistent, in terms of how participants value each attribute. Additionally, other biases discussed include the anchoring bias [McFadden, 2001], inertia bias [Thaler and Sunstein, 2009], hypothetical bias [Murphy et al., 2005] and aggregation bias [Morrison, 2000], as well as attribute non-attendance bias [Hensher et al., 2012].

2.2.2 Concept, Spatial and Network Oriented Data

Other possible sources of data are categorized in terms of concept; spatial and network–oriented data. With regards to concept-oriented data, the collection of data has mainly relied on aggregated data to characterize demand [Bagchi and White, 2005]. For instance data collection for public transport demand was either performed on the basis of ticket counting or with observers who monitored the demand in the vehicles themselves. However, as different types of tickets existed and transfers are in many cases allowed using one ticket, the actual demand is very difficult to be observed. Additionally, data

related to the routes and schedules as well as delays were scarcely available and in many cases (even today) collected on a report-basis, where the driver or the operator has to report an incident and the corresponding delay.

Spatial data is relevant for transportation systems, as the information that they describe essentially drives demand [Efthymiou and Antoniou, 2014]. Spatial data has been originally collected in the form of land use and built-environment characteristics. This has been usually available from registries available by the relevant public authorities, following in most cases troublesome acquisition processes and being scarcely updated. The granularity of the spatial data collected in a centralized way did not allow for evaluation in a spatial scale smaller than the one defined by traffic analysis zones, constituting a dataset of low resolution.

Network–oriented data on the other hand has been collected to monitor demand, primarily for car transport. The instrument in most cases have been the use of loop detectors, cameras and radar or lidars. In some cases plate recognition was used for the extraction of travel time, however, as these efforts were taking place following a centralized deployment approach, data collection takes place on some parts of the network, and inference techniques are deployed for the estimation of network performance metrics [e.g. in Tamin and Willumsen, 1989].

With regards to emerging data collection methods, it should be noted that the data typology is the same, however the data collection instruments introduced enable the collection of richer datasets. In other cases however (such as the case of Social Media) this type of data typology has been completely new, enabling data-oriented research. In addition, users are much more willing to share their data, usually in exchange of services, while the concept of open sharing of data is in some cases viewed as a mean to empower citizens and build the so called digital democracy [Helbing and Pournaras, 2015]. These changes observed and the sources of data available have received wide attention from the transport scientific community towards the utilization of the increasingly available data (Big Data) [Buckley and Lightman, 2015, Reades et al., 2007].

2.3 Emerging Data Sources

2.3.1 User-oriented Collection Methods

For performing surveys, the use of GPS-based data loggers for supplementing the traditional surveys was initially pursued [Doherty et al., 2001]. However, the high investment cost and the unease that carrying an additional device imposes, made them obsolescent against the use of smart-phones [Cottrill et al., 2013]. With their wide spread [Pew Research Center, 2019] and very high penetration rates, research has been performed on the benefits of using smart-phone apps for performing collection of data that comprises both travel survey as well as performance of RP and SP experiments. The benefits of smart-phone are rather clear: users carry them charged everywhere, they comprise an increasing number of sensors, and they offer the opportunity to communicate with the user. Prelipcean et al. [2018] provide a recent overview of the development of such smart-phone apps. Most of these apps are usually closed-source or provided by external

developers and used specifically as an experiment tool, that collects specific data types. The process followed in most of the cases is the collection of GPS traces, automatic or manual data annotation and user validation [Cottrill et al., 2013]. In most of these cases, incentives provided are monetary, in the form of either either directly earning money while using the app or participating in a draw to win a prize.

These apps were mainly designed for passively collecting travel date and are scarcely used to conduct preference experiments that relate to the trips performed by individuals (e.g. asking mode alternatives for a trip performed.) To the best of our knowledge, in the transportation community, the only exception is the recently published paper from Danaf et al. [2019] where they create context-aware experiments for better mode choice prediction. Within the Internet of Things, the development of apps for social data is also growing. Griego et al. [2017] have used an app for collecting spatially aware urban qualities for smart cities. In an innovative data collection concept, users entering a geo-fenced area are asked questions and simultaneously collect environmental and mobility data.

2.3.2 Concept-oriented Collection Methods

One of the key aspects of ICT advancement is that they enabled the collection of data for all transport concepts. An example for public transport is the introduction of smart-card data, where users of Public Transport are required to tap–in (and in some systems, tap–out) whenever using them. This has lead to the definition of a spectrum of methods varying from data mining [Ma et al., 2013], and demand estimation [Jun and Dongyuan, 2013] to individual pattern extraction [Zhao et al., 2018] and metrics of user satisfaction [Ingvardson et al., 2018].

Similar data of use, can be extracted for many of the emerging mobility systems. With regards to bike–sharing there are more than 800 programs around the world with a fleet of more than 900.000 bicycles with trip proposes to be commute to work on weekdays and leisure and/or social purposes [Fishman et al., 2013]. Many of these programs publish their data as open data allowing for data analytics concepts where data from different cities are compared. For instance Chardon et al. [2017] studied the trips per day per bicycle in a city level in 75 SBBS systems in Europe, Israel, United States, Canada, Brazil and Australia, with the independent variables were the operator's attributes, the compactness, the weather, the transportation infrastructure and the geography; Zhao et al. [2014] correlated the logarithm of ridership and turn over rate using data of 69 SSBS systems in China with urban features and system characteristics; and Durán Rodas et al. [2019] developed a data–driven method for estimation of demand models of station–based bike sharing using built environment factors where data of multiple cities are pooled in one data–set.

Car–Sharing is the short-term rental of a car. Subscribed customers are charged to use vehicles based on various pricing schemes (based on vehicle type, kilometer driven, location, time of use), and in some cases a subscription fee. Car sharing essentially allows sharing the use of a private car, on a need basis. The distinction of car-sharing services is among others based on the existence or not of stations. In general there

are Station–Based, Free–Floating or mixed system, where for Station–Based, users have to return the vehicle to a specific location (not necessarily the start location though) while for the Free–Floating systems, users could park their vehicles wherever they want within a business area (and out of this but with an additional fee). The recent wide implementations of car–sharing (in July 2019 Share–Now operates more than 20,000 vehicles in more than 30 cities worldwide – https://www.your-now.com/) have led to the availability of datasets for some of the cities operated. Acquiring the data is on a request to the providers basis (e.g. Share–now, ZipCar) [Schmöller et al., 2015], however attempts have been made on extracting data from the Application Programming Interface (API) [Trentini and Losacco, 2017]. This latter methods is subject to the general data use policy of the provider and bears issues such as inability to define relocation, false/canceled reservations and routes. The exploration of these datasets have yielded interesting results with regards to their use.

2.3.3 Network-Oriented Data

Advances in ICT has also brought new data on a network level. Departing from the conventional traffic count data, efforts have been put through to use different types of sensors for capturing the network based data. This has been directed towards the deployment of networks of Bluetooth or WiFi sensors aiming at capturing the users of smart-phones or cars who bear these features. In most of these cases the deployment of sensors is a rather cheap solution (from a report in Greece the cost for one sensor including installation and 3 years warranty was 2300 Euros [Mitsakis and Iordanopoulos, 2014]). Reliability of such systems relies on the penetration rate of devices who broadcast such signals [Friesen and McLeod, 2015]. In addition placing the sensors and data cleaning are very important to ensure representativeness and reliability. For the former, optimization techniques have been used using different optimization criteria such as observability or flow estimation and OD pair separation [Gentili and Mirchandani, 2012, Zhou and List, 2010, Fei and Mahmassani, 2011]. On the latter, aspects of mode detection and scaling of the bluetooth/WiFi flows to actual traffic flows has been explored [Barceló et al., 2010, Bhaskar and Chung, 2013].

2.4 Data Quality Assessment

The emergence of diverse data sources implies the investigation and assessment of data quality. The discussion however is not new: metrics for data quality have long been present in statistical analyses and methods to quantify data quality has long been discussed (e.g. statistical tests). Research has extensively focused on the impacts of data quality to the information systems. For example Ballou and Pazer [1985] presented a model for the propagation of errors in informations systems, following the extensive work that had been done in accounting literature for evaluating the impact of errors and controls on financial balances [e.g. in Cushing, 1974]. Later the idea of assessing data quality on the basis of propagation of errors has been extended, to cater for the multi-dimensional concept of data quality. Batini et al. [2009] provided a review of

13 prevailing methodologies for data quality assessment and report the used quality dimensions. From a meta-analysis of the Batini et al. [2009] work, the prevailing criteria of data quality (with regards to data) to be presented in Figure 2.3. As it is clearly obvious, the prevailing factors are included in most of the studies examined. A striking difference is that in most cases the evaluation of data in transport takes place on either the accuracy or in some cases for emerging data on interpretability and accessibility, without however a clear and consistent use of metrics that would allow their comparison.

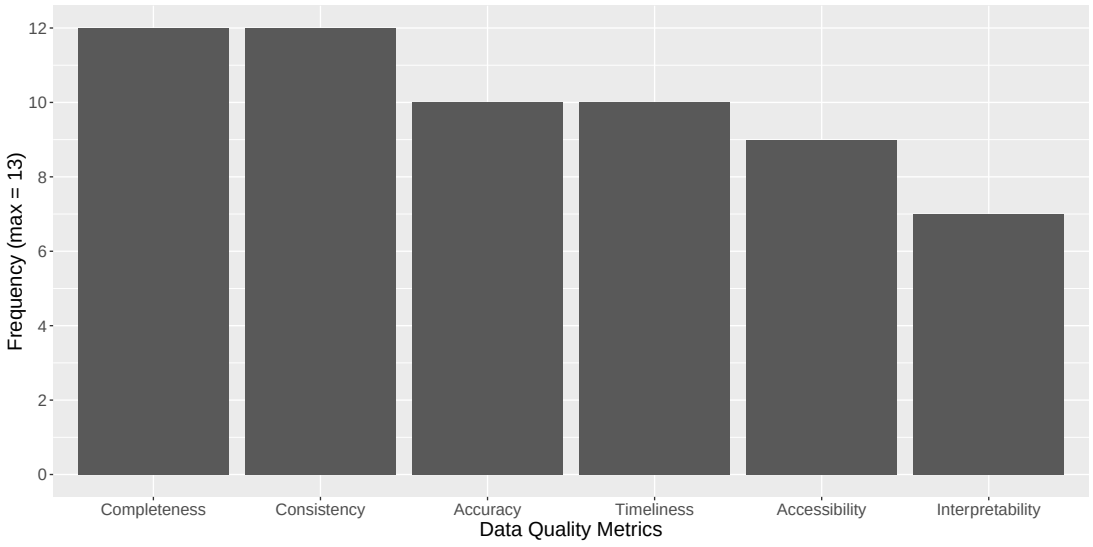


Figure 2.3: Prevailing data quality metrics (Frequency ≥ 5)

Going in detail with each data quality criterion, Batini et al. [2009] provides a comprehensive overview of the definitions that the difference studies examined use. Here we are consolidating the information and providing some definitions and notes related to transportation. Starting with **completeness**, it is generally used to refer to the ability of the data to describe the system for the application at hand. Essentially completeness related to extent that the dataset examined has all the required elements (or variables) to achieve observability and to be used to do analysis and perform modelling. Some researchers have used completeness to describe how much data is available with regards to the maximum possible data (or information) that can be achieved [Biswas et al., 2006], or to evaluate to what extend reports have been completed [Tin et al., 2013] or define the magnitude of missing values [Naumann, 2002]. Clearly there are differences in the above definition and also the reference system [Batini et al., 2009]. Based on the above, for the case of evaluating data completeness for emerging transport systems and their demand completeness should be examined on the basis of exploring the extend to which data can actually describe the system and the interconnections to the remaining transportation system. An example could be bike-sharing data. Assuming that only a sample of 10% of all bookings per day is provided, completeness could be

as high as 10% (but in many cases much lower) of the information needed to describe the bike-sharing system, and even less for describing the transportation system of an area. This could be easily been understood when taking into account that bike-sharing could be as low as for example 5% of the modal split.

Consistency usually refers to compliance with some data–related rules. Several sub–categories of data consistency exist such as inter–relation and intra–relation [Batini et al., 2009] or internal and external consistency. The former categorization refers mainly to the relationships between data (and can be viewed as the examination of internal consistency), while the latter could be viewed as a metric of consistency in terms of data being consistent with data and data being consistent with the real system observed. One example of internal consistency could be that all age data for a particular study refers to individuals of age 20-30 years old, while external consistency could be viewed from the perspective that people of the 20-30 age group are the only people that interact with the system.

Accuracy is one of the data quality measures that has been long discussed; with different different definitions to be found in the literature. As described in Batini et al. [2009], Wang and Strong [1996] "the extent to which data are correct, reliable and certified" and Ballou and Pazer [1985] refers to the data correspondence to real values, Redman [1997] and the proximity of a data value, to some other value that is considered correct. In transportation and particularly with regards to the emerging mobility concepts, measuring accuracy of the data is a rather complicated task, with the main reason to be nonexistence of reference data on the base of which we could compare. For example, many studies explore acceptance and adoption of Shared Autonomous Vehicles. The only way of evaluating the accuracy of these can be based on the comparison of the results of different studies, in many cases performed in different places around the world. For the evaluation of data accuracy, there are various statistical tests that can be deployed for example examining if the results belong to the distribution. For more information, the reader is referred to the comprehensive overview of statistical methods for transportation provided by Washington et al. [2010].

Timeliness is found in the literature to be used to describe the time that it takes (delay) so that changes of the real system are reflected to the data. Timeliness needs in transport differ on the basis of the commonly followed distinction between off–line and on–line use of transport data. Also, depending on the type of the data, timeliness of transport data can vary from years to a few seconds. To give some examples, travel diary surveys happen in a range of 2 to 10 years, traffic counts provide aggregated results every 1-15 minutes; while smart-card reads and trajectory data can be available almost instantly to the database used.

Finally, **accessibility** is defined upon the possibility of getting access to the data and in some cases as the time to get the request, based on the delivery time, request time and deadline time (essentially more referring to the timeliness) and **interpretability** finally, refers to the extend that data is interpretable.

2.5 Transferability

Patterns of time use and travel can be significantly different among individuals at different stages in their life cycle living in places that offer different opportunities for activity participation. It is very important to understand the underlying behavior that governs the choices made in different contexts (i.e. different countries and different cities) for policy definition and this has attracted the interest of research as a potential tool for understanding the outcome of adopting policies and transferring models [Lleras et al., 2002, Timmermans et al., 2003]. This need has become imperative lately with the introduction of new disruptive forms of mobility (i.e. car-sharing, autonomous vehicles) as well as new types of data [Cramer and Krueger, 2016, Belk, 2014]. Comparisons of different travel characteristics and activity patterns are not new in the pertinent literature. Schafer [2000] presented a comparison of around 30 travel surveys in different countries for regularities in time and travel budgets. Lleras et al. [2002] presented an international comparison using travel survey data, using structural equation modelling and latent variables to explain mobility changes, number of trips and total travel time. Simma and Axhausen [2001] performed a comparison of mode choices commitment in a joint relation of car ownership and ticket availability. These studies have shown that transferability of average indicators is in most cases not an adequate solution and should be avoided. de Abreu e Silva and Goulias [2009] performed a comparison between Seattle, WA, USA and Lisbon, Portugal including residential location choice, car ownership, and travel characteristics and Kühne et al. [2018] investigated the car ownership characteristics in Germany and California. Pendyala (14) studied time allocation in different countries finding differences and similarities and this is the only study that examines aspects of daily rhythms across different countries.

Examination of the underlying behavioral characteristics related to activities performed and daily rhythms (such as daily time allocation, activity timing and scheduling, activity frequency) have been found to govern decision making for travel related behaviors and should be included in transportation models de Abreu e Silva and Goulias [2009]. In other words, if people behave the same within clusters of individuals' behaviors, there is a good basis that they would behave the same when it comes to aspects such as new forms of mobility, response to changes in life and new infrastructure. Therefore, studies examining daily summaries in different countries or different cities within countries need to expand their repertory to time of day dynamics.

2.6 Social Media Use in Transport

The emergence of SM has led to many definitions that attempt to capture the diverse services offered. Here, the most generic one is adopted, from Andreas Kaplan and Michael Haenlein, who define SM as "a group of Internet-based applications that are built on the ideological and technological foundations of Web 2.0 and allow the creation and exchange of user-generated content." [Kaplan and Haenlein, 2010]. Sterne [2010] has proposed categories that confine the different SM, including forums and

messages boards, review and opinion sites, social networks, blogging, microblogging, bookmarking, and media sharing.

From their rise, Social Media platforms have received attention from the scientific community, forming a potentially new stream of research. SM-based research is conducted in various different scientific disciplines, such as social sciences, economics, politics, and tourism. The reasons behind this attention can be summarized upon the unprecedented amount of information that can be extracted and the opportunities that Social Media platform use can provide on direct communication with users. The additional information collected from Social Media can complement conventional data collection methods and ultimately provide a better understanding of daily urban rhythms.

From a research perspective, a growing amount of related work has been published in the last few years, showcasing the potential of using Social Media in transportation. The literature on Social Media and Travel Behaviour is vast. [Rashidi et al., 2017] presented a discussion on the evolution of the literature related to Social Media and transport. They performed a Scopus search only for title and abstract, which resulted in 935 papers until late 2015 performing the following query:

("Social media" OR Twitter OR foursquare OR facebook OR yelp OR instagram)
AND ("travel" OR "transport" OR "mobility" OR "geo")

Performing the same search in December 2018, yields 2442 documents, with approximately 400 papers being published every year. To get a better idea of the various focus areas of social media use in transport, we have performed a number of different queries that illustrate the various research trends that emerge (Table 1). For comparison purposes we have used the same query structure for social media (("Social media" OR Twitter OR foursquare OR facebook OR yelp OR instagram)) and we have been changing the transport-related part.

Table 2.1: Number of papers resulting from Scopus search based on keywords used

Keyword*	Papers
Traffic	1096
Travel	838
Tourism	791
Mobility	685
Transportation	403
Trip	227
Travel behaviour	44
Travel demand	21

** Query performed as a combination of (("Social media" OR Twitter OR foursquare OR facebook OR yelp OR instagram)) and the keyword mentioned, only for title and abstract.*

Interestingly, different trends emerge with the establishment of Social Media research in Transport. As presented in Figure 2.4, in the first few years the keyword "traffic" was dominant, there has been a gradual but steady swift towards "tourism" and "travel". The evolution of the topics examined within Social Media research can be attribute to the better understanding of its use, and the ability to use Social Media for the extraction of meaningful information.

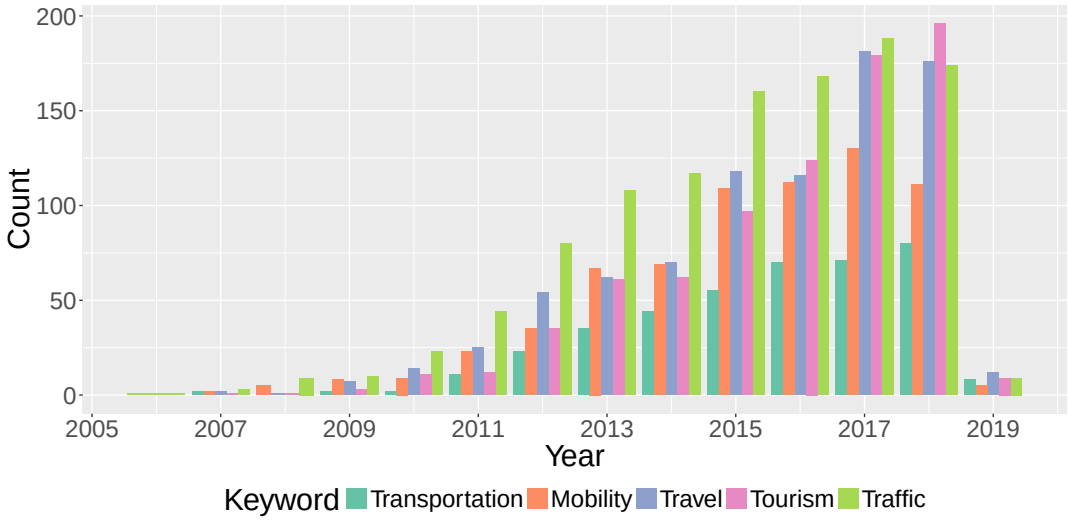


Figure 2.4: Evolution of Literature for different keywords explored, based on scopus search

The most commonly exploited Social Media originated information for transportation is based on the use of the spatial information accompanying posts (that is, geotags) and the language processing of posted content. Social Media data has been exploited for its continuous streaming of information, used to identify disruptions or special events, and for forecasting, or in terms of historic data analysis for extraction of mobility patterns. Additionally, transport service providers are using SM to directly communicate with customers, which differs from the use of cached/downloaded users' data from SM. The areas of transportation research where contributions with regards to Social Media utilization are schematic depicted in the following Mind Map (Figure 2.5); while in the paragraphs below further information with regards to the specific contributions are discussed.

2.6.1 Real–Time Data

On incident, event and disruption detection Wanichayapong et al. [2011] used synthetic analysis to classify traffic related incidents in spatial dimensions, while Gu et al. [2016] performed text analysis for the available tweets to detect accidents in real-time on arterial and highways. Additionally, Abel et al. [2012] presented Twitcident for automated collection and filtering of emergency related information and Pereira et al. [2013] focused on the identification of non–recurrent events from web–pages in order to

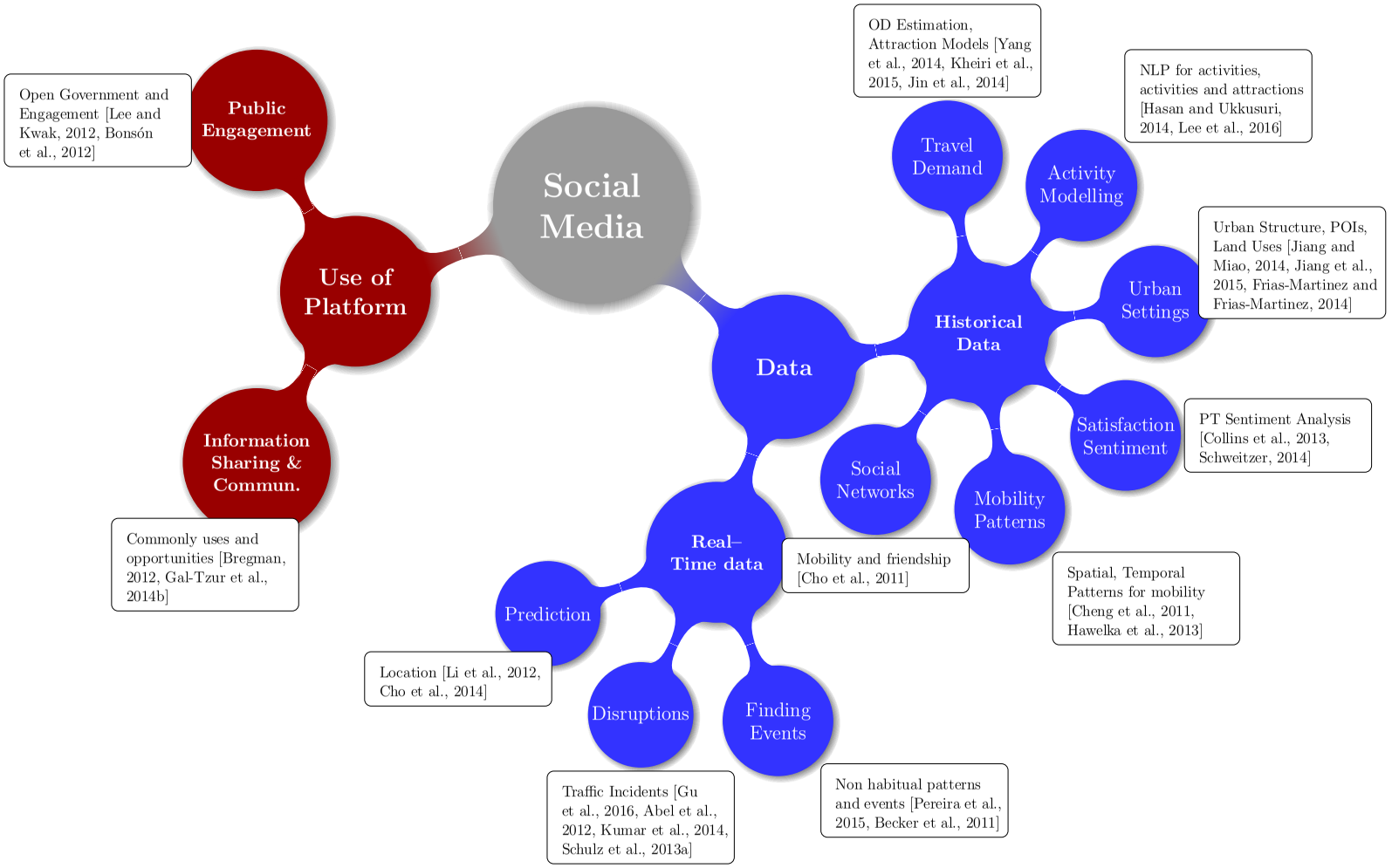


Figure 2.5: Mind–Map of Social Media uses in Transportation, with indicative references

estimate the resulted traffic. With the aim to be understanding the real–time traffic Ni et al. [2016] used social media to predict the subway passenger flow under events occurrence. Zhang et al. [2016] find a correlation between Twitter concentrations (the tweets that are shared by a large number of people) and traffic surges. In the same direction, Schulz et al. [2013b] used microblogs to detect small scale traffic incident in real-time and increase the road users awareness of incident situations, following semantic enrichment and text classification, with this approach to has a 89% success rate. Finally, Kumar et al. [2014] used Twitter to detect poor road conditions in order to be used as a "road hazard alert system".

With regards to traffic prediction, [He et al., 2013] presented a framework for the estimation of traffic intensity based on data from Social Media. In order to extract traffic intensity from Social Media posts, the estimate optimally a transformation matrix for traffic related indicators and incorporate them in traffic prediction using linear regression. Aiming at detecting traffic jams, Zhao et al. [2015] has developed a generative model using Twitter for spatiotemporal event forecasting. Their model is found to be capable of predicting the time and location of congestion on road network.

Social Media has bee also studied as a solution for the challenging problem of real-time congestion monitoring. Chen et al. [2014] proposed a model to monitor the traffic using Twitter, which utilizes textual, temporal and spatial information using topic modelling, probabilistic soft logic and hinge loss Markov random fields. Ni et al. [2016] develop a hashtag-based event detection framework, which is used to develop a fusion model, based on a hybrid loss function that uses passenger data and Social Media data predict subway passenger flow, while Gu et al. [2016] establishes a dictionary of important keywords and their combination to detect traffic–related incidents, using classification methods. The traffic–related incidents identified were found to be mainly produced by highly influential users such as users (such as public authorities) which post this information on Social Media platforms. Zhang et al. [2018] used deep learning for the detection of traffic incidents (i.e. accidents), with a reported accuracy of 80%.

2.6.2 Historical Data

On the other hand, researchers have explored historical time data's use for modeling applications and as supporting actions for gaining theoretical insights for the transportation system. The main applications examined can be distinguished based on the purpose of using Social Media originated data. Rashidi et al. [2017] presents the results of an experts' survey on the usefulness of Social Media data for transportation studies and discusses methods on possible ways of using Social Media data for extracting mobility–related features such as mode of transport, activities and their duration and land use variables.

With regards to identification of Spatial and Temporal mobility patterns, Cheng et al. [2011], Rebelo et al. [2015] presented some first statistics on mobility patterns from social media data. Huang and Wong [2015] presents the use of Space Time Paths (STP) using Social Media data to explore the trajectories of individuals and define STP for Washington DC; while Yao et al. [2018] proposes the extension of STP

using of multiple spatio–temporal paths (mSPT) that allow the exploration of mobility patterns based on the definition of individuals' hotspots. Jiang et al. [2015] explore the definition of hotspots in urban areas using Social Media data for for urban land use classification and disaggregation; Yang et al. [2018] analysis mobility patterns for users from the Chinese University of Geosciences Wuhan (CUG Wuhan) to explore spatio–temporal and activity patterns, finding that the distance of activities follow a power–law distribution and that users with higher check–in activity do not perform more activities on the physical world. Maghrebi et al. [2016] presents an analysis of mode choice using Twitter, by exploring mentioning of modes in tweets text.

Social Media originated data has also been used to identify the movement of population [Liu et al., 2014] and investigate users' social networks and the effect they have on transportation-related behaviour [Cho et al., 2011], the investigation of riders satisfaction and the examination of the relationship between Social Networks and mobility [Collins et al., 2013, Casas and Delmelle, 2017]. The historical use for social media data is also extended to define and verify the accidents black spots on practical levels. The Victorian government in Australia had developed and program using the historical Tweets to detect and verify accidents black spots [Sinnott and Yin, 2015].

Research also focus on the the investigation of the applicability of the Social Media–originated data for travel demand modelling. Towards this direction, research has been conducted on the feasibility of Social Media to be used for the extraction of travel demand patterns using also conventional transportaiton data [van Eggermond et al., 2017, Cheng et al., 2018]. Additionally, Lee et al. [2019] explores the definition of origin-destination and compares it to calibrated statewide demand model using spatial lag Tobit model and latent class regression models. Hu and Jin [2018] explores the potential of defining dynamic travel demand patterns using time–delayed gravity models for dynamic OD estimation.

On the identification of user activities, Hasan and Ukkusuri [2014] presents an activity generation framework using activities extracted from Social Media data using topic modeling. In an extension of the original framework introduces a semi–Markov modeling approach for the prediction of users' next (missing) activity [Hasan and Ukkusuri, 2018]. Lee et al. [2016] presents the exploration of activity spaces using Twitter data in California, while Meng et al. [2017], Paule et al. [2019] explore extraction of activities (trip purpose) sequences using a Dynamic Bayesian Network model and Bayesian Neural Network respectively.

With regards to definition of urban settings and related characteristics, there exist a vast number of articles that aim at the definition of Points of Interest (POI) [García-Palomares et al., 2015, Jiang et al., 2015], urban boundaries [Jiang and Miao, 2014, Yin et al., 2017] and land uses [Zhan et al., 2014, Liu et al., 2017b, Chen et al., 2017, Sun et al., 2016].

2.6.3 Use of Platform

The use of Social Media platforms for information sharing and communication forms a two-way communication channel between transportation operators and individuals.

One of its main advantages is that it reaches large audiences with users to be able to quickly communicate with the operators. Social media has been used for disaster management –mainly during the phases of response and reconstruction– to create the essential channels of communication and utilize information exchange, aiming at a more efficient coordination of actions [Lindsay, 2011, Gao et al., 2011]. To name but a few, prominent cases on the use of social media in disaster management include the Southern California Wildfires during 2007 [Sutton et al., 2008] and the use of Social Media in the Haitian earthquake [Yates and Paquette, 2011]. To the best of the author's knowledge, there is very limited work presented on the exploration of the possible uses of Social Media in the process of evacuation and particularly for the derivation exploration of the human behavior in times of evacuation, while Chan and Schofer [2014] examined the use of Twitter in New York during a hurricane accident.

In the last few years, the this use of Social Media has gained popularity among transit providers as it is cost-efficient, it is reliable and it can serve as a real-time information sharing platform [Cottrill et al., 2017]. According to Liu et al. [2017a], in 2009 66% of the transport agencies in the USA uses at least one platform of SM to communicate with the users, slightly lower that that in 2011, found to be 54% for Facebook, 51% for Twitter, 37% for YouTube of the agencies Liu et al. [2017a]. The success of such endeavor has been tested in several instances with one of interest to be the examination of information handling from multiple users in a coordinated way, during the Glaxo commonwealth games [Cottrill et al., 2017]. There are several transportation operators that are actively exploring mobility–related Social Media data. For example: Southeastern Pennsylvania Transportation Authority (SEPTA) continuously performs sentimental analysis for the users tweets to improve users satisfaction levels [Liu et al., 2016]. In such a form, Social Media also can be thought as form of consumer-generated content, which is a direct effect of the flattened access to the internet [Xiang and Gretzel, 2010].

According to [Haro-de Rosario et al., 2018, DePaula et al., 2018, Bonsón et al., 2019, Manetti et al., 2017, Gal-Tzur et al., 2014b, Bregman, 2012, Schweitzer, 2014, Tang and Thakuriah, 2012, Cottrill et al., 2017] the main uses of SM by transportation operators can be summarised as:

1. Timely updates and information sharing with users, for example advice concerning services disruptions.

2. Public information about routes, fares, new projects

3. Citizen engagement between operators and users in informal way benefiting from the interactive nature of the SM, for example in receiving complaints and travel questions handling, response to questions about Social Media

4. Seasonal messages for goodwill

On the other hand users communicate with transportation operators in order to share their opinion concerning a transportation service, report an incident and make transportation related questions [Gal-Tzur et al., 2014b]. Bregman [2012] performed a survey distributed among public transport operators and found that the main goals

of using Social Media were to communicate with current riders, improve customer satisfaction and agency image.

3 Mapping Social Media for Transportation Studies

Contents

Components of this chapter are presented in:

E. Chaniotakis, C. Antoniou, and F. Pereira. Mapping social media for transportation studies. *IEEE Intelligent Systems*, 31(6):64–70, Nov 2016b. ISSN 1541-1672. doi: 10.1109/MIS.2016.98

Emmanouil Chaniotakis, Constantinos Antoniou, and Loukas Dimitriou. *Digital Social Networks and Travel Behaviour in Urban Environments*, chapter Social Media and Travel Behaviour., pages 0–0. CRC Press, Cham, 2010a. ISBN 9781138594630. URL `https://www.crcpress.com/Digital-Social-Networks-and-Travel-Behaviour-in-Urban-Environments/Plaut-Pinsly/p/book/9781138594630`

3.1 Transportation oriented Social Media Taxonomy

The ability to extract information concerning the transportation system from SM is influenced by two factors: the functionalities and focus of the SM and the data availability. Concerning the functionalities and focus of SM, the leading companies are built on market strategies shaped to fulfill different needs of their costumers. Smith[1], Webb[2] and Butterfield blogged their ideas on the needs fulfilled by each SM, developing in an essence a mapping of their market strategy pursued within a 7 blocks honeycomb functionality framework. In this article, we focus on each functionality's practical implications for transportation research. On data availability, indicators – such as the researchers' ability to collect or acquire data, the availability of georeferenced locations, and social network–related or textual information – provide a detailed mapping of the capabilities offered.

We examined this framework's building blocks in the context of transportation research. The description is based on the work of Kietzmann et al. [2011]. These seven building blocks (presence, sharing, relationships, identity, interactions, groups, and reputation) indicate why individuals use each SM.

Presence This describes the functionality that tells another user when someone is accessible and where that individual is located. In the virtual world, presence lets other users know that the individuals in question are online, whereas in the real world it lets other users know the location of their friends (or others if they share information in public). This functionality, and especially its reflection on the real world, is interesting in transportation research because it lets researchers analyze traces from individuals to study mobility and activity patterns.

Sharing This refers to the extent that a SM lets individuals share content. Depending on the platform, content can be almost anything, from documents, pictures, and videos to executable files and compressed folders. Some platforms (such as Flickr, Instagram, Myspace, and YouTube) are built around a particular type of content, whereas others (such as Facebook and Twitter) embed content within other functionalities. To the best of our knowledge, transportation studies have not explored SM content sharing, although it could allow for valuable information on the individual identity and also on some of the personality characteristics. However, such an endeavor would require the use of advanced image and video processing algorithms and would raise privacy issues

Relationships SM platforms commonly allow for the definition of social networks. Some let users define a group of acquaintances or individuals they generally want to connect with. Others use the strength of ties by letting users define groups of family, friends, and coworkers and then choose which content to share with those different groups. The inclusion of relationships as a key functionality provided by SM platforms

[1] `http://nform.com`
[2] `http://interconnected.org/home/`

allows for the exploration of online social network structures and interactions, which are factors that affect mobility [Arentze and Timmermans, 2008, Axhausen, 2008].

Identity This describes the extent to which SM allow and require users to reveal their true identity by including information, such as their real name, age, gender, education, place of birth, and profession. Identity can also be perceived as the act of disclosing thoughts and feelings that describe an individual's preferences [Kaplan and Haenlein, 2010]. Some SM (such as Facebook and LinkedIn) allow the creation of profile pages wherein individuals can describe themselves. Identity characteristics are particularly interesting for transportation research, for the identification of SM user samples for modeling purposes. Most studies on SM for transportation do not account for this aspect of SM, and instead present their work on the SM sample space from which it is not always clear how to depict the generalization to the population.

Interactions These let users communicate (via messages, pokes, and posts) with people in their SM social network and with strangers (in some cases, from groups defined within each SM platform). SM conversations can be performed in public or with private messages. Depending on the nature of the SM posting process and other functionalities, conversations can vary significantly from one SM platform to another. For example, Facebook allows lengthy posts and comments that can facilitate an exchange of political views, whereas Twitter allows only 140 characters (including links for sharing content) and is most commonly used for microblogging. The interactions that take place among individuals in SM platforms and their tone, form, and content can reveal each individual's social network and, consequently, can allow transportation researchers to get insights in the social network and mobility relation. They also allow for the identification of identity and personality characteristics that have been found to be factors affecting mobility.

Groups In many cases, SM platforms let users form communities and groups to exchange information with the community members privately or publicly. User management tools let group managers maintain a community that might be open to everyone, invitation based, or based on an approval process.

Reputation In SM platforms, reputation can be evaluated on two levels: the individual and the content posted. In many cases, SM platforms provide statistics on reputation used to provide suggestions to users –for example, the number of likes a picture receives or the number of friends or followers a user can indicate the individual's reputation. The reputation indicators that a SM platform provides can be indicative of the SM use and the user's real-life profile, adding information on the personal characteristics of users (identity).

Figure 2 presents the categorization of SM for the 10 dominant SM platforms. Darker colors indicate a stronger emphasis paid by SM providers to a specific functionality.

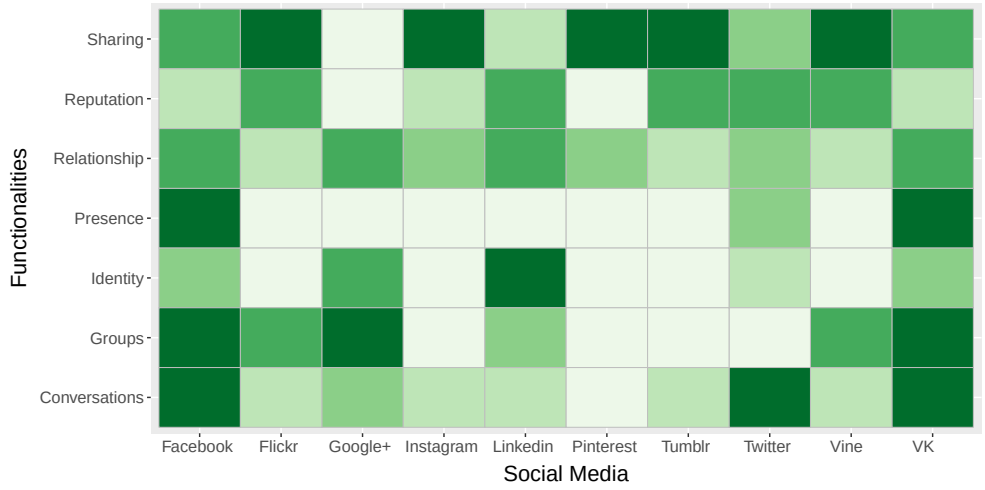


Figure 3.1: Social Media Functionalities summary table

3.2 Data Availability

The second factor to consider when designing a strategy to use SM-originated data is the data's availability. Data availability is defined as the ability to access sample data from an SM platform and use the data for research. A collection of indicators is specified to understand the data typology and the current limitations of SM platforms. Most SM platforms have created APIs through which researchers can access the data. However, the degree to which this is possible, and the type of information that someone can use, can differ significantly from one SM platform to another and is also subject to terms of use that can change without prior notice.

Although most of the research performed includes the collection of data from APIs, companies owning SM platforms in some cases share datasets for research. Chaniotakis and Antoniou presented a generic methodology for collecting SM data from the available APIs. They illustrated that in some cases a public API can allow for collection of random data, whereas in other cases, researchers must build an application and ask users to join and give privileges to the application in order to collect data from those specific users. Each SM platform generally has supporting libraries (such as Twitter4J (http://twitter4j.org) and Facebook4J) in a wide range of computer languages (for example, Java, Python, and R) that allows for data collection and the use of available data.

Figure 3.2 presents the data availability indicators for the 10 dominant SM platforms. Darker colors indicate more possibilities to extract the indicated type of information or perform indicated actions.

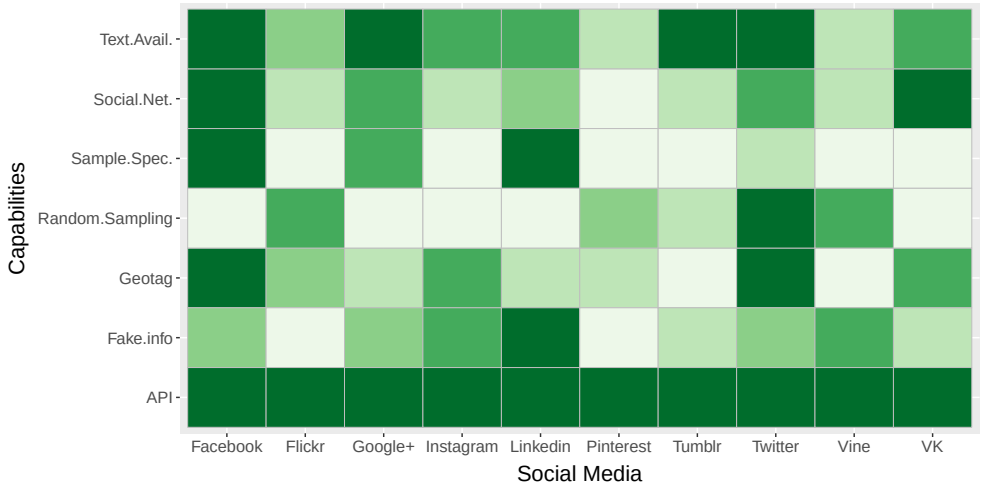


Figure 3.2: Social Media Data Availability summary table, darker indicated higher scoring on data availability

3.3 The future of Social Media Research

In light of the two critical factors that define the exploitation potential of SM in transportation studies, we focus in this section on the SWOT of using SM data in transportation studies (see Figure 4). This analysis is based on findings from the pertinent scientific literature [Kietzmann et al., 2011, Kaplan and Haenlein, 2010]) and online blogs that discuss SM-related matters, such as Buffer Social (`www.blog.bufferapp.com`), Jenn's Trends (`www.jennstrends.com`), and Socially Sorted (`www.sociallysorted.com.au`). Starting with strengths, SM offers an opportunity to obtain a combination of user-generated textual, temporal, and spatial information (user-generated content and information provision) that does not rely on user recollection, while also expressing the current state of mind at generation. The activity-oriented use, within a scope of rich data provision, strengthens researchers' ability to investigate the activity space, taking into account the social network dimension. Furthermore, the intense streaming of information allows for dynamic approaches on an unprecedented amount of data produced. Finally, the cost of collecting data and performing research can be significantly reduced while ensuring a comparatively large sample size.

Numerous opportunities arise: researchers can increase their knowledge on mobility–related behavior, supplement or redesign explanatory and prediction models, and use SM data for transportation management. Finally, SM could provide a space for tailored surveys that would require less time to be completed and would be accompanied with users' characteristics.

However, the use of SM data for research could have disastrous weaknesses. For example, the data ownership scheme may contradict conventional expectations that data belongs to public entities. SM providers generally support research, but data collection, storage, and use is a gray area, and data exchange is prohibited. Moving

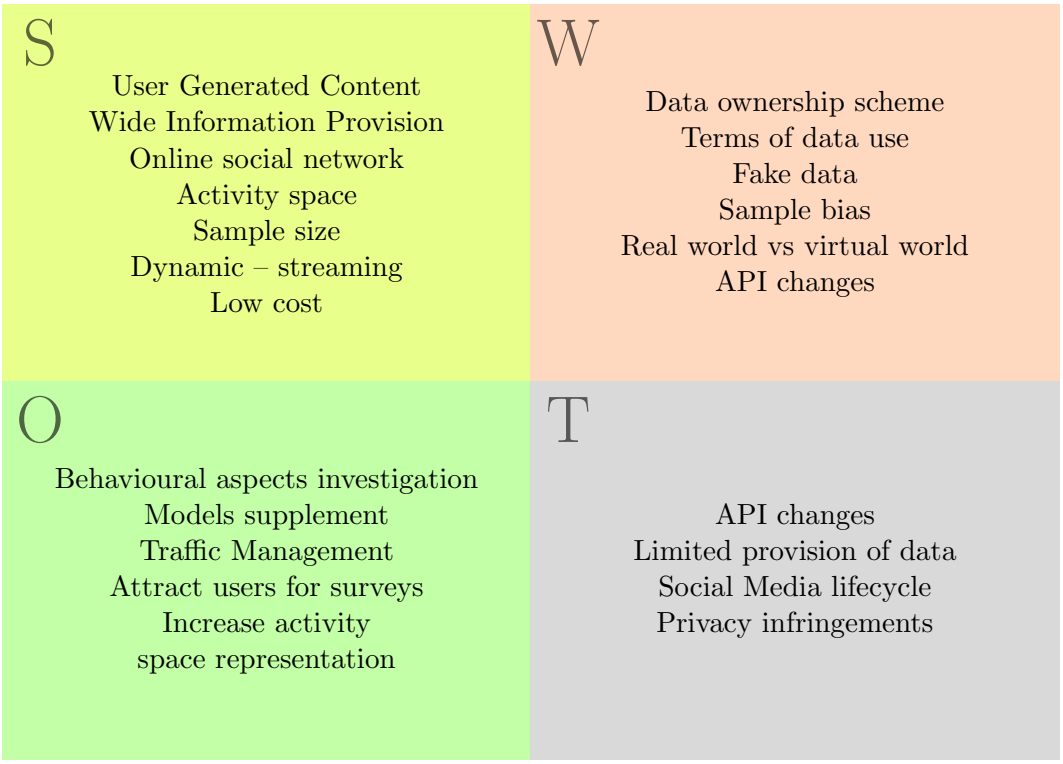


Figure 3.3: SWOT Analysis

along, SM users post intentionally misleading information (fake data) that degrades the information quality that can be achieved. Additionally, researchers have to deal with high information noise (such as nonsense posts) to acquire useful information. On another level, SM users form a rather misrepresentative sample of the population (sample bias). Researchers should consider the online/offline life differences and the ego bias that user-generated content bears; often, users can have a second (online) life that is different than their "real" life. Finally, SM providers in many cases redesign the API or the data it provides. This could result in quickly outdated research contributions that might be based on API data provision.

The main threats identified express issues of privacy, SM usage, data provision that might limit the use of SM, or make it inaccessible for research. SM emerged at a rather dramatic pace in a very limited time. We cannot yet be sure that they are going to be able to form a representative sample of the entire population or even if its use will reduced or abandoned. Furthermore, although research is considered as a neutral cause of collection, storing and processing data, it is not certain that SM would continue data provision as it might be identified by many as an invasion of privacy.

3.4 Privacy and Data Implications of Social Media use

Privacy implications of Social Media have long been examined by different angles. Two main streams can be identified: a) potential of privacy infringements through social media and b) changes of privacy perceptions through Social Media. Both are widely connected to transportation research and the use of social media in transport. Starting from the latter, it has been evidenced that Social Media has changed the culture of sharing information, especially for younger generations [Madden et al., 2013] Users are more likely to reveal their real name, talk about their interests and be open in revealing their thoughts and experiences. The reason why this is important for transport is that it relates to the difficulties associated with data collection. Users revealing true information could lower the value of privacy (Antoniou and Polydoropoulou, 2014); increasing the response rate on transportation related surveys and allow for the combination of Social Media (which provide an inexpensive stream of information) with other sources of data. On the potential of privacy infringements through social media, the case of Cambridge Analytica [Greenfield, 2018] was one of the cases that changed the landscape on Social Media privacy awareness and the need of safeguarding user privacy. One could argue that this was only the tip of the iceberg. Users providing geolocated information or information concerning activities could have adverse impact on their safety and their ability to safeguard their interests e.g. insurance premiums and employment [Sánchez Abril et al., 2012]. Although numerous measures have been taken from Social Media providers to safeguard user privacy, the possibility of privacy infringement is more relevant than ever, as new methods of user identification collection of user data emerge [Smith et al., 2012, Zheleva and Getoor, 2009]. Users of social media in many cases stand helpless against identification of possible treats and issues, reducing trust and overall creating problems, even to a healthy use of Social Media. With regards to the use of data from Social Media, there is a long-lasting discussion on the premises of data provision and inherent biases. Data provision is tightly connected to issues of privacy and the code of contact of using sensitive data. Data availability relies on the Social Media platform providers' policies and one of the main disadvantages of its use, is the uncertainty with regards to the continuation of its provision. Already in the last few years, and due to privacy concerns, many providers changed the policies concerning the use of APIs and limited the data provided (e.g. provision of number of visits from Foursquare, events visited data from Facebook) in ways that could restrict its use in Transport. With regards to biases, the main point raised with the sampling bias that related to the socio-demographics of Social Media users. With the establishment of Social Media platforms this seems to be alleviated (Rashidi et al., 2017); however, there is still the issue of why and when people use social media, which poses the question of the capabilities that Social Media data have to increase the observability of the transportation system.

4 Methods

Contents

Components of this chapter are presented in:

Emmanouil Chaniotakis and Constantinos Antoniou. Use of geotagged social media in urban settings: Empirical evidence on its potential from twitter. In *IEEE 18th International Conference on Intelligent Transportation Systems, (ITSC)*, pages 214–219. IEEE, 2015

Emmanouil Chaniotakis, Constantinos Antoniou, and Evangelos Mitsakis. Data for leisure travel demand from social networking services. In *4th hEART Symposium*. European Association for Research in Transportation - hEART, DTU, Denmark, 2015

Emmanouil Chaniotakis, Constantinos Antoniou, Georgia Ayfantopoulou, and Dimitriou Loukas. Inferring activities from social media data. In *Transportation Research Board 96th Annual Meeting*, 2017a

Emmanouil Chaniotakis, Constantinos Antoniou, and Konstantinos Goulias. Transferability and sample specification for social media data: a comparative analysis. In *Proceedings of the mobil.TUM 2017 Conference, 4-5 July, Munich Germany*, Jan 2017b

4.1 Data Collection

4.1.1 Generic Data Collection Methodology

There are some Social Media sites that can facilitate data collection, with each Social Media site imposing different limitations on the data structure, collection and terms of use, which might change without notice (see Chapter 3). However, in most cases covered by the definition of Boyd and Ellison [2007], there is a basic structure (of users, activities and relations) that could allow the collection of mobility-related data.

Data collection from Social Media is different in comparison to the structured data collection methods that are usually met in mobility. The data that is collected is of event-based nature i.e. whenever users post on Social Media. There is a basic distinction between different data collection models: some require the creation of an application or a group, while others, allow the collection of publicly available information. The process of selecting the data collection approach is presented in Figure 4.1. In case of an application-based data collection, users should be invited to share their activities, forming –in a way– a sample of users. In case of non-application based data collection, real-time data is collected for the formation of a users sample from which the activities is collected.

Although data from different Social Media platforms is explored in this thesis, the vast majority of data has been collected from Twitter. For this reason, the following paragraph is dedicated to Twitter based data collection.

4.1.2 Twitter Data Collection

Twitter is a micro-blogging site on which users publicly share information and discuss. The idea behind twitter is that users post their statuses and communicate with other people using twitter using short text messages, pictures and videos. Twitter allows for a generic random data collection, data collection based on geographic boundaries and getting users' timeline. For the former two, queries return a list of randomly selected latest tweets posted, while with the later, Twitter returns the user timeline, which in its basic form, refers to a twitter page, but can be extended using pagination. The Twitter specific methodology presented here is designed in order to overcome the limitations of the Twitter Application Programming Interface (API). Given the focus on mobility-related applications, we focus on the geographically bounded data collection and the collection of users' timeline.

The first data collection component (Real-time Data Collection) mainly focuses on the random data that is provided by the Twitter API. Extracting this data, is useful to a) evaluate data collection process and extract descriptive characteristics of the sample (for example Section 5.2); b) definition of a sample of users for mobility–related studies and for event –incident– identification. Specifically, this first component of the methodology (Figure 4.2) includes the coding and execution of queries at the Twitter API, in order to collect geotagged tweets, given a central point in the area to be examined and a radius. The latest tweets from the examined area are returned in JavaScript Object Notation (JSON) that is parsed to get information of the users that

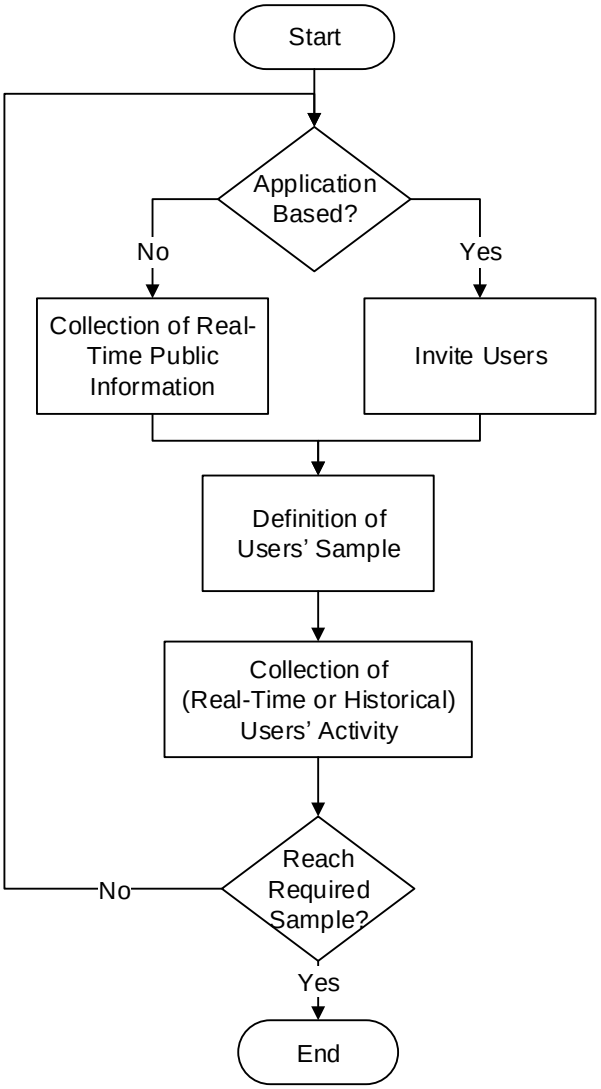


Figure 4.1: Generic Social Media Data Collection Methodology

tweeted and the related tweets. Aiming at collecting an adequate amount of information a selection of the JSON variables is pursued.

The second data collection component (Historical Data Collection) is used to collect historical data from the users that were sampled by the first data collection component (Figure 4.3). It is used periodically in order to collect a predefined number of latest tweets (N) posted by the users that were collected during the first phase of the data collection. This component is particularly important when the number of tweets posted by users at real-time in a certain area is higher that the tweets that can be collected (given the Twitter API limitations), as given the random selection of tweets to be returned, the vast majority of them is never queried. However, with the

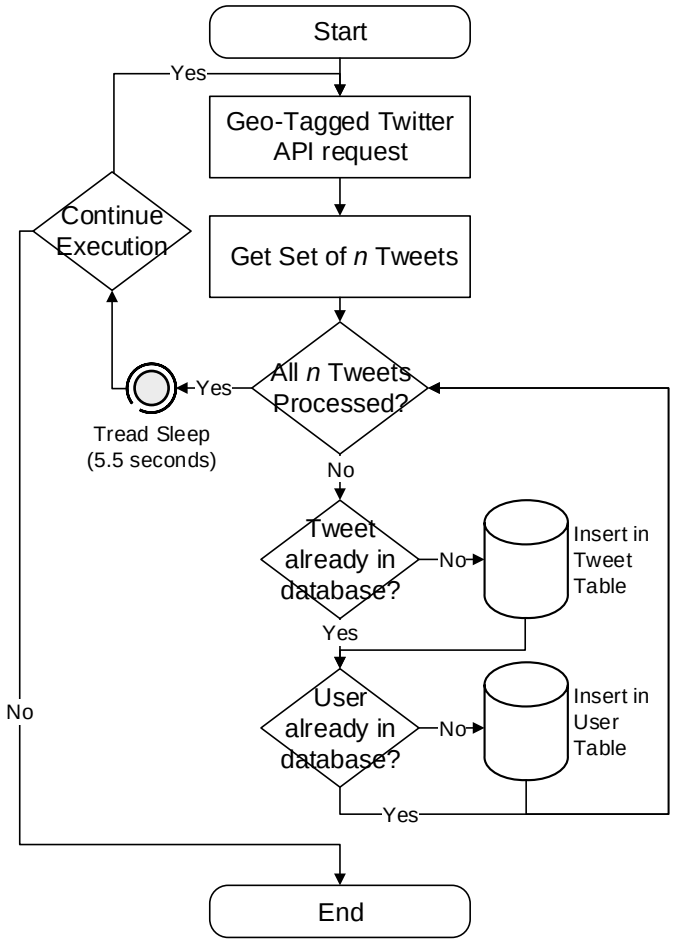


Figure 4.2: Real–Time Data Collection Method

random–time data collection process being executed for a long time, the historical data collection could return a larger number of tweets. In addition, it offers the opportunity to perform user-based analysis, and extract information useful for transportation and mobility (Section 6). The Historical Data Collection methodology follows the same collection principles as the Real-Time Data Collection methodology: queries are coded and executed with the main search term to be the user identity (`userID`). Again, due to limitations of the Twitter API, it is only possible to retrieve up to 200 tweets with each query. This limitation can be overcome using multiple queries per `userID`; using paging.

In this thesis the implementation took place following the documentation of the Twitter API and based on the use of the Twitter4J [Yamamoto, 2007] java library. The code the coding is implemented under the Java Runtime Environment, having

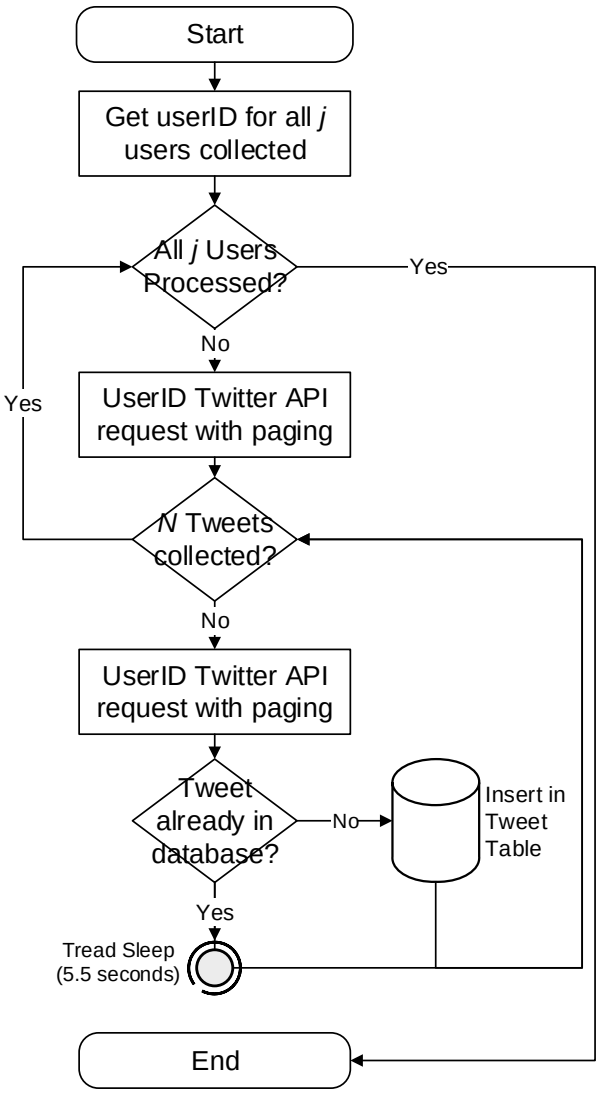


Figure 4.3: Historical Data Collection Method

different classes for different twitter–related objects (e.g. `user` or `tweet` class). A scheduler has been programmed to run a query every t seconds, with t to be defined based on the allowed number of queries per hour.

Various data validity tests need to be performed in order to store properly the data in a structured database. The first one is duplicates tests. Tweets and users are checked for already existing in the database or being new (based on unique IDs provided by Twitter). In case of being new, they are inserted in the corresponding tables of the database. This is performed by having a hashmap for the users and the corresponding

tweet identification number, which is updated with every query. Additionally, another validity evaluation test is performed to explore the location of users. In some cases, although geofenced the tweets return do not have coordinates. For initial implementations, only tweets with geographical information had been collected.

The resulted program has a configuration file which allows for variation of the data collection process or for querying based on specific arguments. The database in which the data is stored is a Structured Query Language (SQL) database, which consists of two main tables: the *User* table and the *Tweet* table.

4.2 Descriptive Analysis

Given the diversity of data available from Social Media, a wide spectrum of descriptive measures can be used to evaluate the information that can be acquired from Social Media data. As the pertinent literature on the use of Social Media data in transport has been fragmented, the descriptive evaluation of the potential of using Social Media data has been considered as one of major importance. The exploration of the descriptive characteristics takes place on the basis of spatial, temporal and textual characteristics. Apart from the widely used descriptive statistics (average, standard deviation, density and correlations) a comprehensive framework has been developed that encompasses the whole data analysis process followed.

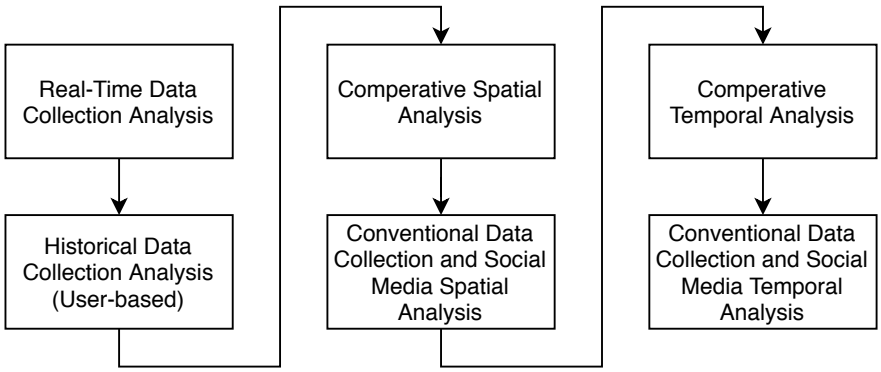


Figure 4.4: Descriptive Analysis Outline

The starting point is the exploration of the data collection characteristics. As the characteristics of use are different in different places around the world, a comparative study of different areas around the world is pursued. This step encompasses the exploration of the number of tweets collected, the number of users posting, the frequency of posts and the percentage of geotagged posts. Based on the sample collected for each area, a user–based analysis is performed to understand the user–related characteristics (such as number of tweets and percentage of geotagged tweets) for the cities compared. Towards this end, and given the vast use of Social Media for leisure and tourism activ-

ities, the categorization of users and the analysis of the characteristics is distinguished between residents and tourists.

Spatial analysis of the geotagged tweets is also performed to get insights with regards to the coverage of Social Media in the examined cities and the potential of using them for exploring urban characteristics and areas of activities. This is performed on the basis of spatial density, as well as the exploration of the global patterns of the users collected. For a selection of specific cases the descriptive analysis also focuses on comparing the different destination types and activities recorded in different data collection methods. This is primarily targeted towards the investigation of patterns of similarity and differences between conventional mobility–related datasets and data from different Social Media platforms, including aspects of exploring Traffic Analysis Zones (TAZ) attractions.

Finally, temporal analysis reveals similarities or difference between the different cities examined on the basis of posting distributions. These are explored on the basis of geotagged or not-geotagged posts. Of interest is again the comparison of the temporal distributions between different data collection sources, which is performed on the basis of temporal correlations in different time–of–the–day periods.

4.3 Data Enrichment

Data from Social Media can represent a variety of similar or different things. For that reason a major methodological step, considered to be important for the extraction of mobility patterns from Social Media data is the development of data enrichment methods. Let $\mathbf{S}$ be a complete Social Media dataset. Given a specific amount of information $I_{\mathbf{S}}$, data enrichment is defined as the process to increase the information by either combining the dataset with external datasets or extracting information from it such that the posterior information $I^k_{\mathbf{S}}$ is higher than the original $I_{\mathbf{S}_{post}} > I_{\mathbf{S}}$. The simplest form of information enrichment is adding in the analysis variables which are already included in the dataset, for example assume a dataset which includes spatio-temporal information. The inclusion of a "user" variable to distinguish the different users of the system is a rather straightforward data enrichment process. This notion of data enrichment can be rather useful in Social Media research. One could imagine a situation where multiple people visit the same location and post similar (for example food related) statuses. When combined (based on spatial and textual information) these posts could provide a larger set of documents describing this location visited.

Information enrichment takes place on a measure of connection (e.g. similarity) which measure the extend to which variables can be considered similar or not. Similarity is commonly measured in terms of correlation and distance (e.g. Euclidian Distance in Equation 4.1 or Pearson Correlation in Equation 4.2).

$$D = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(A_i - B_i)^2} \tag{4.1}$$

$$P = \frac{\sum_{i=1}^{n}(A_i - \tilde{A})(B_i - \tilde{B})}{\left(\sqrt{\sum_{i=1}^{n}(A_i - \tilde{A})^2}\right)\left(\sqrt{\sum_{i=1}^{n}(B_i - \tilde{B})^2}\right)} \tag{4.2}$$

It becomes obvious that data enrichment can be performed on the basis of clustering. This could be done for data enrichment of two or more different datasets or for data enrichment within the same dataset, based on a collection of variables which illustrate within variable similarities. For Social Media data enrichment can be pursued on various different variables. These could be defined upon spatial, temporal and textual variables or a combination of them. On a user's basis, it is possible to extract habitual patterns and activities. Let $\mathbf{S}$ be the set of data collected from social media and $\mathbf{s_k} \subseteq \mathbf{S}$ the subset for the random user k. Clustering of the locations commonly visited by an individual could result on the identification of the various areas from which someone frequently posts on Social Media. Given a set of observations $\mathbf{s}_{(1,k)}, \mathbf{s}_{(2,k)}, \ldots, \mathbf{s}_{(n,k)} \in \mathbf{s_k}$ one come up with clusters of locations for each user (e.g. using k-means as presented in Equation 4.3).

$$\arg\min_{\mathbf{K}} \sum_{i=1}^{c} \sum_{\mathbf{s_k} \in K_i} \|\mathbf{s_k} - \boldsymbol{\mu}_i\|^2 \tag{4.3}$$

where μ is the average of points in K_i with K_i to be representing a set of clustered points and c is the number of clusters. For the case of spatial data, a much more informative way of clustering, is the use of Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [Ester et al., 1996]. Locations are characterized based on their spatial distance and their density, in order to specify locations that are visited by each user more than a specified number of times. Locations visited less than a specified number of times and under a specified density are characterized as noise. Towards the direction of enriching the information that can be extracted from Social Media data, it is possible to combine the textual data available by each user from a specific location, assuming that users visiting a location more than m times are actually performing a similar activity at that location (e.g. being at restaurant or at work).

Additional data enrichment activities could be pursued with the use of links that are commonly present as textual information. For example, based on a Twitter dataset, extraction of information from Foursquare posts integrated in Twitter can be performed to achieve the enrichment of activities–related information. In order to parse the information, the extraction of URL patterns using regular expressions, the extension of shortened links to long links and Foursquare queries is performed. The queries' response is of JavaScript Object Notation (JSON) format and it can be either a venue, or a check–in. In some cases it requires for the link to be (within the Foursquare query) resolved. Finally, the response can be parsed in tabular form and stored with the initial tweet information.

The generic process followed for the data enrichment is presented in the following diagram. First the target variable is identified and analysis of the influencing factors

(independent variables) takes place. Then Exploration of similarities is performed within the dataset or from external sources. Finally, different observation are combined.

One of the novel methodologies included in this thesis is directed towards the data enrichment for predicting activities performed using textual data, namely the **User–Centric Activity Enrichment**. Given a dataset for a user, we perform user–based enrichment to increase the amount of textual information available for each location visited frequently. Let $\mathbf{s_k}$ be the collection of Social Media post of user k, where $t_{(i,k)} \in \mathbf{s_k}$ is the text and $l_{(i,k)} \in \mathbf{s_k}$ the location of post i, of this user. Also let $\mathbf{s'_k} \subseteq \mathbf{s_k}$ exhibiting a high degree of internal similarity of some variables (e.g. $l'_{(i,k)} \in \mathbf{s'_k}$ or $t'_{(i,k)} \in \mathbf{s'_k}$). The textual information available for this similar spatio-textual doublet could be combined in an effort to have more text, representing the same activity performed. The textual information available with regards to an activity performed by a user is defined:

$$a_k = \begin{cases} t_{(1,k)} \cup t_{(2,k)} \cup \cdots \cup t_{(i,k)} & \forall i \in \mathbf{s'_k} \\ t_{(i,k)} & \forall i \notin \mathbf{s'_k} \end{cases} \tag{4.4}$$

The resulting dataset from the data collection and dataset definition methodology is used for the extraction of information on the nature of activities users perform and post about in Social Media. The methodology is based on the notion that there are places that users visit frequently to perform activities (e.g. home, work) namely recurring activities, and other places which are scarcely visited or only visited once, for non–recurring activities (e.g. leisure). This distinction of places visited and –as a consequence– activities performed is considered of high importance, for the data generalizations that are required aiming at both data reduction – in terms of the identification of locations that are commonly visited – and enrichment – in terms of generalization of activities identified.

4.4 Feature Extraction

One of the main objectives of this research is the extraction of information with regards to the transportation system. As such methodological and practical contributions are included which allow for the use of Social Media in transportation studies. One of the latest major transportation modeling necessities is the development of activity–based models. Although at the frontier of transportation demand modelling, activity–based models present many challenges especially concerning the data required for representing individuals activities [Cascetta, 2009]. In this direction, the incorporation of Social Media is appealing as their data can be utilized to provide the information on the activities users perform.

4.4.1 Areas of Interest

A starting point of this analysis is the extraction of areas where activities are concentrated, with the aim to be the evaluation of areas where similar activities take place.

This has been performed by extending the concept of Natural cities [Jiang et al., 2015] in order to extract natural Areas of Interest (AoI) using Social Media spatial data.

The natural definition of cities is based on the fact that in many cases, data illustrate an imbalance that can be described by a heavy–tailed distribution (L-shaped) [Jiang and Miao, 2014]. The data is divided in classes iteratively, until the data from classes do no illustrate a heavy–tailed distribution. The break point in every iteration is the arithmetic average. This methodology is named head/tail methodology. It should be mentioned that this method is a classification technique that is primarily intended to be used for the extraction classes for spatial data. In this study we focus on the part of the data characterized as the head of the distribution and we use them to identify locations which have a high density of Tweets. This method can me applied for any subset of the original data, e.g. for specific types of activities (such as leisure and shopping) or for specific user classes (such as residents and tourists).

The implementation is based on the distances between different spatial points. At first, the definition of Triangulated Irregular Network (TIN) is performed, which connects all points with lines and creates triangles. Then the length of those lines is used for the identification of areas, defined by lines of small length and essentially indicating places of high density posting activity.

4.4.2 Text Based Feature Extraction

As discussed above, the existence of spatial and temporal characteristics can by default provide information concerning the transportation system. Not limited to that, textual and user related information (available either from posts, or from the user profile) enrich the information provided and increase the potential to be utilized towards newer solutions to mobility–related problems [Grant-Muller et al., 2014, Gal-Tzur et al., 2014a]. Pertinent literature suggests a wide spectrum of Social Media analysis in transportation. However, to the authors' knowledge, only a few focus on users' activities [Lee et al., 2016, Noulas et al., 2013, Hasan and Ukkusuri, 2014], with most of the studies to focus on the extraction of information following what we will refer to as a post–centric approach. In post–centric analysis, each Social Media post is viewed as an entity, ignoring in most cases the user perspective.

This is believed to disorient from the necessity to merge the online and offline space that could potentially offer useful information towards a better representation of the transportation system and specifically individuals' activities. On the other hand, user–centric approaches can provide valuable interpretations on elemental research questions, such as why a user posts on a Social Media platform, which activities users perform and how online activity is linked with users' offline life.

For this reason, an efficient user–centric framework for the inference of users' activities is developed. The framework is based on the process of data enrichment to a) automatically retrieve labels from external sources (e.g. Foursquare) and b) enrich the textual information related to each label, by combining places visited frequently by the same user (Section 4.3). Based on the known activities and the Twitter text, classification algorithms allow for the exploitation of activities from Twitter text.

The outline of the suggested methodology is presented in Figure 4.5. First, locations are characterized based on their spatial distance and their density, in order to specify locations that are visited by each user more than a specified number of times. The classified locations receive a characterization based on the venue categories they belong to. The characterized –based on recurrence– tweets combined with venue and temporal characteristic are then used for the definition of document–term matrices specified to each type of activities that can be performed in venues. This allows for direct classification of each tweet describing an activity, in an activity cluster using classification algorithms. It should be noted that not all tweets are used to describe activities.

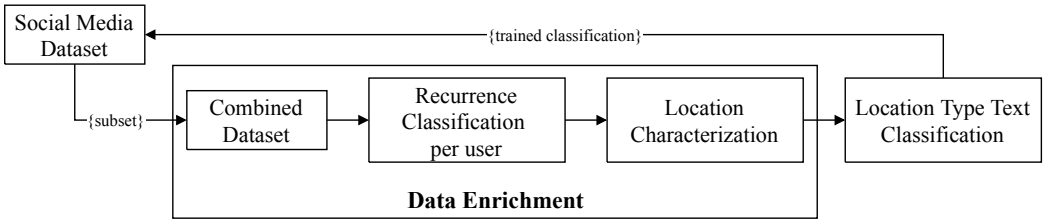


Figure 4.5: Overall Method including Data Enrichment and Activity Extraction

The main feature upon which we perform the extraction of activities is text. Thus a number of steps needs to be followed before performing the actual text classification (such as normalization, tokenization and removing stop words). In the following paragraphs the specifics of text (pre)processing and text classification are discussed.

4.4.2.1 Text Processing

The process to be followed when dealing with text is rather standard. It includes a variety of pre–processing measures that allow for the extraction of meaningful information. The process to be followed is presented in the following diagram. It is assumed here that the text to be used for classification is the outcome of the data enrichment method (a_k – see Equation 4.4), however the process followed is rather standard with any type of documents / textual information.

Tokenization Tokenization refers to the process of splitting text into smaller entities, which in many cases are words. In most languages each word is considered to be the minimum entity that has a semantic meaning, thus each token refers to a specific word. In most cases this is performed with rather straightforward operations, such as splitting on text space, or commas or dots, which indicate word boundaries. It is worth noting that in some cases the exploration of sequences of words could be useful as they might have a distinct meaning (e.g. expressions or phrasal verbs). This is commonly referred to in the pertinent literature as n–grams with n to refer to the number of words combined. For the case of social media this is generally straightforward although in some cases (e.g. with hashtags) there might be the need to have a dictionary based tokenization process.

Normalization Normalization commonly refers to the operation of transforming tokens in a more standardized format. The need of normalization emanates from the need to properly extract words of same semantic meaning (e.g. “eating” and “eat”; “As” and “as”) and also have a vocabulary of reduced size.

The first and probably simplest normalization is making all letters of same case (lower– or upper–case). The second and rather more sophisticated method used is called stemming and refers to the normization of words which have different morphological variations, but they actually refer to the same word. This process is referred to as stemming. Several stemming methods exist, from the basis rule–based stemming (e.g. if word ending with “ing”, remove “ing”), or with techniques such as lemmatization, which searches for the root of each word in a dictionary.

Stop–words Another important pre–processing step is the removal of stop–words. Stop–words commonly refer to words of low semantic importance such as “the”, “a”, “an”, “on”. These words are removed to reduce the size of the vocabulary used and the computational time. The removal of stop–words is most commonly performed on a rule–based approach, based on collections of stop–words commonly available.

Document Term Matrix Documents and generally text needs to be represented in a mathematical form for it to be used in classification or modelling. One of the most common ways of doing this is referred to, in the pertinent literature, as Document Term Matrix (DTM). DTM represents each document (or each social media post or each collection of tweets) as a bag–of–words. Meaning that the document has a collection of words (corpus) that is structured based on the occurance of each word. A DTM is a matrix which has the documents as row and the tokens as columns. An example is presentedin Figure 4.6.

Document 1: A nice day today!

Document 2: I will go out

day	go	nice	out	lets	today	will
1		1			1	0
	1		1			1

Figure 4.6: An example of a Document Text Matrix

The main advantage of such a representation is that mathematical operations, such as algebraic operations and modeling is possible; essentially changing the format from strings to numbers.

4.4.2.2 Text Classification

After following the steps of text processing and the definition of the Document Term Matrix, the textual information is now ready to perform the necessary text classification. There is a basic distinction on performing classification: a) supervised or b) unsupervised. The distinction is related to the availability or not of a labeled variable which is actually the variable to perform the classification upon. For example in the case that we are focusing, the labelled variable is the location type visited by individuals, while a commonly unsupervised classification technique is topic modelling, which, based on a set of documents finds the topics of each document (in terms of most representative words).

In the following paragraphs the methods used in this thesis are shortly described. These refer to some of the most widely used Machine Learning Techniques, thus there is a large number of textbooks which describe these processes and as a consequence, for more information the reader is referred to them [e.g. Shalev-Shwartz and Ben-David, 2014].

Support Vector Machine Support Vector Machine is a supervised classification algorithm for linear and non–linear classification and regression problems that works with the definition of optimally selected hyperplanes [Friedman et al., 2001]. The Support Vector Machine algorithm works on the basis of margins. Hard margins are defined in case the data is separable in a perfect way; with the objective to be the maximization of the margin (thus selecting the parallel lines with the largest halfspace). Soft margins are introduced in the case of data not being directly separable, in a way that the constrains can be violated to some extend.

Generalized Linear Models with Penalized Likelihood Function Another very widely used family of classifiers is the Generalized Linear Models with Penalized Likelihood Function. These methods extend the classical linear regression estimation (e.g. Ordinary least squares - OLS) to perform regularization. Regularization for linear regression is generally of two forms: a) Lasso and b) Ridge [Friedman et al., 2001]. Ridge regression is shrinks the coefficients of correlated predictors towards each other; while Lasso tends to pick one and ignore the rest. A method comprising both is named elastic-net [Friedman et al., 2010], which defines a compromise between the two regularization. Using the usual regression setup, let $\mathbf{x}^i, y_i, i = 1, 2, \ldots, N$ be the predictor ($\mathbf{x}^i$) and response variables y_i variables, where $\mathbf{x}^i = (x_{i1}, x_{i2}, x_{i3}, x_{ip})^T$. Generalized Linear Models with Elastic Net Penalization solves the following optimization problem:

$$\min_{\beta_0,\beta} \frac{1}{2N} \sum_{i=1}^{N} (y_i - \beta_0 - \beta x_i^T)^2 + \lambda \left[(1-\alpha)\frac{1}{2}||\beta||_2^2 + \alpha||\beta||_1 \right] \tag{4.5}$$

where the first part ($\frac{1}{2N}\sum_{i=1}^{N}(y_i - \beta_0 - \beta x_i^T)^2$) refers to the common linear regression optimization problem and the second part ($\lambda\left[(1-\alpha)\frac{1}{2}||\beta||_2^2 + \alpha||\beta||_1\right]$) refers to the elastic-net penalty. As mentioned above, this second term is responsible for the

regularization of the likelihood function in a compromise between the ridge ($a = 0$) and lasso penalty ($a = 1$).

Maximum Entropy Classification Maximim Entropy classifiers estimate the conditional distribution of the label variable to the actual document terms [Nigam et al., 1999]. It is a probabilistic Machine Learning technique that uses least informative priors (e.g. uniformly distributed) and imposes priors with regards to available information. For example assuming that there are n classes of activities that can be identified from Social Media data and that in a dataset 90% of Social Media posts that include the word "food" referring to a visit to a restaurant. It could be understood that when the word "food" is present there is a 90% chance of this post to be of the restaurant activity class. If not, then there is no information about it and the post could have a uniform prior class distribution ($1/n$). As Nigam et al. [1999] mentions the complexity increases with the increase of the dictionary used, thus the constraints derivation needs to be defined, with an example to be the use of word–class combinations (if a word is included in more than one Social Media post, then it has a higher weight for the specific word–class pair).

4.4.3 Activity Space

Another interesting field to which Social Media could be useful involves the investigation of the activity space that individuals visit. Conventionally, activity spaces are estimated as the convex hull of the locations visited by individuals. This allow researchers to investigate the actual area in which individual live, thus it can be considered as a measure of mobility coverage. For many years, and given limited resources the definition of the activity space was based on a number of observations that were performed usually in the time–span of a few days (e.g. 14 days). As discussed in Lee et al. [2016] short time–spans might not really represent the actual activity space evolution; thus the exploration using other sources of data could be assisting towards this direction, as Social Media provide information from the same individuals for a course of years. Consequently, each user sub-dataset could include traveling in different countries or visiting places once in a few years or just being at a particular location once in a lifetime.

Although rather important, this could also have negative effects to the estimation of activity spaces, as there are mainly derived to explore habitual travel patterns. For this reason, a methodology is introduced that defines activity spaces through a two steps process:

1. investigates the characteristics of frequently visited places by individuals (location recurrence) using clustering techniques, and

2. performs the same clustering analysis on a local scale for the definition of the total activity space.

These two different analyses combined provide a much better understanding of the activity space and the resulting habitual patterns of individuals, allowing for a clearer evaluation of the travel patterns that we observe. In both cases the application of the Density-based Spatial Clustering of Applications with Noise (DBSCAN) algorithm is used. This is performed on the basis of different user groups (i.e. Residents; Tourists and as a whole) to better understand where exactly these people move with regards to each city.

5 Social Media and Transportation

Contents

Components of this chapter are presented in:

Emmanouil Chaniotakis and Constantinos Antoniou. Use of geotagged social media in urban settings: Empirical evidence on its potential from twitter. In *IEEE 18th International Conference on Intelligent Transportation Systems, (ITSC)*, pages 214–219. IEEE, 2015

Emmanouil Chaniotakis, Constantinos Antoniou, and Evangelos Mitsakis. Data for leisure travel demand from social networking services. In *4th hEART Symposium*. European Association for Research in Transportation - hEART, DTU, Denmark, 2015

E. Chaniotakis, C. Antoniou, J. M. S. Grau, and L. Dimitriou. Can social media data augment travel demand survey data? In *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1642–1647, Nov 2016a. doi: 10.1109/ITSC.2016.7795778

Emmanouil Chaniotakis, Constantinos Antoniou, Georgia Ayfantopoulou, and Dimitriou Loukas. Inferring activities from social media data. In *Transportation Research Board 96th Annual Meeting*, 2017a

Emmanouil Chaniotakis, Constantinos Antoniou, and Konstantinos Goulias. Transferability and sample specification for social media data: a comparative analysis. In *Proceedings of the mobil.TUM 2017 Conference, 4-5 July, Munich Germany*, Jan 2017b

5.1 Introduction

As discussed in Chapter 4, the exploration of the generic characteristics of Social Media use and the connection to the usually investigated transport related features is very important for understanding the potential of their use in Transport Studies. This is performed on the basis of descriptive analysis 4.2, including comparisons to conventional transport–related data.

Initially a generic analysis of Social Media data takes place, with the focus to be the comparison of the data collection and the data characteristics in different places around the world, using the Real–Time Data Collection method for Twitter (Section 5.2). This allows the evaluation of the potential in terms of data availability, and provides a generic understanding of the factors that affect Social Media. Additionally, the more thorough exploration of one data collection case (the case of the city of Athens) is pursued as an example of what is to be expected when collecting Social Media data. The historical data collection is then explored, to gain insights with regards to the way that people post.

The exploration of the temporal and spatial distributions of Social Media posts is included for the various cases, to better understand mobility patterns within cities and on a global scale with regards to space and per hour and per day of week with regards to time. Finally indicators of the connection between Social Media and Travel Survey is presented. This allows for the investigation of the differences and similarities that are observed from the different data collection types and also gives indication of the merits of using social media in transportation.

5.2 Social Media Data Collection

The data collection for the comparison of the different cities was performed by first deploying the Random Data Collection (Section 5.2.2.2) process that collects random tweets from Twitter (essentially forming a users' dataset), and then based on some filtering criteria, proceeded with the Users-based Data Collection (Section 5.2.2) that collects a number of the latest tweets from each user. For the extraction of information concerning the Social Media usage, data has been collected for 4 major cities in the United States of America (Los Angeles, New York, Orlando, Seattle) and 6 major cities in Europe (Amsterdam, Athens, Copenhagen, London, Munich, Paris). The selection of the particular cities was based on (a) the ability to extract information from textual characteristics (based on known languages), (b) the indicated Social Media usage, (c) the relatively large size of the cities, and (d) the diverse characteristics (in terms of size, demographics and Internet penetration).

5.2.1 Real–Time Data Collection

5.2.1.1 Real–Time Data Collection from Twitter: Comparison of Cities

In order to make the data collection as homogeneous as possible at the time of this study, a rather small Random Data Collection period was specified (approximately 2 months); resulting in mostly collecting a users sample. The data collection was performed using the Twitter REST Application Programming Interface (API) and by utilizing the Twitter4j library within a Java program that automatically collects data based on the latitude and longitude of a central point and a radius (as presented in Section 2.1). It should be noted that the Twitter API returns both geo-referenced (geotagged) tweets as well as tweets without geo-reference (not-geotagged). Additionally, the Twitter REST API returns a limited amount of tweets per query (200 tweets) and has a time-quota of 180 queries per 15 minutes (1 query per 5 seconds). In Table 5.1 the results from the initial data collection are presented. As it is clearly evidenced, the use of Social Media in the USA (at least within the examined period of time) is much higher than the use in Europe, something that agrees with the statistics (pewinternet.com).

5.2.1.2 Real–Time Data Collection from Twitter: the Athens Case

The data was collected in 35 days during January and February 2015 (30-01-2015 to 24-02-2015). The implementation was programmed in Java Runtime Environment using the Twitter4J library [Yamamoto, 2007]. During that period, approx. 550 thousand queries were made to the Twitter API in an cyclical area centred at the city centre of the Athens and a radius of 30 kilometres. The analysis bellow aims at investigating the way that a users' sample of tweeter users is formatted, focusing on the temporal characteristics of the data collection process.

The empirical distribution of the number of new users is fit by a Negative Binomial distribution and the results are presented in Figure 5.1 (both Probability Density Function and Cumulative Distribution Function). As it is clearly indicated, there is a high probability collecting tweets from users already inserted in the database (approx. 80%), while only approx. 20% of the queries receive one or more new users.

In total, 6737 unique users were collected, posting 89,692 (unique) geotagged tweets. Regarding the users that were collected, an increased number of inserted users was observed during weekends with the peak being on Friday. On the contrary, the tweets inserted in the database, illustrated a nearly uniform distribution with a slight decrease during weekends. The performed hourly analysis revealed a peak around 14:00 and 15:00.

The new hourly collected tweets were found to follow a different pattern, with approx. 25% of the total number of new tweets to be collected between 19:00 and 21:00 (peak observed between 20:00 and 21:00). The differentiation among the number of new users and the number of new tweets is interpreted by the fact that the increase of the users' sample size is not related to the number of users who tweet systematically. For this reason, we further the analysis of the tweets posted by each user during the period of the data collection. It is found that the mean of the number of tweets posted by each

Table 5.1: Tweets Data Collection

City	Data Collection Days	Geotagged Tweets Per Day	Not Geotagged Tweets per Day	Percentage Geotagged Tweets per day
Amsterdam, NL	14	1171.7	117597.6	0.987
Athens, GR	594	928.8	NA	NA
Copenhagen, DK	30	459.9	54328.7	0.839
Orlando, USA	199	454.2	NA	NA
Seattle, USA	123	773.9	61783	1.237
Munich, DE	11	3551	447178.7	0.788
New York, USA	664	16959.2	NA	NA
London, UK	106	202	30417.1	0.660
Los Angeles, USA	14	6385.8	1208561.3	0.526
Paris, FR	195	3.6	233085.9	0.002

NA: Refers to random data collection processes collecting only geotagged data

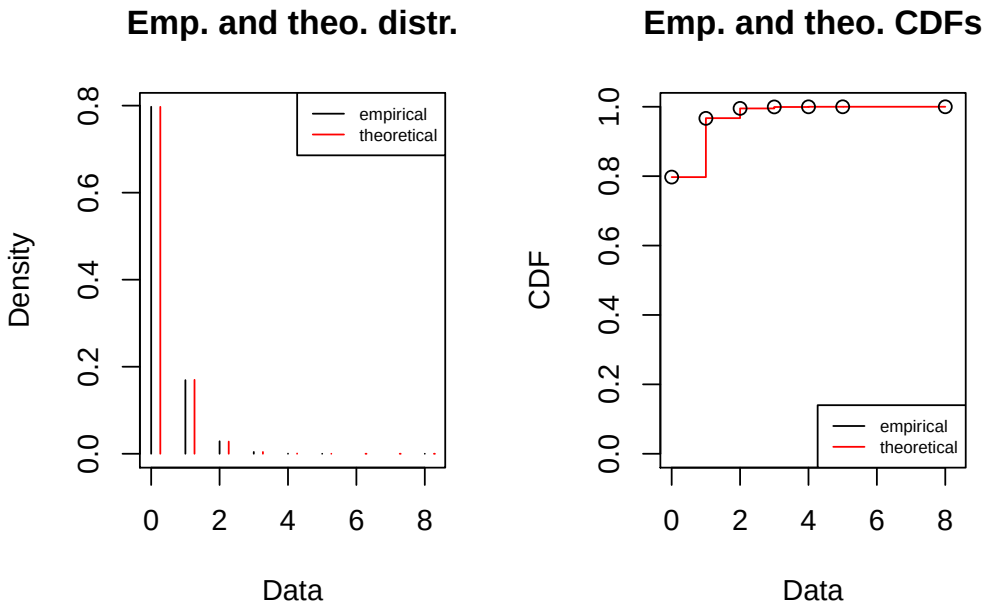


Figure 5.1: Probability Density Function and Cumulative Distribution Function of New Users

user is 12.86 tweets with a standard deviation of 65.01. It is important that there are some users who post tweets on a considerably higher rate that reach 2130 tweets in the examined period (35 days).

Given the data collection characteristics and in order to get a better understanding of the factors which affect the data collection process, we derive a Single Output Multiple Input Regression Model using solely temporal factors. The number of new users collected at time t (U_t) is thus given by (t–values given in parenthesis below the estimated coefficient values):

$$U_t = \underset{(37.78)}{1.81} - \underset{(-34.73)}{0.200} \cdot \log C_{t-1} - \underset{(-5.87)}{0.036} \cdot W_t + \underset{(35.45)}{0.213} \cdot H_t \tag{5.1}$$

where C_{t-1} is the cumulative number of users collected until time $t-1$, W_t is a dummy variable for the set of days E which includes Monday, Tuesday, Wednesday, Thursday; equal to 1 in case of $t \in E$ and H_t is a dummy variable for a set of hours (Q) which includes the hours after eleven o'clock in the morning to midnight; equal to 1 in case $t > 11$. Note again that the model is developed solely for matters of understanding the collection process and cannot be used for any application such as prediction, as the sampling period is rather small and there might be patterns of seasonality that are not included.

5.2.2 Historical Data Collection

Based on the collected dataset, a random sample of at least 1000 users was selected for each city, to collect their twitter history. The selection of the particular user sample was solely based on one criterion: the users should have posted at least two geotagged tweets within the examined data collection period. The selection of this minimum number of geotagged tweets was based on the general data collection characteristics, allowing to collect, for all cities, at least the required users to examine, while confirming that they are using their account for —among others— posting geotagged tweets. The small number of users and the random sample selection aim at the exploration of the potential of using Social Media in transportation studies. The data collection was performed by extracting the latest tweets from the time-line of each user [Chaniotakis and Antoniou, 2015]. For each user the last 600 posted tweets were collected. The data collection process was again performed using the Twitter REST Application Programming Interface (API) and by utilizing the Twitter4j library within a Java program for the collection of tweets using Twitter pagination. It should be noted that the user-based data collection does not include the tweets collected from the random data collection process. This might result in users which have not posted a geotagged tweet in the last 600 tweets, but on the other hand allow us to better compare users and cities.

5.2.2.1 Historical Data Collection from Twitter: Comparison of Cities

For the analysis of the characteristics several indicators were selected to be used in order to allow for the adequate comparison of the users. Descriptive statistics were explored for the generic understanding of Twitter use in the examined cities (Table 5.2). As it is clearly observed, the users that were collected are in general active twitter users with a large number of tweets posted. The percentage of the number of geotagged tweets posted in each case differ with cities in the USA to have a range of geotagged tweets that is higher than that observed in European cities (32.9% to 48.4% in the USA vs. 11.7% to 29.2% in Europe). When comparing the mean percentage of tweets posted in each city respectively to examine the number of users that are using Twitter to post geotagged information in the city of residence, it is rather clearly evidenced that again, there is a clear difference between the USA and Europe. Specifically, the highest percentage of geotagged tweets performed in the examined city (to the total geotagged tweets) is found in New York (73.9%), while the lowest is found in Copenhagen (24.3%). Another apparent difference between the collected data in Europe and the USA is the percentage of the users who did not post any geotagged tweets. The maximum percentage of the non-geottaged tweet in European cities is 21.1%, London, while in the USA, it is in Orlando with only 2.1%.

5.2.2.2 Historical Data Collection from Twitter: The Athens Case

The implementation of the second methodology presents some interesting results, as well for the case of Athens. In the data collection setup, the maximum number of

Table 5.2: User Based Data Collection

City	Number of Users	Users with non–geotagged Tweets (%)	Mean Number of Tweets	St.Dev. Number of Tweets	Mean Geotagged (%)	St.Dev Geotagged (%)	Mean In Examined City (%)
Amsterdam, NL	1127	0.2	556.9	124.1	29.2	27.9	33.9
Athens, GR	2092	12.9	576.7	87.8	22.6	24.8	31.9
Copenhagen, DK	1739	4.1	575.1	90.3	28.3	26.7	24.3
London, UK	2153	21.1	591.4	58.2	11.7	17.6	42.6
Los Angeles, USA	2313	0.0	532.1	160.3	44.3	34.1	64.4
Munich, DE	1389	1.9	545.2	140.3	26.8	27.5	28.5
New York, USA	1997	0.1	566.2	119.3	48.4	32.3	73.9
Orlando, USA	2748	2.1	545.4	142.1	34.3	29.9	35.6
Paris, FR	3856	5.7	583.8	71.3	25.04	25.5	35.3
Seattle, USA	1852	0.1	532.3	155.9	32.9	30.7	47.8

tweets to be collected from each user (N) was defined to the 1000 last tweets for a intermediate sample of 4082 users, collected the first 20 days of the data collection period. This analysis allows for a better understanding of the temporal dimensions of the tweet-posting process, as the sample of tweets requested is well distributed among users of the sample (about 1000 per user). In absolute numbers 2,622,113 new tweets were collected, from which 17% (460,003 tweets) were geotagged. The deviation from the target number of tweets (4,082,000 tweets) is attributed to the fact that not all users had posted 1000 tweets. Note that the second methodology component does not impose a restriction on the geographic area (in the radius defined at the first step of the methodology); as such, tweets from areas visited by users were collected.

The oldest geotagged post collected was posted on September 2010 while the oldest post (not geotagged) was posted on June 2007. This observation allows for same basic assumptions on the usage of Twitter: there are users that use periodically the service who have not posted 1000 tweets in 8 years while there are others who post 1000 tweets in a week.

On the analysis of posting habits per day, most geotagged tweets per day are posted during weekends (61650 average total geotagged tweets collected per day on week days, 71134 average total geotagged tweets collected per day on weekends). On the contrary, most not–geotagged tweets are posted on week days (326105 average total not–geotagged tweets collected per day on week days, 71134 average total not–geotagged tweets collected per day on weekends).

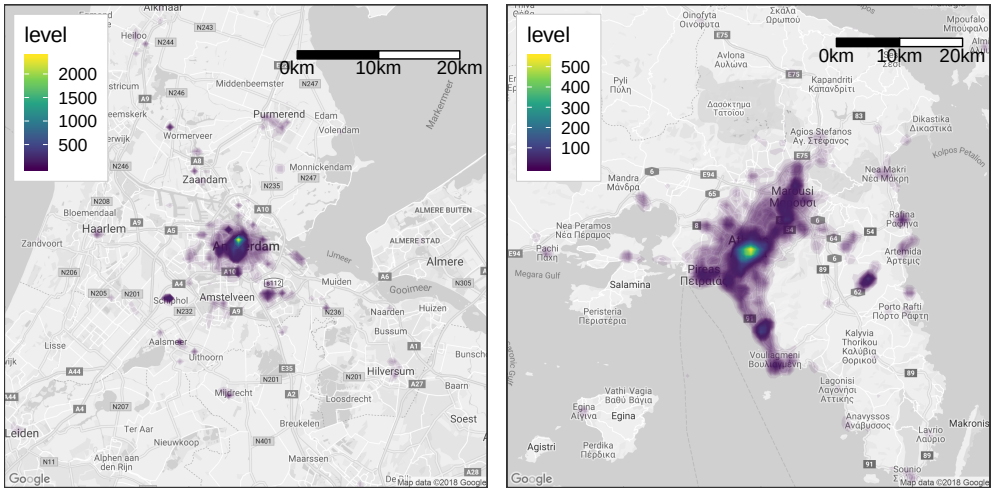
5.3 Spatial Analysis

5.3.1 Comparative Analysis

The geotagged tweets seem to widely cover the cities examined, as presented by the spatial density plots in Figure 5.2 and 5.3. The observed coverage confirms the fact that even with a small number of users, Social Media (and particularly Twitter) could be used to extract information with a rather large concentration in the central areas of cities, but also on suburban areas.

Interestingly, the collected tweets illustrate the differences in the way that cities in the USA and Europe are formed. In particular, it is evidenced that cities in Europe are structured within a denser populated center, while cities in the USA are spread in space, creating multiple centers where individuals dwell and perform activities. This is particularly evidenced in Los Angeles, Orlando, and Seattle, while New York illustrates a concentration of tweets in the Manhattan area and it is clearly a reflection of urban form and business establishments structures in the different regions examined.

Such an analysis of the mono/polycentric urban structure observed is particularly interesting for mobility, as there is strong evidence that urban form is related to travel patterns [Stead and Marshall, 2001]. In particular, the mono/polycentric structure has been found to relate to mode choice and distance traveled [Lin et al., 2015, Schwanen et al., 2001].

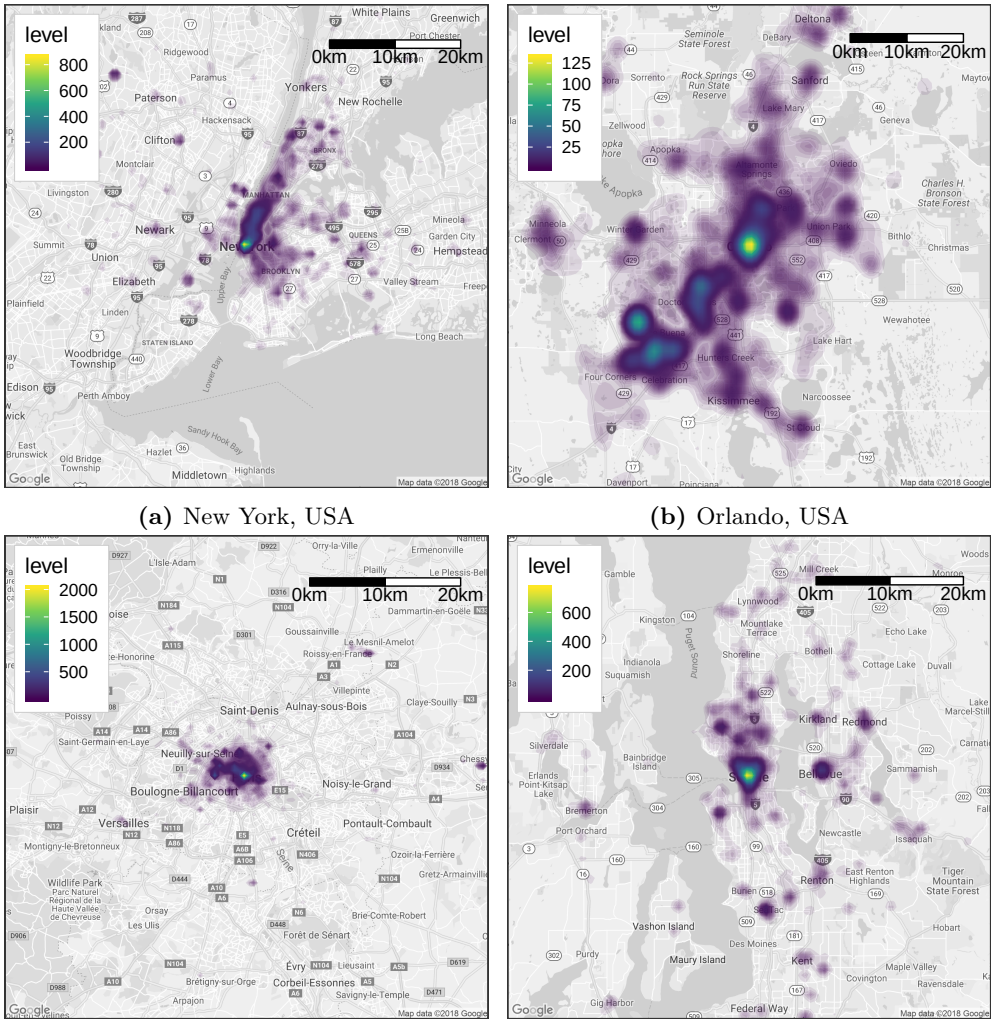


(a) Amsterdam, NL **(b)** Athens, GR

(c) Copenhagen, DK **(d)** London, UK

(e) Los Angeles, USA **(f)** Munich, DE

Figure 5.2: Density Plots of the tweets performed by the collected users, near the examined city



(a) New York, USA **(b)** Orlando, USA

(c) Paris, FR **(d)** Seattle, USA

Figure 5.3: Density Plots of the tweets performed by the collected users, near the examined city

These city characteristics, and particularly the areas that attract activities, cannot be considered static; thus the frequent investigation of such changes needs to take place, to identify and cater for areas of interest that attract additional demand. Given the scarcity of traffic count locations, and their granularity, the indication of these changes from other sources is of particular interest, and as it is shown, this is possible to be achieved in a rather cheap and easy way from Social Media data. It should also be mentioned that these observations are becoming increasingly important when considering the differences in the observed travel survey and Social Media patterns (Section 5.5).

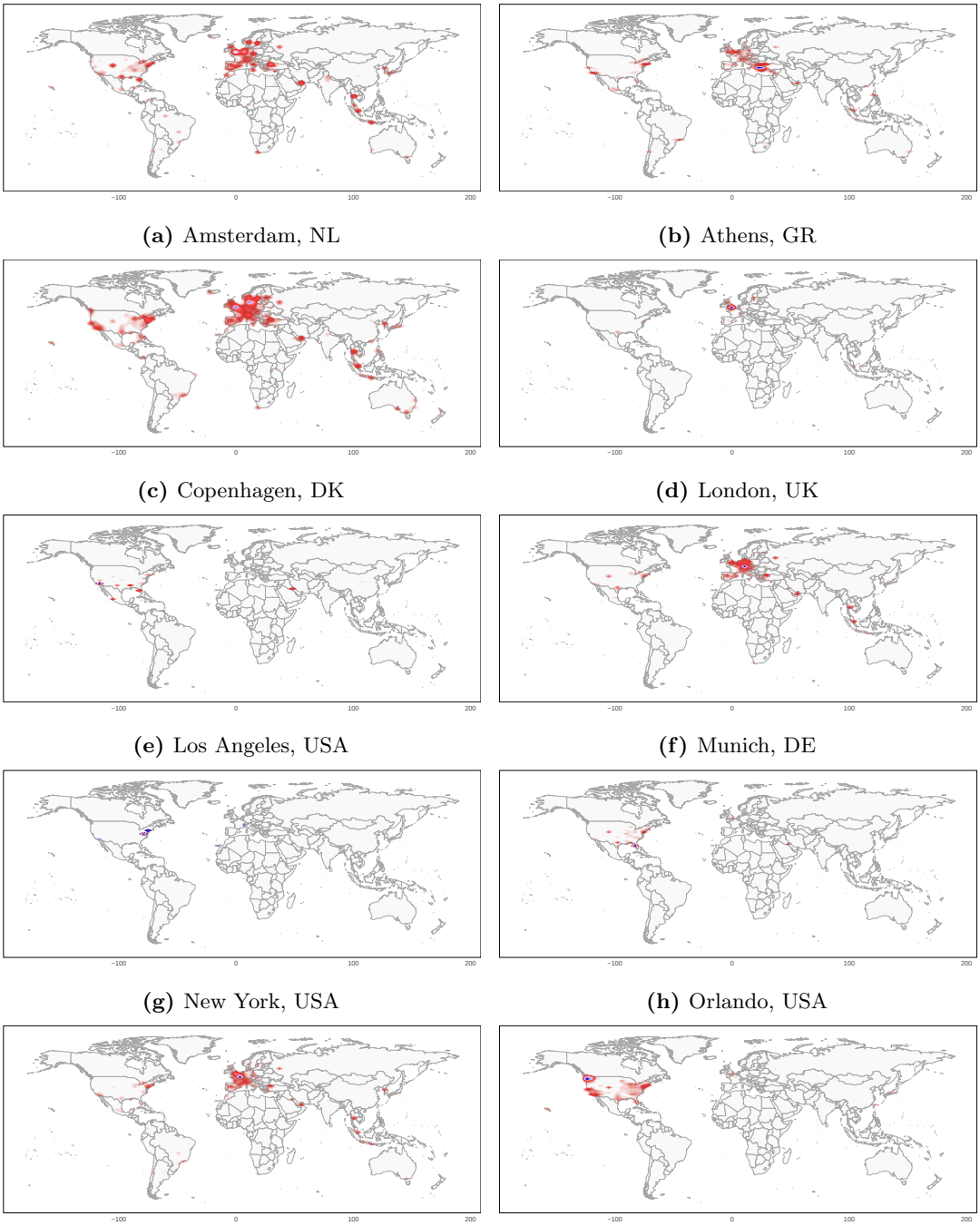
Another interesting stream of research is the global mobility patterns and the potential of extracting them from Social Media. The seminal work of Hawelka et al. [2013] concluded that Twitter can be used as a proxy of human mobility, especially at the country to country level. In our case, the User-based Data Collection (Historical Data Collection) is not spatially restricted, allowing for the collection of the places visited by individuals around the world. This allows us to first of all undestand which are the places that attract mobility from the examined cities and also where demand is produced and is captured in the cities examined. Figure 5.4 presents the density plots of areas that users in each examined city have visited. It should be noted that we do not distinguish between tourists and residents of each city.

As it is evidenced, the locations visited differ in Europe and the USA. More specifically, users from Amsterdam illustrate a concentration of posts in Europe, while also showing posts in other areas, mainly the east coast of the USA and regions in Asia. Users from Athens illustrate a higher concentration of posts in Europe; however, there is a much lower dispersion of tweets in Europe in comparison to users from the rest of the European Cities. Another compelling case is the case of London, where there is a very high concentration of posts in the United Kingdom, illustrating a much lower number of tweets posted in other areas. The most extensive spread of posts is observed by users originally collected in Copenhagen. On the other hand, in the USA there is an apparent concentration of users in the areas collected (Orlando, Los Angeles, New York), while only Seattle illustrates a rather wider spread of posts mainly in the USA.

5.3.2 Athens

Spatial analysis was performed for the identification of the areas where users post tweets from the Real–Time Data Collection Methodology. It was found that larger numbers of twitter posts were posted in areas with leisure–related land uses. Table 5.3 presents some Twitter–related characteristics for the 10 municipalities that were found to have the highest number of tweets (labelled in Figure 5.5). It was found that the highest number of geotagged tweets have been posted from municipality of Athens (approx. 30% or total), which is the centre of the Athens region; attracting travellers from all the other municipalities. The other municipalities that are identified to illustrate a high number of posts are either municipalities with high income (i.e. Glyfada, Marousi, Chalandri) or municipalities with leisure–related or transportation–related land uses (i.e. Kalithea, Vari, Voula, Vouliagmeni, Peraias). The finding on income is further confirmed but the positive linear statistical correlation (0.31) of the number of tweets posted at areas with high income.

The density of the tweets to the population is also presented in Figure 5.5 for all the municipalities. Note, that there is a rather very high density of tweets per population at the Athens airport.



(a) Amsterdam, NL **(b)** Athens, GR

(c) Copenhagen, DK **(d)** London, UK

(e) Los Angeles, USA **(f)** Munich, DE

(g) New York, USA **(h)** Orlando, USA

(i) Paris, FR **(j)** Seattle, USA

Figure 5.4: Density Plots of the tweets performed by the collected users on a world scale, as collected from each city.

Table 5.3: Tweets and Characteristics of the 10 Municipalities with highest number of tweets in Athens Region

Municipality	Tweets	Population	Tweets/ Pop.	Percentage of total Tweets	Income (Euro/ year)
Athens	85325	664046	0.13	29.5%	16959
Glyfada	12483	87305	0.14	4.3%	18100
Marousi	12064	72333	0.17	4.2%	20508
Pireas	11293	163688	0.07	3.9%	14631
Kifisia	9935	70600	0.14	3.4%	24518
Chalandri	7008	74192	0.09	2.4%	20030
Kalithea	6328	100641	0.06	2.2%	13784
Vari – Voula – Vouliagmeni	6257	48399	0.13	2.2%	19349
Nikaia – Rentis	5767	105430	0.05	2.0%	10227

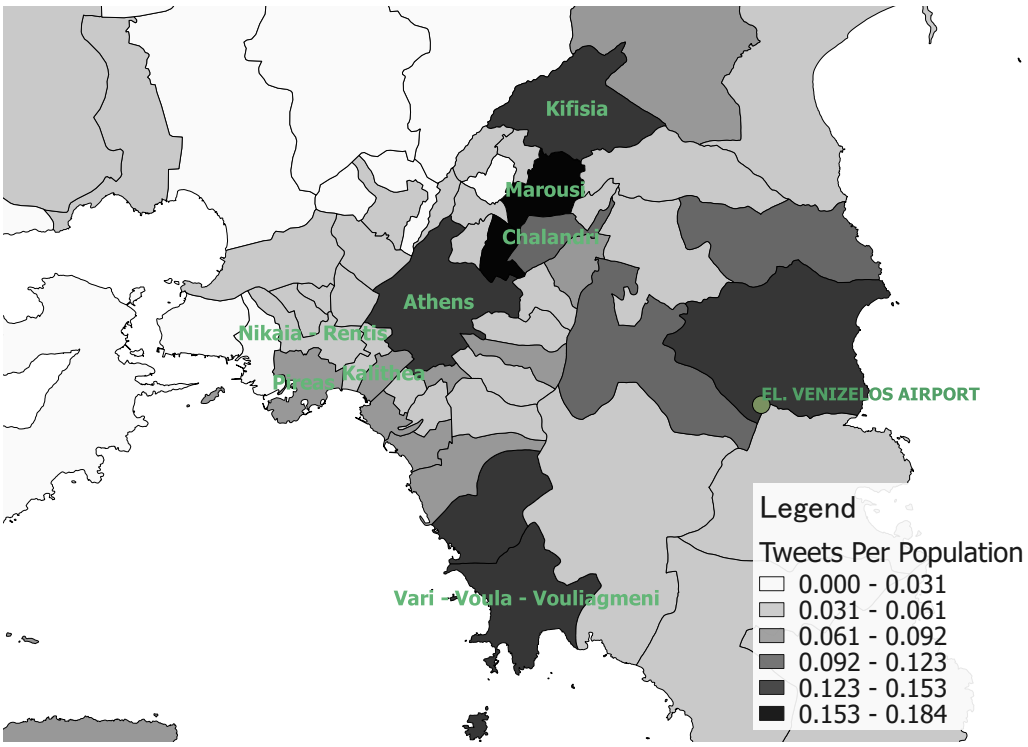


Figure 5.5: Tweets per Population in the Region of Athens (10 denser municipalities labeled)

5.4 Temporal Analysis

5.4.1 Comparative Analysis

The analysis of the temporal dimension of Twitter posts has been performed in the form of a direct comparison of the temporal distributions for both the geotagged and the not-geotagged tweets. For matters of clarity, it should be noted that the hours in the distribution were adjusted for the different time zones, taking into account the summer time difference when necessary. Figure 5.6 presents temporal distributions in different days and hours of the week.

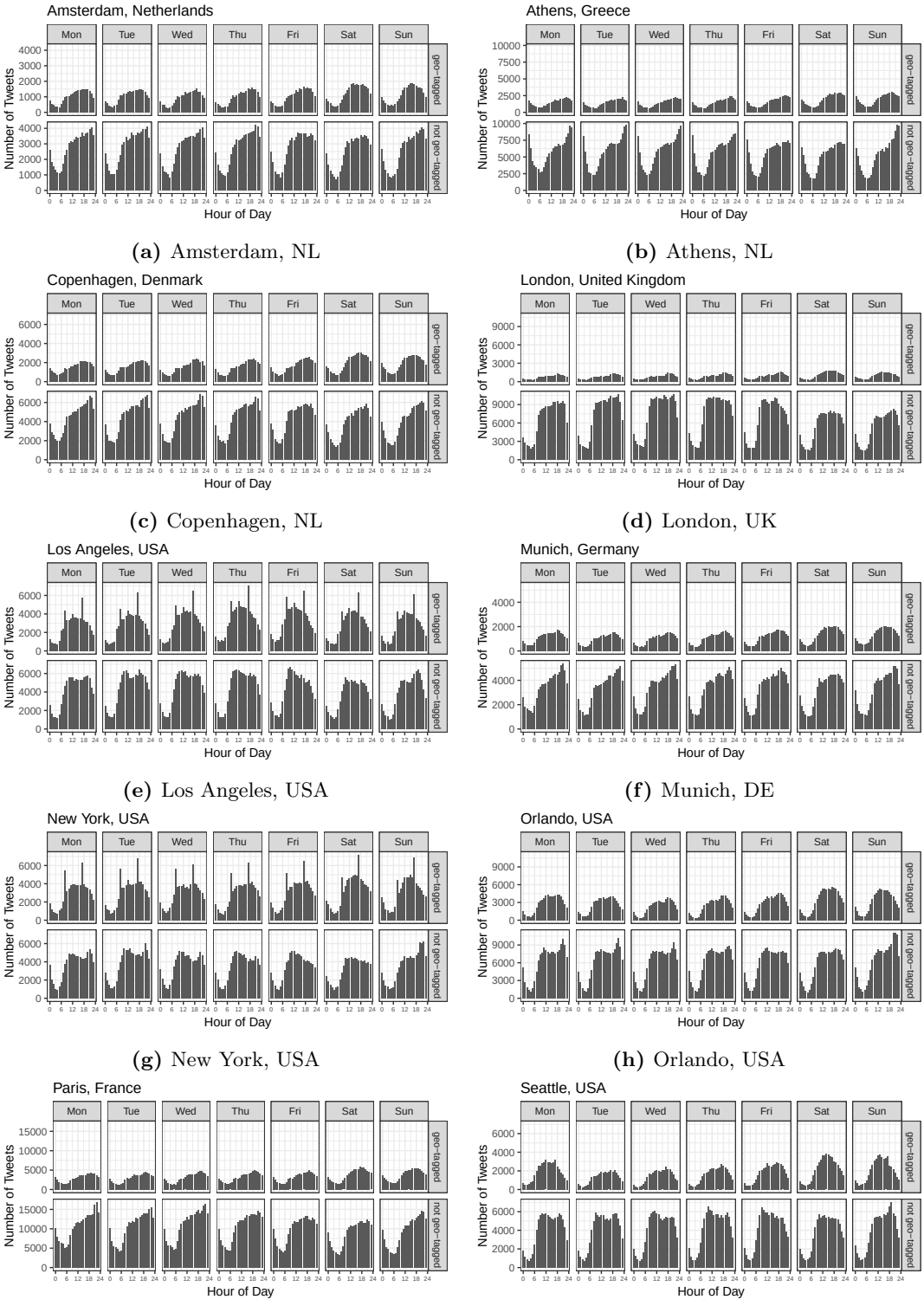
With regards to regularity in time, it is observed that it is much more pronounced than what we can observe for space. This illustrates an almost habitual use of Social Media, which makes it very interesting for the exploration of mobility patterns. Additionally, it is evidenced that there is a rather increased posting activity during weekends, especially for geotagged tweets, and also during evening hours with the peak to be usually around 17:00-20:00. The lowest points for all examined cities is during night hours. Another interesting characteristic of the data collected from New York is the peaks that are observed in most cases two hours in the day (around 8:00-9:00 and 17:00-18:00). This type of peaks is also observed in the case of Los Angeles.

5.4.2 Athens

Further analysis on the temporal dimension illustrates that weekdays follow a similar pattern of posting geotagged tweets (Figure 5.7). On the other hand during weekends (Friday, Saturday and Sunday), the number of geotagged tweets deviates from the pattern observed during weekdays. On the hourly patterns observed, there is a peak in all days during the evening/night concerning both geotagged and not geotagged tweets. The findings above, point towards the direction that there is a connection of Twitter use with non–work related activities.

5.5 Social Media and Travel Surveys

Aiming at understanding the connection between Social Media and mobility–related data, the extraction of data from three major SM platforms and the direct comparison to the resulting travel demand with available conventional travel–diary survey responses from the same region is pursued. Thus, the aim is to investigate the degree to which SM data can contribute towards the generation of data that would be directly useful for mobility studies, in general, and Intelligent Transportation Systems (ITS) applications in particular. Facebook and Foursquare are used for the collection of check-ins in venues near a location defined by users, while Twitter is used for the collection of public tweets that users post in a specified area. A case for Thessaloniki, Greece, is presented from which conventional travel survey data was available for comparison. The comparison is meant to illustrate the advantages and disadvantages of using each data source and the potential of combining those different data sources for the more adequate definition



(a) Amsterdam, NL

(b) Athens, NL

(c) Copenhagen, NL

(d) London, UK

(e) Los Angeles, USA

(f) Munich, DE

(g) New York, USA

(h) Orlando, USA

(i) Paris, FR

(j) Seattle, USA

Figure 5.6: Examples of Temporal Distributions for geotagged and not-geotagged tweets

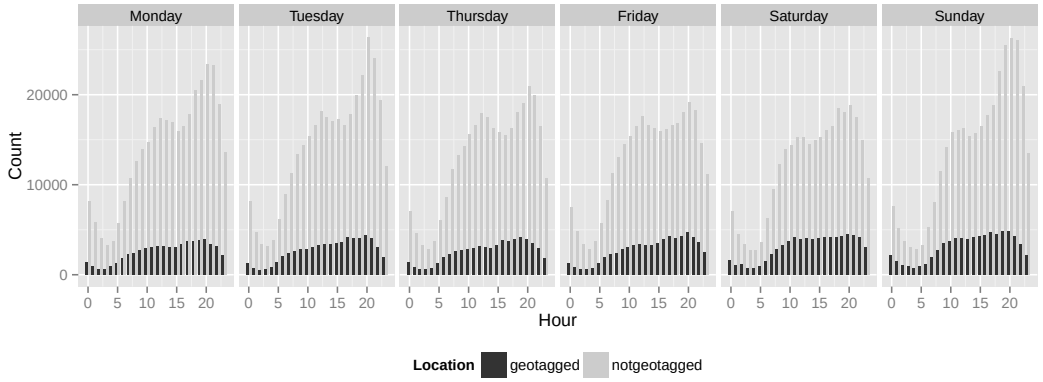


Figure 5.7: Daily and Hourly Distribution of Tweets collected for Athens case

of travel demand. In the following subsections, an analysis of the destination types defined in Facebook, Foursquare and conventional travel survey data is then presented, followed by the presentation of temporal and spatial characteristics of all data sources.

5.5.1 Destination Types

The venue categories from Facebook and Foursquare were manually aggregated, after being corrected in case of wrong category entry. Venues with wrong categories were usually those that were defined as Local Business, while they could be cafeterias or bars. The categorization took place in a way that would allow for a comparison to the destination type recorded in the conventional travel survey. The first identified difference lies on the fact that SM originated data was more disaggregated and dense in leisure–defined destination types (such as bars, cafeterias, parks, museums); thus activities performed. In the conventional travel survey those destination types were by default aggregated in one class (leisure). On the other hand, work–related destination types and activities could not be inferred by SM–originated data, as users do not indicate the activity performed at the venue visited. For the direct check–in definition, Twitter is not taken into account due to the requirement of indirect extraction of activities and destination types. For the data sources examined the percentile distribution of check–ins (for Social Media) and trip–end (for the conventional travel survey) the selection of destination type is presented in Figure 5.8. As illustrated, there is a similar percentile distribution for SM–originated data while there is a clear difference in the representation of most destination types among the conventional travel survey and SM–originated data.

5.5.2 Temporal Distribution

The clear difference of the activity representation by each data source is also evidenced by the temporal distributions of the examined data sources. In Figure 5.9 the average percentile distribution of within day variations for each data source is presented. As

Figure 5.8: Distribution of a selection of destination types for Facebook, Foursquare and conventional travel survey

illustrated, the peak of the percentile distribution for conventional data occurs during the morning hours, something that does not coincides with SM originated data, which illustrate peaks during evening hours. This can be connected to the fact that the work–related activities are not represented.

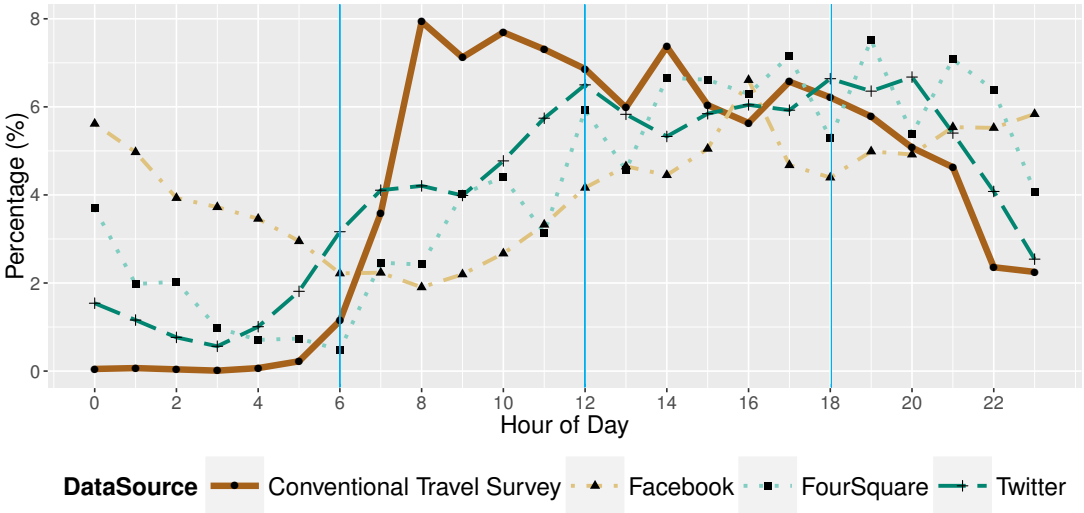


Figure 5.9: Within day temporal percentile distribution of recorded activities across examined data sources

For the identification of possible relationships among the data, Pearson correlation (Figure 5.10) was estimated for the 4 examined periods (i.e. [0–5], [6–11], [12–15], [16–23] Figure 5.9). In most periods there is a clear positive correlation among Social Media data sources that illustrate that the actual use of Social Media is mostly similar among the platforms examined. The most prominent correlation is found for the case of Facebook and Foursquare (*FB-4SQ*), while Twitter in some cases illustrate a negative (although not very strong) correlation mainly with Facebook. Strong positive correlation is not evidenced in the case of the Conventional Travel Survey (*TrS*) with the exception of mainly the morning period ([6-11]) where all data sources illustrate increasing activity.

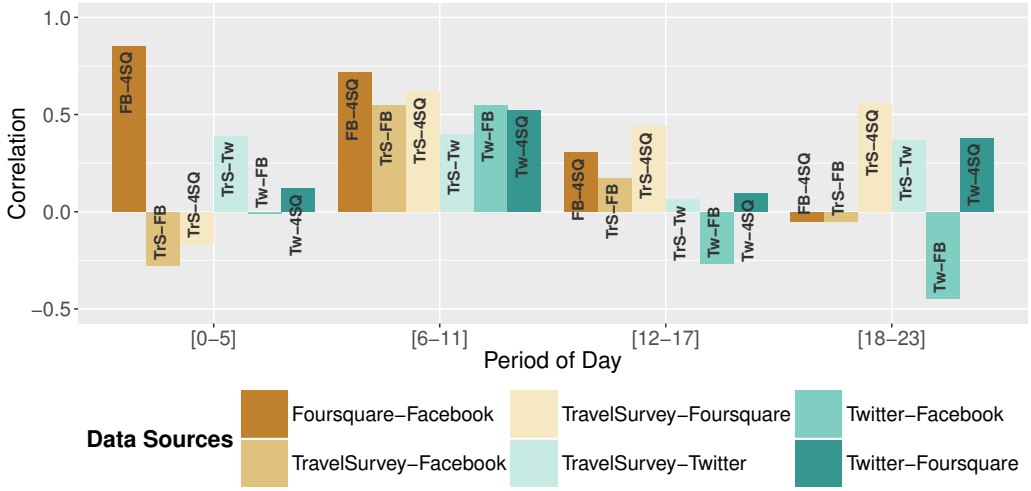


Figure 5.10: Pearson Correlation among different data sources

5.5.3 Spatial Analysis

On a spatial level, the distribution of locations collected and characterised as areas with high Social Media activity (Figure 5.11 and 5.12) are also known to be characterised as high recreational land-uses areas in the city of Thessaloniki. Those areas are the coastal line of the Thessaloniki city where many restaurants and cafés are located, and also the main shopping area of the city (north-east to the coast). This is evidenced in 3 Social Media platforms data, which in the case of Facebook and Foursquare is expected given the venue oriented check–in. Note, that venues can in some cases be altered to include someone's house, however not very common in the data collected (also evidenced in Figure 5.8).

On the other hand, the conventional data collected illustrates a more evenly distributed concentration of attractions with little fluctuations in the city centre. The only location attracting a higher number of trips (dark blue spot) is found to be on

one of the major squares (Aristotelous Square), which is the centre of the down-town business district, a common meeting point and a public transport transfer location.

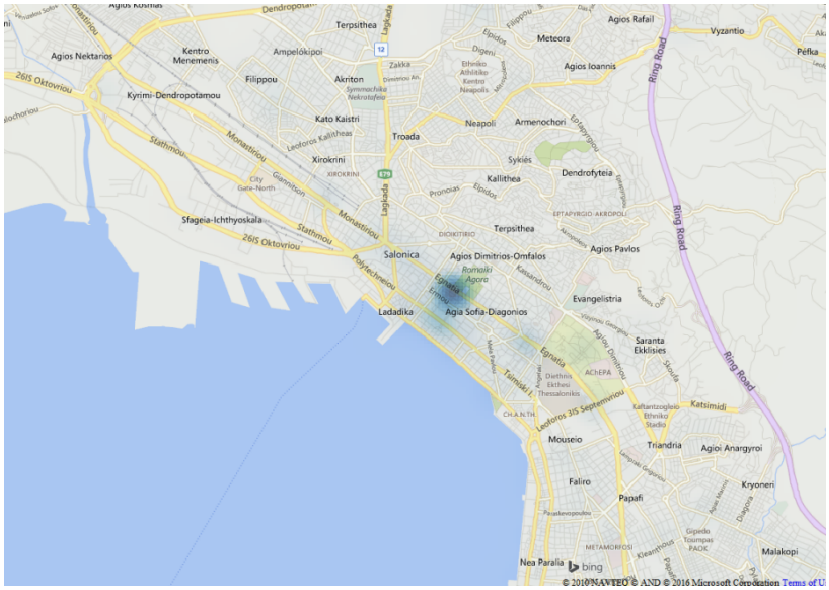


Figure 5.11: Conventional Travel Survey Attractions

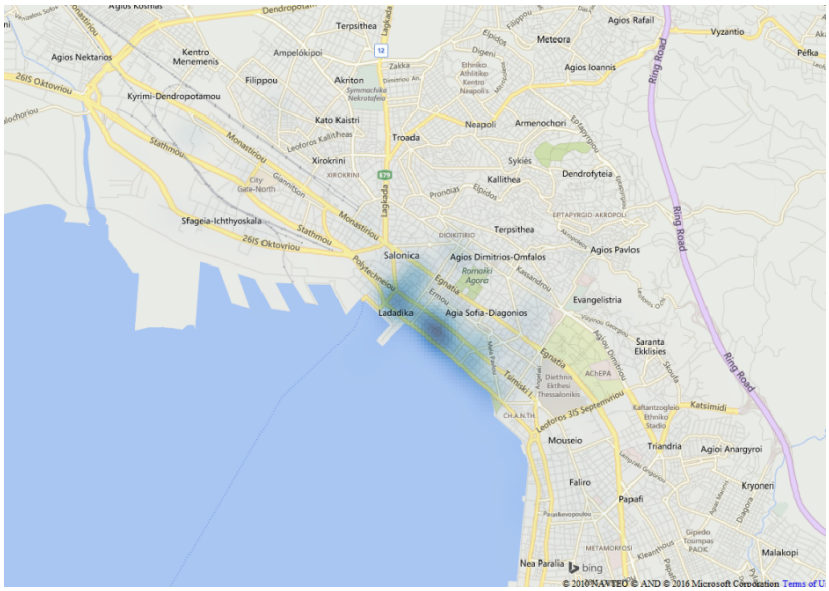


Figure 5.12: Facebook Check–ins

A comparison of the percentile distribution of Transportation Analysis Zones (TAZ) was also performed in order to identify areas with high activity, given a common spatial indicator. This analysis was performed only for the case of data originating from

conventional travel survey and Twitter, because for Facebook and Foursquare only a fraction of the total area is covered, thus not allowing for an indicative representation. As it is illustrated in Figure 5.13, there are many cases where there is a strong difference of trips starting from, or ending to each TAZ. More specifically, for TAZ that are denoted to be within the city centre, a higher percentage of Tweets posted are recorded, while conventional data projection to TAZ follows a smoother distribution (although some peaks are identified in a few zones). It should be noted that all zones are generally well represented in terms of percentile distribution of Twitter and the conventional travel survey population.

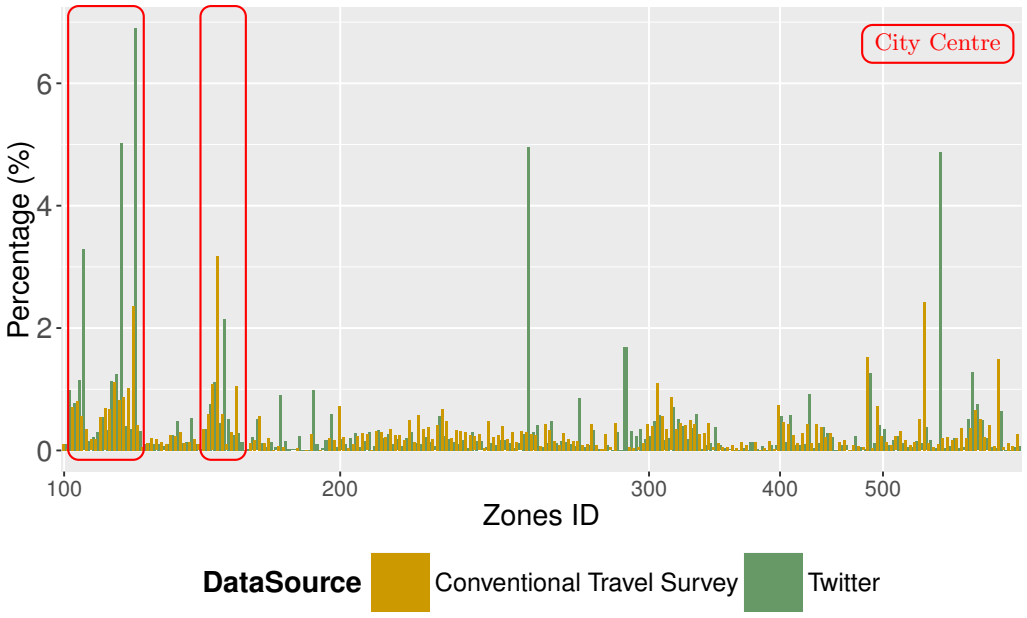


Figure 5.13: Comparison of Twitter and Conventional Data locations

6 Extracting Transport Features from Social Media Data

Contents

Components of this chapter are presented in:

Emmanouil Chaniotakis and Constantinos Antoniou. Use of geotagged social media in urban settings: Empirical evidence on its potential from twitter. In *IEEE 18th International Conference on Intelligent Transportation Systems, (ITSC)*, pages 214–219. IEEE, 2015

E. Chaniotakis, C. Antoniou, and F. Pereira. Mapping social media for transportation studies. *IEEE Intelligent Systems*, 31(6):64–70, Nov 2016b. ISSN 1541-1672. doi: 10.1109/MIS.2016.98

Emmanouil Chaniotakis, Constantinos Antoniou, and Loukas Dimitriou. *Digital Social Networks and Travel Behaviour in Urban Environments*, chapter Social Media and Travel Behaviour., pages 0–0. CRC Press, Cham, 2010a. ISBN 9781138594630. URL `https://www.crcpress.com/Digital-Social-Networks-and-Travel-Behaviour-in-Urban-Environments/Plaut-Pinsly/p/book/9781138594630`

6.1 Introduction

Based on the findings from the descriptive analysis performed and the comparison of the Social Media data with transport–related data, it has become obvious that the exploration of the potential to extract activities from Social Media would fill a gap in the current transport data. Thus the exploration of feature extraction for activities is pursued in this chapter. For the feature extraction applications presented, different datasets from different cities have been used as examples. This is primarily targeted towards the evaluation of Social Media in different conditions.

The main feature extracted is the activities themselves based on the combination of the data enrichment and text–based feature extraction methodologies presented in Sections 4.3 and 4.4.2. With a combination of data from different Social Media platforms labels are extracted for activities performed and then text modeling is deployed to extract activities from the related text (Section 6.2. Another interesting feature that indicates the areas of interest for different user groups is referred to as the Areas of Interest 6.3. The case of Athens is examined for the definition of the different areas from which individuals post on Social Media, for two user classes: residents and tourists. The process followed is described in 4.4.1. Finally, the exploration of activity space on the basis of Social Media takes place based on the method presented in Section 4.4.3. This application allows for the evaluation of the use of clustering techniques for the estimation of activity spaces and also to compare the difference different residents and tourists in the cities examined.

6.2 Activities Extraction

The Social Media data examined originates from Twitter and was collected using the Twitter Application Programming Interface (API) service (Section 2.1). It should be noted that the application of the second phase of the data collection methodology from Twitter is not bounded in a specific area as it collects all the tweets that individuals post. Based on the Twitter dataset, extraction of information from Foursquare posts integrated in Twitter is performed to achieve the enrichment of activities–related information (Figure 6.1. In order to parse the information the extraction of URL patterns using regular expressions takes place including the extension of shortened links to long links and Foursquare queries. The queries' response is of JavaScript Object Notation (JSON) format and it can be either a venue, or a check–in. In some cases it requires for the link to be (within the Foursquare query) resolved. Finally, the response is parsed in tabular form and stored with the initial tweet information. The required queries are subject also to Foursquare APIs impose. At the time this methodology was applied (June 2016), the maximum number of queries was 500 queries per hour. The application of the methodology for enriching twitter data was performed in R [R Core Team, 2017] using related packages (ThinkToStartR, stringr, rjson, RCurl, httr). The only exception was the URL extension component that was applied in Java.

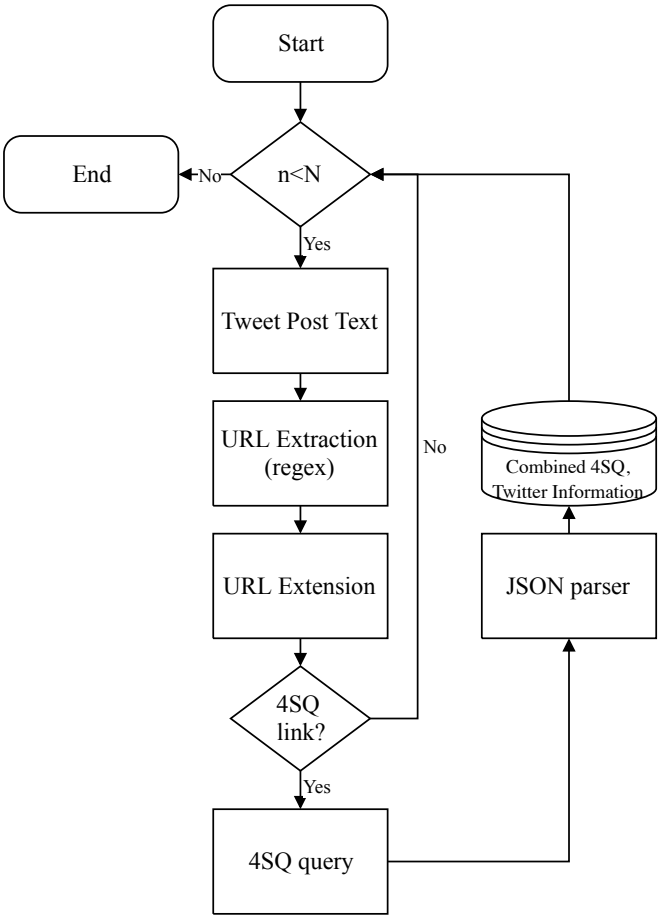


Figure 6.1: Parsing process for Twitter data enrichment from Foursquare

The resulting dataset from the data collection and dataset definition methodology is used for the extraction of information on the nature of activities users perform and post about in Social Media. The methodology is based on the notion that there are places that users visit frequently to perform activities (e.g. home, work) namely recurring activities, and other places which are scarcely visited or only visited once, for non–recurring activities (e.g. leisure). This distinction of places visited and –as a consequence– activities performed is considered of high importance, for the data generalizations that are required aiming at both data reduction – in terms of the identification of locations that are commonly visited – and enrichment – in terms of generalization of activities identified. The methodology is presented in Figure 6.2. First, locations are characterized based on their spatial distance and their density, in order to specify locations that are visited by each user more than a specified number of times. The classification algorithm used is the Density–based Spatial Clustering of Applications with Noise (DBSCAN) [Ester et al., 1996], which characterizes locations visited less than a specified number of times and under a specified density as noise. This is very

convenient, as it allows for the distinction of locations that are frequently visited and can be all together characterized by only one Foursquare post, in a robust and fast way. The classified locations receive a characterization based on the venue categories they belong to. The characterized – based on recurrence – tweets combined with venue and temporal characteristic are then used for the definition of document–term matrices specified to each type of activities that can be performed in venues. This allows for direct classification of each tweet describing an activity, in an activity cluster using classification algorithms. It should be noted that not all tweets are used to describe activities.

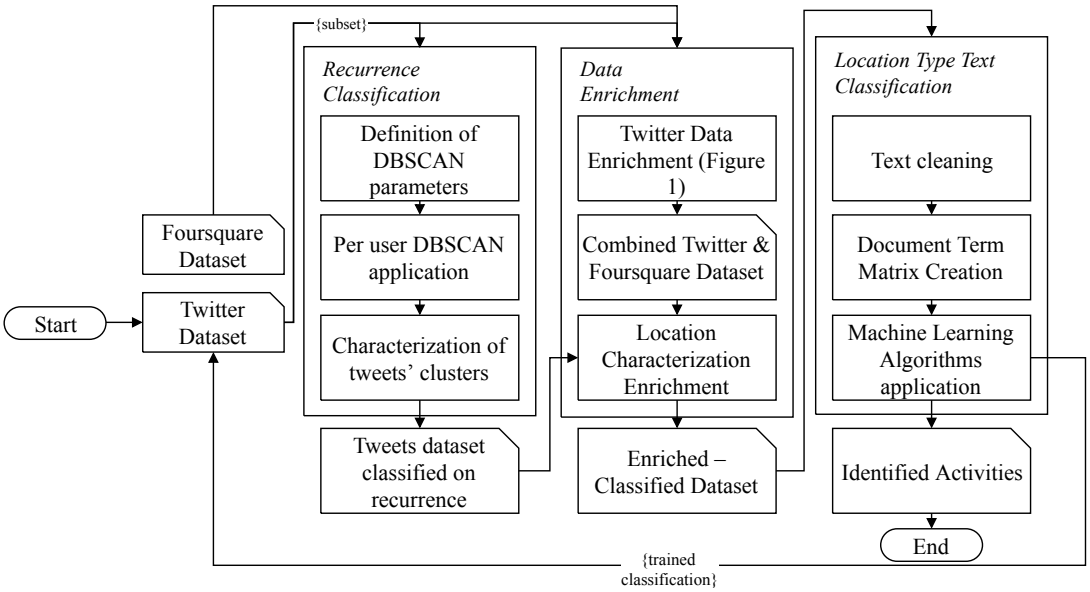


Figure 6.2: User–Centric Activity Enrichment methodology

6.2.1 Dataset Construction

The first phase twitter data collection method was used for continues data collection in a period of one year for the greater London area. The resulted dataset consists of 482,883 unique users and approx. 4.5 million tweets. For a random sample of those users (90,000 users) the last 200 tweets were collected using the historical data collection of the Twitter data collection methodology. In total, the database included 11,060,814 tweets. From those, 8,141,996 were tweeted in the greater London area. From the geotagged tweets in the greater London area, 3,764,230 (46.2%) included a link that could be parsed, 220,118 of which originated from Foursquare (2.7%). The spatial and temporal characteristics of the sampled tweets are presented in Figure 6.3 and 6.4 respectively.

Concerning the spatial distribution, it is clearly evidenced that urban structures can be identified, as tweets seem to also be concentrated in the urban areas around the

Figure 6.3: Spatial Coverage of extracted dataset in London

city of London. On the temporal level, the distributions presented indicate a peak among evening hours, while the highest number of unique tweets takes place during the weekends and specifically Saturday afternoon. This trend is found to be similar in other cities examined (Section 5.4), indicating a tendency towards posting during non–working hours.

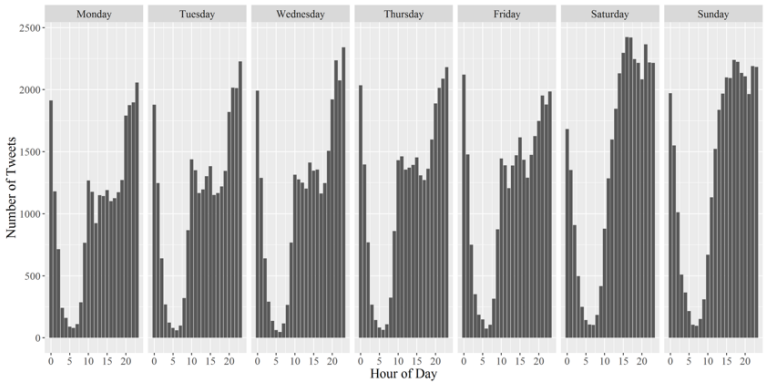


Figure 6.4: Temporal Distribution of extracted dataset in London

As clarified in the literature, data collected from Twitter contains a large number of users who have either only a few tweets or post automated messages from the same location [Gal-Tzur et al., 2014a]. For that reason, a subset of users characterized based on their above average posting activity and a threshold of average twitter trip length (average direct distance between tweets of a user's determined as the average of the distance matrix) was selected, for the application of the User–Centric Activity Enrichment methodology. The subseting data flow and the subset procedure is presented in the Figure 6.5. The average number of tweets was found to be 22 and the average distance threshold was defined to be 1 km. The subset included in total 50,344 users who have posted in total 5,080,362 geotagged tweets.

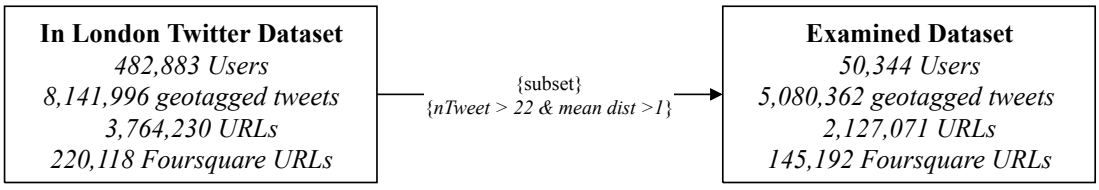


Figure 6.5: Subset characteristics, in the greater London area and concerning users with above average Social Media use

6.2.2 User Characteristics and Recurrence Classification

The selected with above average posting activity users were examined upon their posting and classification for location recurrence characteristics. On the posting characteristics, the average number of posts each user was posting was found to be 101.8 with a standard deviation of 311.4, indicating that there is a number of users that are posting a very large number of tweets. The standard deviation of the percentage of tweets posted per day was found to be 0.8% suggesting a close to uniform distribution and indicating that on average the users selected are frequent Twitter users. From the initially selected 50,344 a fraction of 4,185 users (8.3%) were found to have posted at least one tweet that include a foursquare link. The average percentage of posts including activities for these users was found to be 39.0%.

On the classification for location recurrence, the parameters were selected after examination of various settings taking into account the GPS accuracy [Schaefer and Woodyer, 2015] and the number of tweets that each individual posts. For this reason, the neighborhood of a point parameter (Eps) was defined to be 0.002 and the minimum number of points to be 5. The average number of clusters was found to be 2.41 with a standard deviation of 3.18. The percentile distributions of the locations that belong in a cluster and those characterized as noise per user are presented in Figure 6.6.

It is indicated that there is a significant amount of users, who tend to only post in locations that were not be included in any cluster. This agrees with the fact that some people only use social media in order to post for activities performed scarcely, or places visited only a few times. This finding further supports literature on the use of social

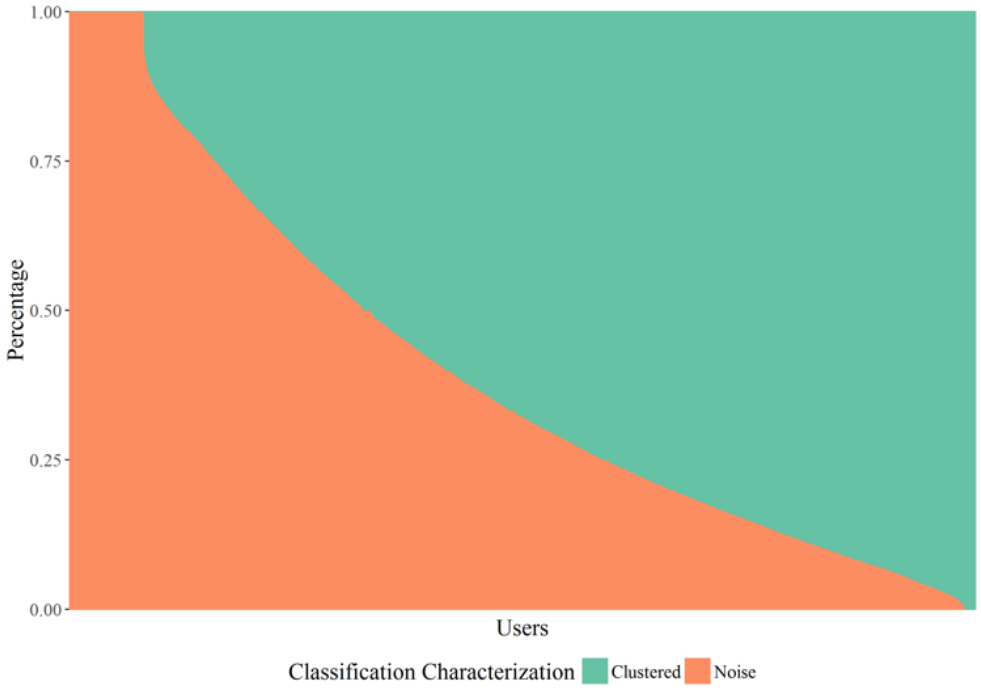


Figure 6.6: Percentile distributions of the locations that belong in a cluster and those characterized as noise per user

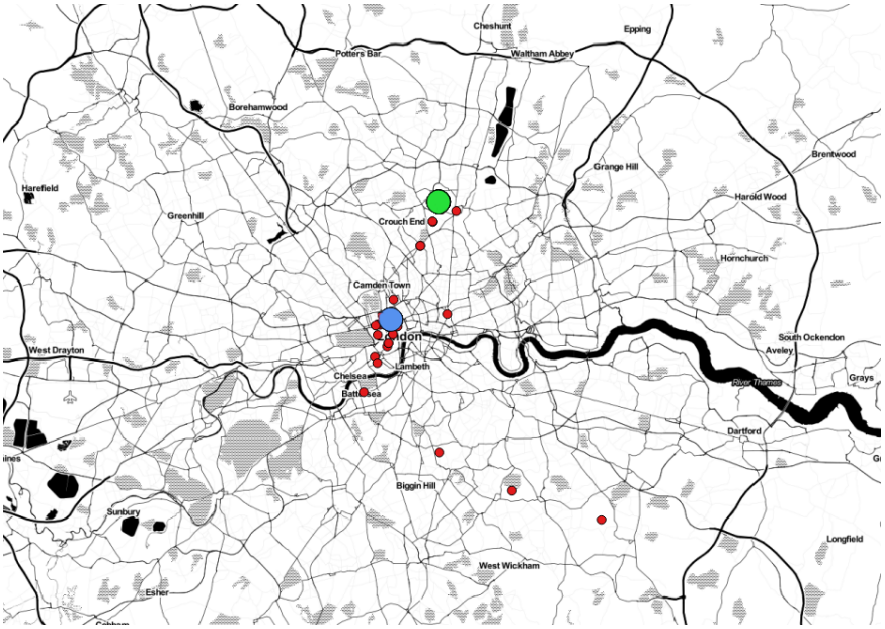
media, and can be used in profiling of individuals who use Social Media based on the type of things that they tend to post.

On the following figure (Figure 6.7), two examples of the classification results are presented: one for which the locations visited cannot be clustered and a second one for which the algorithm identifies two classes.

6.2.3 User–Centric Activity Extraction

The categories provided by Foursquare on the locations visited by users were manually aggregated in 14 representative categories and examined upon their temporal characteristics for verification. As it is illustrated in Figure there was a tendency towards leisure activities that was present for all days of a week. Activities such as education and work were clearly higher represented during week days and less represented during weekend days. The activities with the highest representation were found to be the "Bar – Pub" and "Restaurant" activities.

The application of the data enrichment methodology increases the information concerning the activities performed by individuals. First the identified as clustered destinations were enriched based on location characterization (activity). More specifically, for the dataset of the above average users defined; from the initial number of Foursquare posts, the number of posts that were assigned in recurrence clusters was found to be



(a) User with no cluster identified

(b) User with two clusters identified

Figure 6.7: Examples of classification using Density Based Spatial Clustering of Applications with Noise

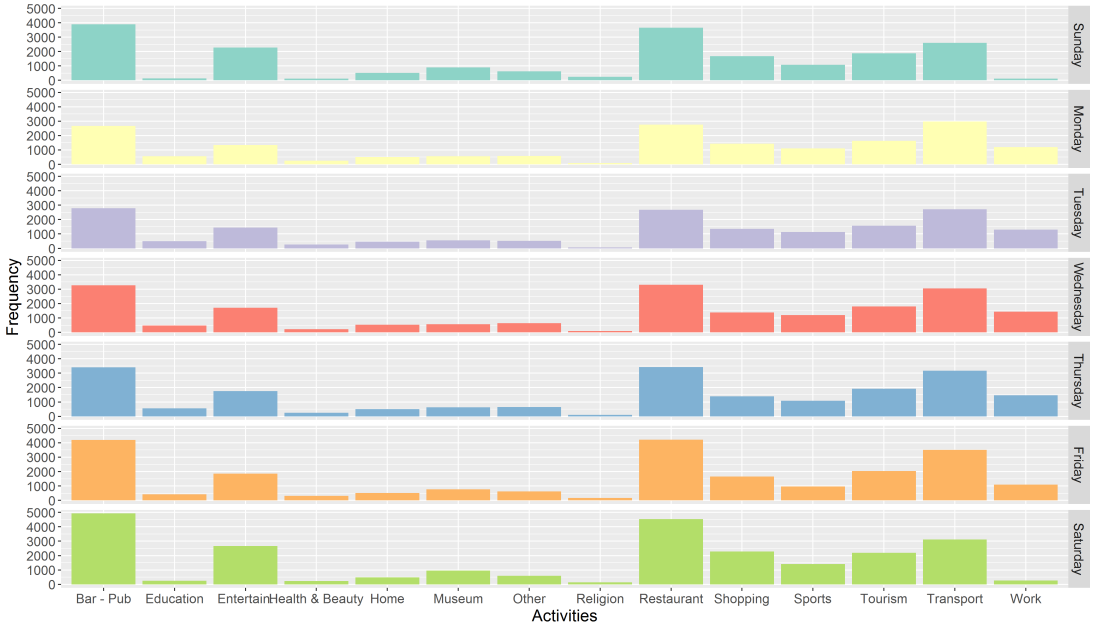


Figure 6.8: Daily distribution of Foursquare labeled activities

172,675. This is based on the initial 65,806 tweets that included known location characterization and were associated with a known cluster. It should be noted that this data enrichment step further enriched the classification process presented below, as it allows for posts that were not meant to denote visitation at a specific location to be associated with a cluster location characterization and as a consequence an activity.

The last step of the data enrichment methodology is the training of classification model that can represent the activity space based on the Twitter text that individuals post. This approach would greatly improve the data availability on which researchers build their analysis. A set of classification methods believed to better fit the methodological framework proposed were selected and tested on cleaned tweet texts (removed stop words, punctuation): a) Support Vector Machine (SVM), b) Generalized Linear Model via Penalized Maximum Likelihood (GLMPML) and c) Maximum Entropy (MAXENT).

The test of the best algorithm took place 10 times on a subset of data that included the major activities identified (Bar–Pub, Entertainment, Restaurant, Sports, Shopping, Other, Work, Tourism, Transport) and for 20,000 randomly selected cases. Each time, the dataset was split randomly in a training (85% – 17,000 entries) and testing set (15%, 3000 entries). The overall accuracy results from the classification are presented in the Table 6.1 and Figure 8. Table 1 includes a summary of the results from the examined classification methods on the metrics of precision, recall and the F–score. Results in Table 1 suggest that both the Maximum Entropy and the Generalized Linear Model via Penalized Maximum Likelihood are considered capable to be used for the identification of activities from Tweets. Furthermore the precision of both algorithm indicate that

the results would provide a sufficient accuracy, especially taking into account the short number of characters allowed in each tweet.

Table 6.1: Classification Performance

	Average Precision	Precision St. Dev.	Average Recall	Recall St. Dev.	Average F–Score	F–Score St. Dev.
SVM	0.6374	0.0104	0.5808	0.0132	0.6006	0.0117
MAXENT	0.806	0.0112	0.7773	0.0081	0.7881	0.008
GLM	0.8392	0.0078	0.6752	0.0068	0.7249	0.0064

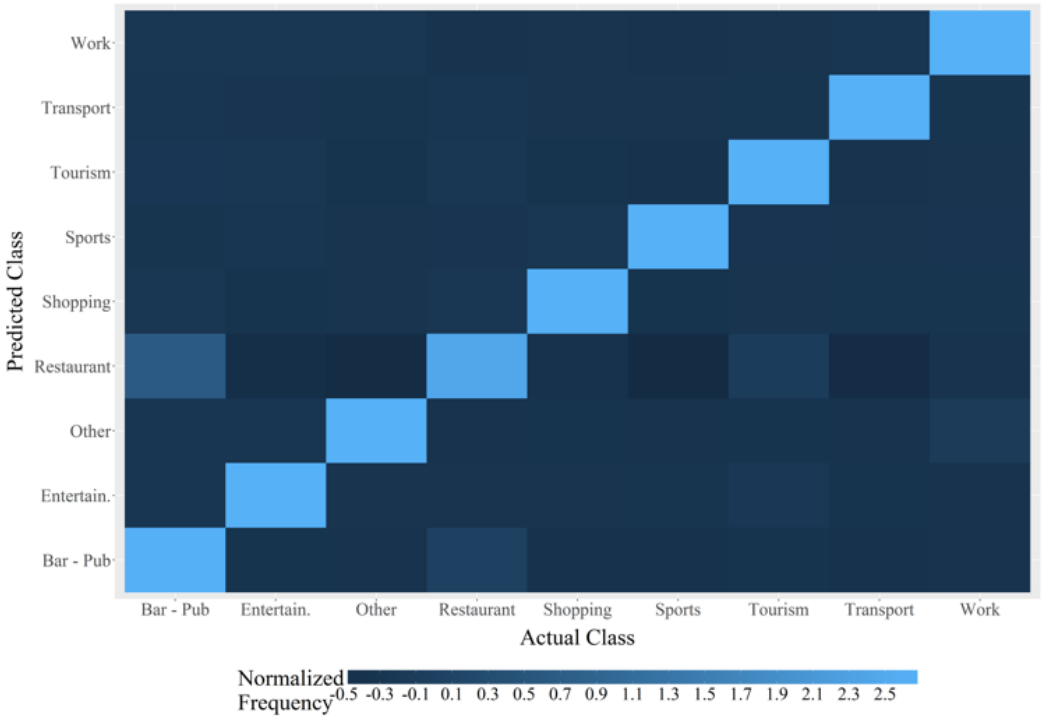


Figure 6.9: Confusion Matrix for the Performance of Max Entropy Classification

Figure 6.10 and 6.9 on the other hand provides an overview of the classification in terms of confusion for the GLM and MAXEXT case. As it is clearly indicated for both classification methods, the pair that is mostly confused is the Bar–Pub and restaurant pair, which is logical when taking into account that both location categories might involve same activities performed by users (for example drinking wine, as commonly met).

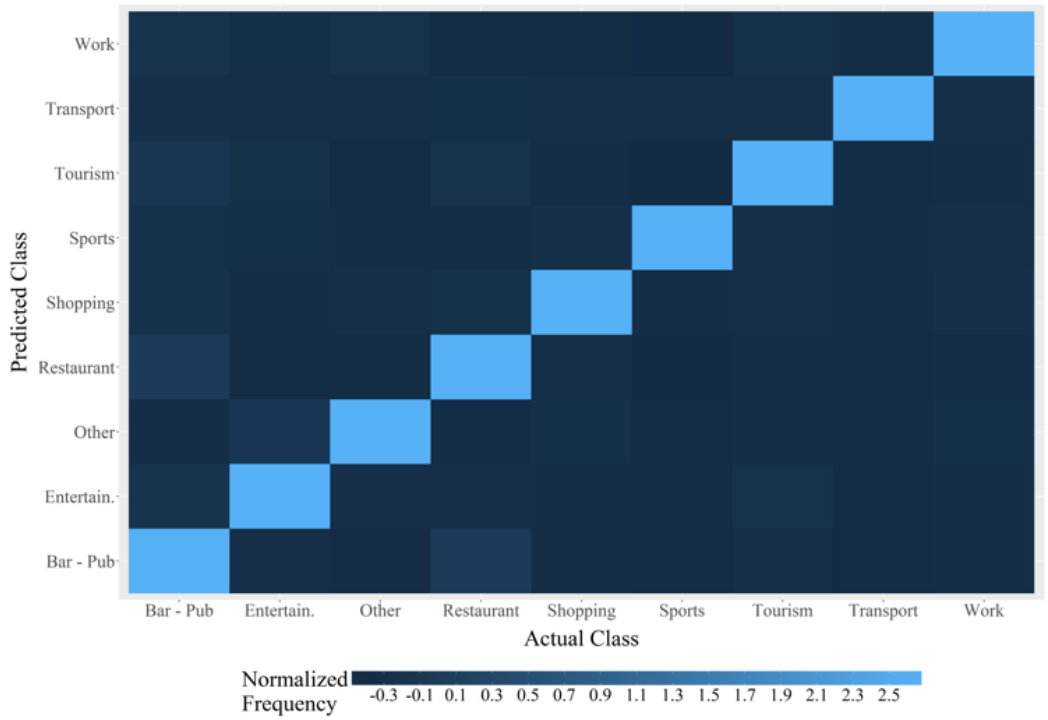


Figure 6.10: Confusion Matrix for the Performance of GLMNet Classification

6.3 Areas of Interest

For the exploration of Areas of Interest, the sample specification took place upon the distinction between inhabitants and tourists as well as the choice of users with high Social Media activity. This characterization is not always straightforward as information on the place of residence is not provided by Twitter. For that reason and in order to avoid misconceptions due to users' posting only a very small number of Tweets; users with above average posting activity (total number of geotagged Tweets posted larger than 97 Tweets) were selected. For the subset of users selected (that included 3917 users) the frequencies of the number of Tweets in the Metropolitan Athens' area to the total number of Tweets per user were estimated (Figure 6.11). As it is illustrated in Figure 6.11, there is a number of users of whom Tweets are mainly posted outside the boundaries of the Metropolitan Geographical Area, that can be characterised as tourists (percentage of in examined area Tweets <0.25), a set of users of whom Tweets are mainly posted inside the Metropolitan Geographical Area that can characterised as residents (percentage of in examined area Tweets <0.75) and a set of users of whom the residence location is unclear (percentage of in examined area Tweets >0.25 and <0.75). This classification in 3 user classes lead to 3 subsets –namely *tourists*, *residents*, *unclear*– that contained 2232 users, 1169 users and 516 users respectively. Note that in this study we only focus on the following two user classes: *residents* and *tourists*.

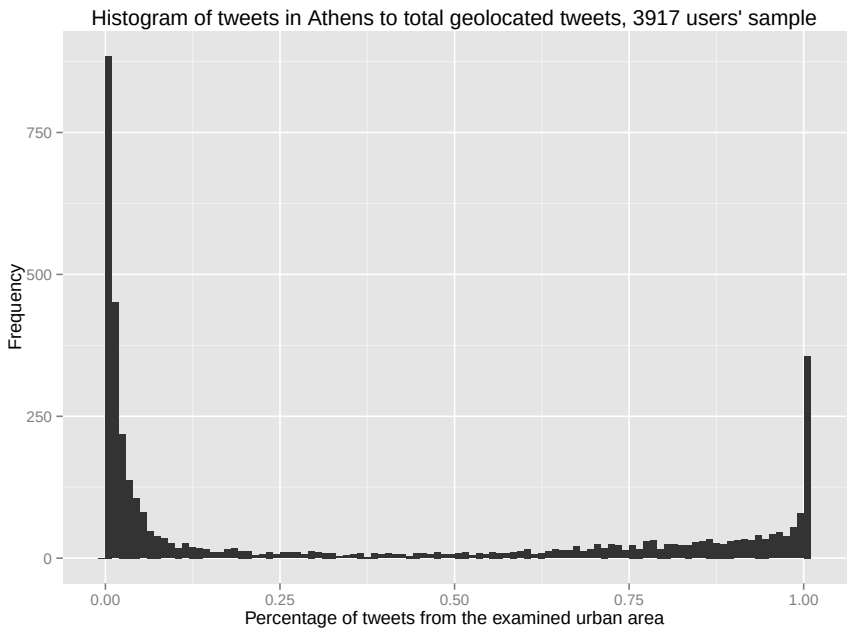
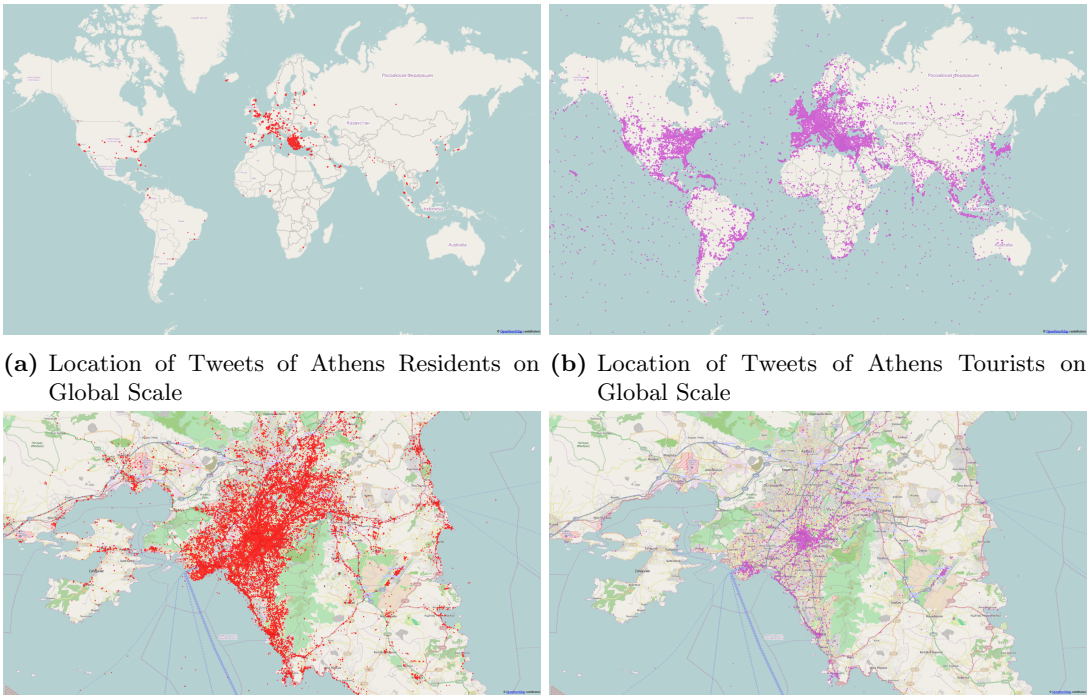


Figure 6.11: Frequencies of the number of Tweets in the Metropolitan Athens' area to the total number of geo–tagged Tweets per user (bins width = 0.01) for the above average posting activity sample

The *residents* and *tourists* classes derived were examined upon the spatial distribution of their posting activity. As the Historical Data Collection was applied geotagged Tweets were collected from places all over the world. Figure 6.12a and 6.12b illustrate the different spatial distribution of posting activity on a world scale. Users categorised as residents are found to have a higher density of Tweets in the area of Greece.

On the other hand, users categorised as tourists for the area of Athens are found to post Tweets from various places around the world with higher density of Tweets in Europe and the USA. This difference is also evidenced when examining the spatial patterns of posting activity, in the metropolitan area of Athens. Tweets from users characterised as residents are distributed in the city area, while Tweets from users characterised as tourists are mainly gathered at places that are common tourist attractions, or ports and airports.

With regurds to the actual evaluation of the Areas of Interest, Figure 6.13 illustrates the implementation of the head/tail methodology for the *tourists* class. It should be noted that especially for the *residents* class the inference of Areas of Interest should take into account the fact that those users might post repeatedly from their homes or places of work, however not included here. The finally selected sample included lines with lengths under 12.83 and 41.09 meters for *residents* and *tourists* user classes respectively.



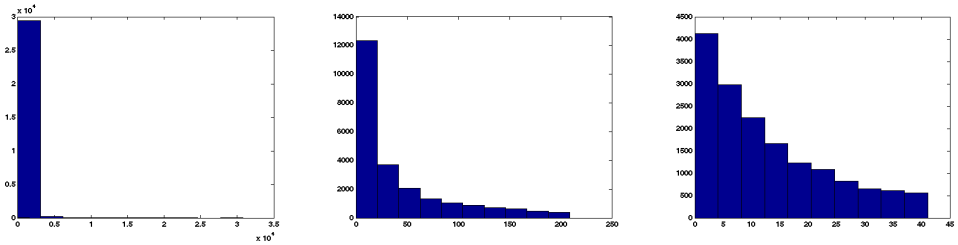
(a) Location of Tweets of Athens Residents on Global Scale

(b) Location of Tweets of Athens Tourists on Global Scale

(c) Location of Tweets of Athens Residents in Athens Metropolitan Area

(d) Location of Tweets of Athens Tourists in Athens Metropolitan Area

Figure 6.12: Location of Tweets collected for *residents* and *tourists* user classes



(a) TIN Length

(b) TIN Length under 208.3 meters

(c) TIN Length under 41.09 meters

Figure 6.13: Selected histograms illustrated the head/tail distribution

Those were compared with Points of Interest (POIs) for the city of Athens from Openstreetmap (OSM) based on its *tagging* system. Figure 6.14 and 6.15 illustrate the inference of Natural Areas of Interest derived for residents and Points of Interest from OSM. It is found that Natural Areas of Interest include a vast majority of POIs from OSM that are characterized as bars, cafe, cinema, fast–food, music–venues, night clubs and restaurants. Furthermore, there are areas, that are not described as POIs, which are located at areas defined as recreational land uses areas. Those AoI can be safely characterized as areas that are visited by users but have not yet been added in OSM.

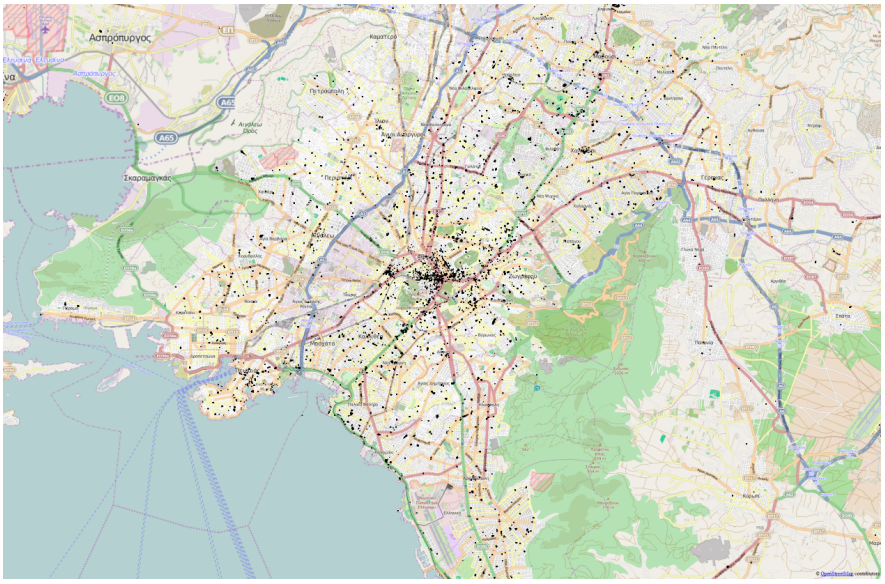


Figure 6.14: Natural Areas of Interest in the city of Athens for residents of Athens

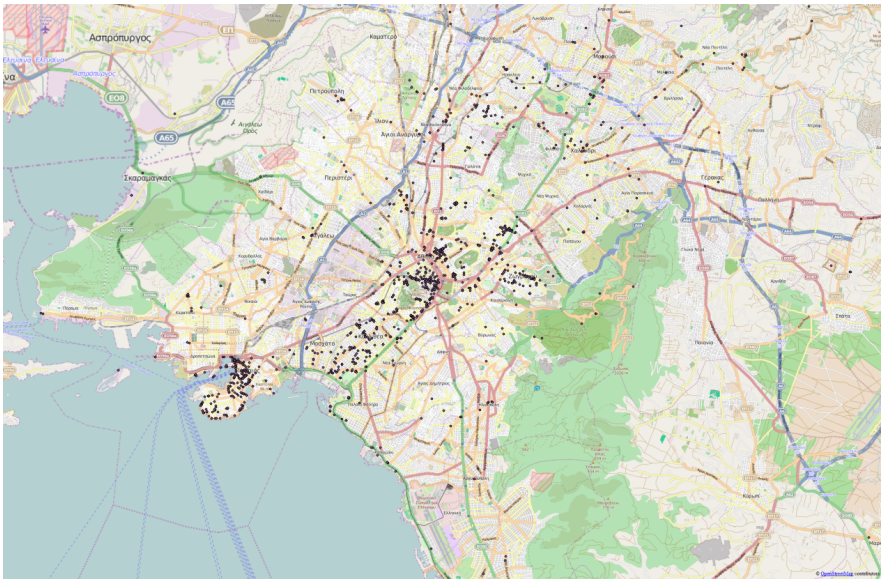


Figure 6.15: Leisure POIs from OSM for residents in the city of Athens

Concerning tourist the inference of activities resulted on larger Areas of Interest, due to the smaller number of tweets in the city of Athens (approx. 15400). In this case the activities were not examined upon leisure amenities but tourists attractions again defined as POIs from OSM. The results of the inference are presented in Figure 6.16. As it is clearly presented a large majority of tourists attractions (approx. 71%) are within the boundaries of the Areas of Interest. This finding further supports the fact

that Social Media and specifically Twitter is used to share leisure activities in the areas visited.

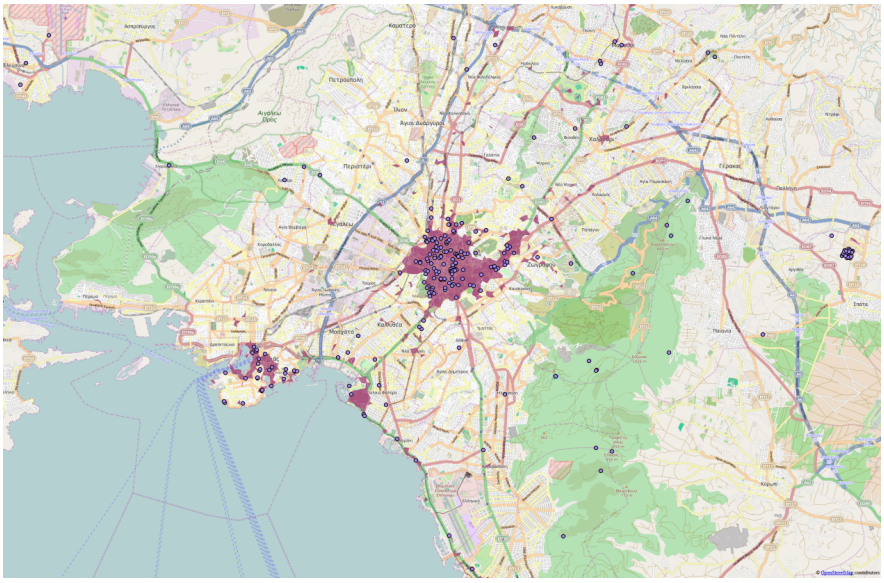


Figure 6.16: Natural Areas of Interest and Points of Interest from Social Media Data and OSM for tourists of Athens

6.4 Activity Space

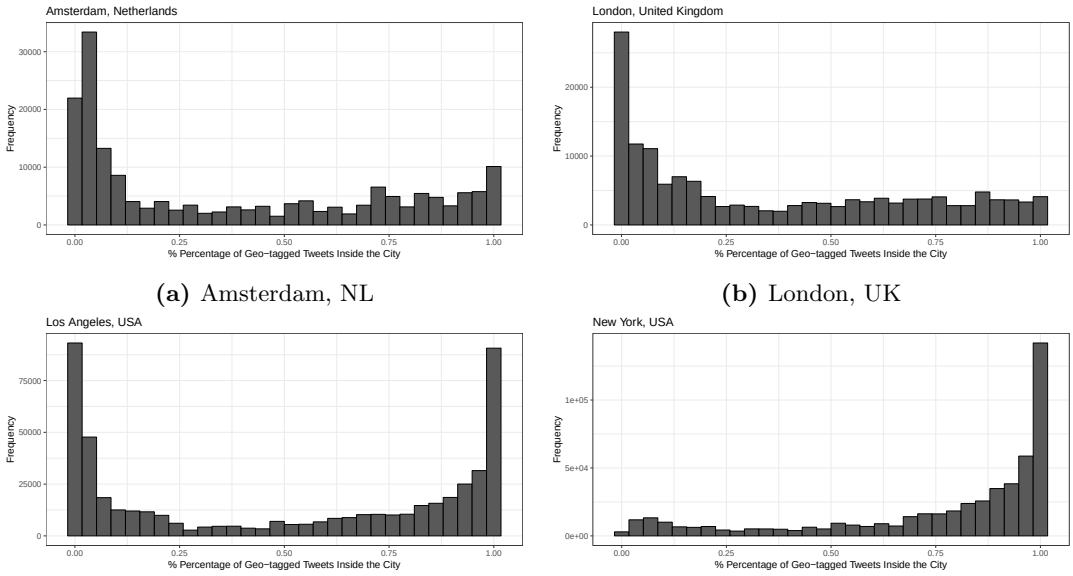
Further explored in this thesis is the activity space of individuals. As described in Section 4.4.3, we could define activity spaces from individuals on the basis of their use of Social Media. This is performed using the dataset defined for different cities around the world described and explored in detail in Sections 5.2.1.1, 5.2.2.1, 5.3.1, 5.4.

The analysis of the location recurrence and activity space is performed on the complete dataset, city residents, and tourist user groups. However, as the collected data does not contain any information concerning the residency location of the different users. Therefore, the identification of the various users' groups residency is decided based on the percentage of the geotagged tweets inside the city boundaries to the total tagged tweets (as also applied int he case of Athens in Section 6.3). Aiming at reducing the uncertainty in this classification, a rule–based approach is followed: if an individual have posted less than 25 percent (<0.25) of the total geotagged tweets inside the cities' limits they are considered tourists, and if the percentage of geotagged tweets within city boundaries is greater than 50 percent (>0.50) they are considered residents. The status of the third group with geotagged tweets outside of the chosen range of 0.25 and 0.50 is unclear. Table 6.2 shows the percentage of each user group for the collected data for each of the subject cities.

Table 6.2: Percentage of the Different User Groups per City

City	% of Resident	% of Unclear	% of Tourist
Amsterdam, NL	58.39	27.60	13.84
Athens, GR	53.35	11.47	22.28
Copenhagen, DK	57.16	14.38	24.38
London, UK	46.68	12.22	20.02
Los Angeles, USA	47.00	28.06	24.90
Munich, DE	66.74	16.49	14.83
New York, USA	49.87	25.54	24.49
Orlando, USA	56.30	20.92	20.67
Paris, FR	55.63	16.88	21.84
Seattle, USA	58.75	28.40	12.74

Examples of the distributions for different cities is presented in Figure 6.17. As it is clearly illustrated, there are major differences between the different cities, which is an indication of both the posting activity of residents and the extend to which a city is a tourists destination. It has to be noted, that as it is presented in Table 5.1 for most of these cases there has been given a limited time–window for the collection of users. Thus, given the random character of the data collection, these results might not entirely reflect the actual posting distribution. However, it is still considered l valuable to understand, under similar data collection processes and for a limited data collection period, what would be the actual distribution of tween within the city examined.



(a) Amsterdam, NL

(b) London, UK

(c) Los Angeles, USA

(d) New York, USA

Figure 6.17: Examples of Inside City User Geo-tagged tweets

6.4.1 Location Recurrence

Starting from the location recurrence, for each user the geotagged tweets are investigated for the formation of spatial clusters of different in time posts. This sheds some light on aspects of transferability of methods and solutions which use Social Media. When examined, we could identify if there are strong differences or similarities with regards to habitual posting from specific locations that can be associated with specific activities (e.g. the activities performed as in Section 6.2). The analysis of location recurrence has been performed by specifying the characteristics of the clusters based on the GPS accuracy. The analysis was implemented in R using the dbscan library, which applies the density-based algorithm for discovering clusters in large spatial databases with noise originally developed by Ester et al. [1996]. The parameters were selected after examination of various settings taking into account the GPS accuracy [Schaefer and Woodyer, 2015] and the number of tweets that each individual posts: the neighborhood of a point parameter (Eps) was defined to be 0.002 and the minimum number of points to be 5. The usability of this analysis is based on the investigation of the way users use Twitter, distinguishing the use of Twitter to post extraordinary locations visited (in terms of the users' "mean" activity space) from the ordinary Social Media use. The analysis of the location recurrence yield interesting results (Table 6.3).

First, for the aggregated data representing all the users groups, the mean number of clusters varies from 3 (London) to 8 (New York). For the resident group, the mean number of clusters ranges between 2 (London) and 6 (New York and Los Angeles). The tourist group mean number of clusters varies between 4 (London) and 11 (Seattle). For the total aggregated data, there is a clear difference between Europe and the US, with US cities illustrating a larger mean number of clusters, except for Seattle that illustrates a mean number of clusters close to the European mean. This primarily illustrates the difference in the cities' structure and the fact that users in the USA post (recurrently) from more locations that users in Europe. The resident user groups represent the same pattern for the different cities with a lesser mean number of clusters compared to the total data. The tourist user groups show a slightly different pattern with the increase of the mean number of clusters for all the cities compared to the previously described two groups. The larger number of clusters for the tourist groups replicates the expected tourist behavior of visiting more locations compared to the city residents. The difference between the US and Europe is still evident for the tourist groups except for Amsterdam having a mean of 10 clusters, which is higher than Los Angeles and Orlando mean number of clusters.

The same characteristics are observed in the mean number of points in clusters and the opposite in the mean number of points characterized as noise. Finally, the number of clusters inside the city seem to follow the same pattern (more frequent visits at the same areas) in the USA in comparison to Europe. Considering all the above, the Users' Location Recurrence analysis illustrate the differences in use of Social Media in the examined areas and particularly between Europe and the USA. It is found that in Europe, users mainly post geotagged tweets when visiting locations that can be characterized as not frequently visited (presumably for leisure activities or special

Table 6.3: Analysis of User Location Recurrence for Different Groups

City	Mean Number of Clusters			Mean Point in Cluster			Mean Noise Point			Mean Cluster In City		
	Total	Res	Tour	Total	Res	Tour	Total	Res	Tour	Total	Res	Tour
AmsterdamL	5.11	3.92	10.26	82.05	82.21	121.38	77.57	55.15	163.47	1.26	2.00	0.08
Athens	5.32	4.35	5.67	69.46	59.44	72.42	88.03	67.34	89.56	1.09	1.97	0.00
Copenhagen	5.74	4.63	6.09	81.38	69.91	86.61	92.19	72.73	89.22	0.98	1.61	0.01
London	3.04	2.35	3.76	39.42	30.97	53.36	60.15	50.37	59.18	0.81	1.61	0.01
Los Angeles	6.14	5.88	5.40	155.65	185.76	152.10	70.18	54.37	78.16	3.78	5.28	0.85
New York	7.95	5.88	10.95	190.37	226.08	144.63	73.63	38.84	124.04	5.23	5.42	3.15
Munich	4.78	3.54	7.86	66.93	49.92	112.88	78.43	56.69	135.43	0.91	1.30	0.04
Orlando	6.42	5.64	7.23	108.07	99.64	130.75	71.39	56.23	89.99	2.02	3.32	0.04
Paris	5.54	4.62	6.44	73.52	62.00	92.22	86.60	72.07	90.08	1.53	2.55	0.05
Seattle	5.47	4.45	11.36	92.57	92.72	147.73	70.23	45.56	171.24	2.20	3.02	0.21

Table 6.4: Power-Law Properties of Location Recurrence Clusters - All Users

Total Data									
City	Points in Cluster			Noise Points			Number of Clusters		
	p-value	x_{min}	α	p-value	x_{min}	α	p-value	x_{min}	α
Amsterdam	0.46	302	1.43	0.07	196	1.27	0.16	18	1.73
Athen	0.28	321	1.45	0.34	257	1.25	0.00	16	1.70
Copenhagen	0.49	364	1.43	0.04	212	1.25	0.93	30	1.68
London	0.37	220	1.54	0.01	178	1.28	0.37	22	1.91
Los Angeles	0.00	164	1.35	0.08	158	1.27	0.27	26	1.7
Munich	0.83	406	1.47	0.28	172	1.26	0.00	18	1.74
New York	0.00	233	1.31	0.95	234	1.26	0.77	33	1.60
Orlando	0.00	317	1.39	0.29	130	1.27	0.00	21	1.66
Paris	0.00	277	1.44	0.86	293	1.25	0.00	11	1.68
Seattle	0.00	201	1.41	0.00	113	1.28	0.61	26	1.74

events). On the other hand users from the USA tend to post geotagged tweets from locations they visit frequently.

Apart from the generic analysis of the location clusters and in order to extract information concerning the transferability of the findings to other cities, the examination of the classification-related variables distributions properties took place. Particularly, given the observed resulting distributions, in addition to the findings of the literature, the fitting of Power-Law distribution was found to be the most prominent distribution. The examination of the applicability of the Power-Law properties in the Location Recurrence Clusters data was performed using the poweRlaw library in R [Clauset et al., 2009]. The results of the fitting are presented in Table 6.5 with an example of the log-log plots for the number of users location recurrency clusters is presented in Figure 6.18.

For the total data-set, it is clearly evident that in half of the cases and only starting from a large number of x_{min}, we were able to detect Power-Law distributions (p-value$\leq$ 0.05 where 0.05 is the significant level examined – using the Kolmogorov Smirnoff test), indicating that the examined variables (Number of Clusters, Number of Points in a Cluster and Number of Noise Points) cannot be characterized by the Power-Law distribution [Clauset et al., 2009]. For the resident group data, the Power-Law distribution is evident in most of the cities (7 cities). For the tourist group data, Power-Law distribution is not evident in half of the cases. This finding further supports the previously examined characteristics that indicate the strong difference in Social Media use across the two examined continents and even across different cities. The finding of the Power Law distributions are related to the first law of Geography. Essentially, what we observe is a different degree of relatedness in different cities. This is a function of the people that live and visit those places, the infrastructure and spatial distribution of

Table 6.5: Power-Law Properties of Location Recurrence Clusters - Residents and Tourists user classes

Resident data									
City	Points in Cluster			Noise Points			Number of Clusters		
	p-value	x_{min}	α	p-value	x_{min}	α	p-value	x_{min}	α
Amsterdam	0.26	282	1.42	0.88	132	1.30	0.00	13	1.82
Athens	0.01	165	1.48	0.01	171	1.27	0.00	6	1.80
Copenhagen	0.01	198	1.47	0.00	150	1.26	0.10	20	1.78
London	0.49	159	1.60	0.02	128	1.29	0.00	10	2.09
Los Angeles	0.00	156	1.33	0.06	154	1.28	0.00	8	1.74
Munich	0.44	273	1.53	0.00	72	1.28	0.00	14	1.90
New York	0.00	178	1.30	0.18	114	1.31	0.00	16	1.75
Orlando	0.00	306	1.40	0.02	125	1.28	0.90	26	1.70
Paris	0.00	138	1.47	0.00	182	1.26	0.00	16	1.75
Seattle	0.00	91	1.42	0.00	119	1.31	0.00	7	1.85
Tourist Data									
City	Points in Cluster			Noise Points			Number of Clusters		
	p-value	x_{min}	α	p-value	x_{min}	α	p-value	x_{min}	α
Amsterdam	0.00	126	1.36	0.62	218	1.29	0.64	17	1.51
Athens	0.00	131	1.43	0.07	189	1.32	0.17	18	1.64
Copenhagen	0.33	344	1.41	0.31	310	1.26	0.01	15	1.62
London	0.00	110	1.48	0.06	106	1.30	0.00	6	1.76
Los Angeles	0.00	59	1.35	0.00	106	1.28	0.02	14	1.79
Munich	0.96	420	1.36	0.39	147	1.29	0.28	18	1.53
New York	0.55	386	1.33	0.27	232	1.29	0.95	27	1.47
Orlando	0.11	390	1.36	0.00	102	1.28	0.00	12	1.64
Paris	0.32	281	1.40	0.52	308	1.26	0.01	18	1.60
Seattle	0.04	217	1.34	0.00	127	1.22	0.53	21	1.51

business establishments, predominant cultural traits, and reporting habits using social media.

6.4.2 Cluster-based Activity Space

The activity space from Social Media was examined on clustered data based on points mutual distances again using the DBSCAN algorithm. This choice was based on the need to identify the areas visited by individuals on a local scale, avoiding the formation of very large activity spaces that could result from long distance traveling for tourism or business purposes, as this is out of the scope of this work. The parameters for this analysis were selected taking into account the upper level of commuting distance for

the countries examined: the neighborhood of a point parameter (Eps) was defined to be 1.2 and the minimum number of points to be 3. Additionally, in order to extract characteristics of the city examined, the clusters were characterized as near the city examined (in case the distance between the center of the cluster and the center of the city was smaller 120 km). The cluster analysis was performed on the basis of the cluster in the city examined, as well as including the two additional largest clusters (in terms of points in the clusters). Table 6.7 presents some aggregated activity space characteristics. This analysis also confirms the differences in the use of Social Media in Europe and in the USA, while it is worth noting that all European cases have an significantly larger activity space in comparison to the activity space of the American cases for the different users groups. Besides, the area of the tourist groups activity space is more extensive than the resident groups' activity space which confirm the rational travel pattern of tourists.

As presented in Table 6.8, in the majority of the cases the number of geotagged tweets belong to the examined city's cluster. It should be noted that the fact that a large number of tweets is classified as of being in the city might seem contradictory, when compared with the percentage of in city tweets, however it illustrates that in many cases the strict administrative areas of the cities do not necessarily represent the individuals who commonly use the city; while it should also been taken into account that in extreme cases some tweets could even be posted almost 300 km away from the city center and still belong to the classified as in the city classification (as we only consider the distance of the class center to the city center). The largest percentage is observed in New York city, while the lowest in Copenhagen and Munich. Another interesting fact is that apparently only a small percentage of geotagged tweets do not belong to a cluster. This finding is subject to the low minimum number of points specification and the large Eps parameter used. Finally, in most cases a vast majority of geotagged tweets are included in either one of three examined clusters.

Table 6.6: Activity Space Characteristics, for Different Groups

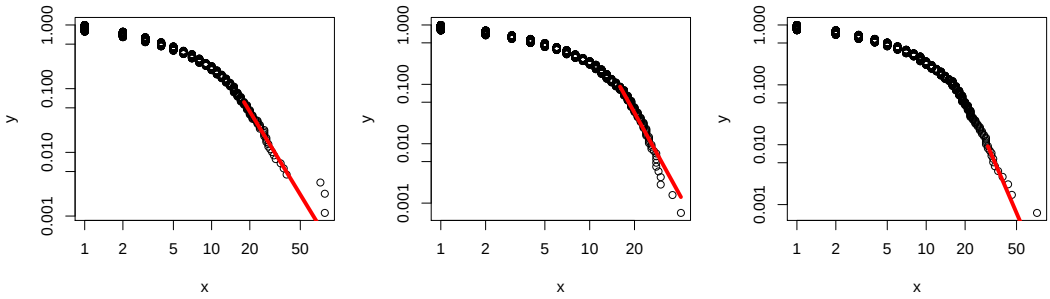
City	Mean Number of Clusters			Mean Points in Exam. City Cluster			Mean Points in Second Larger Cluster			Mean Points in First Larger Cluster		
	Total	Res	Tour	Total	Res	Tour	Total	Res	Tour	Total	Res	Tour
Amsterdam	5.24	2.96	13.95	77.43	112.36	12.77	48.81	14.67	119.95	19.7	15.07	25.90
Athens	5.59	3.78	6.19	63.11	76.19	8.06	58.02	33.56	75.79	24.55	16.39	26.13
Copenhagen	6.26	4.88	5.98	58.98	74.13	13.06	63.47	36.78	87.07	25.83	15.99	38.3
London	3.82	3.07	3.98	54.04	60.27	37.87	35.45	10.46	65.78	16.31	9.77	24.31
Los Angeles	3.86	2.23	5.55	180.27	223.78	172.18	52.99	12.32	98.87	20.89	10.5	33.8
New York	3.54	1.46	7.34	212.06	258.7	110.05	36.18	8.79	73.64	19.48	7.53	28.78
Munich	5.73	4.06	10.63	42.37	52.11	14.59	53.69	30.07	99.22	22.66	16.9	32.35
Orlando	4.23	2.88	6.93	105.48	123.71	97.93	61	23.55	100	21.2	12.64	30.85
Paris	5.62	4.64	5.81	61.32	80.24	10.27	54.53	22.27	95.51	23.83	15.52	36.38
Seattle	4.92	2.32	16.00	96.15	123.6	30.25	34.83	12.55	85.46	19.18	9.66	38.48

Table 6.7: Activity Space Characteristics, for Different Groups

City	Mean Number of noise point			Activity Space in Exam. City (1E+07km2)			Activity Space in First Larger Cluster (1E+07km2)			Activity Space in First Larger Cluster (1E+07km2)		
	Total	Res	Tour	Total	Res	Tour	Total	Res	Tour	Total	Res	Tour
Amsterdam	4.62	2.35	14.53	2.1	1.32	4.95	2.56	1.87	4.88	3.21	2.81	5.04
Athens	4.29	3.05	4.99	2.05	1.03	6.44	2.46	1.33	2.72	3.07	1.76	3.35
Copenhagen	5.13	4.44	4.89	2.84	1.9	5.32	2.84	2.19	2.33	3.28	2.55	2.85
London	3.54	3.24	2.98	1.89	1.65	2.14	2.24	2.39	1.44	2.93	3.18	2.08
Los Angeles	4.31	2.39	6.44	1.79	1.31	2.22	2.18	2.03	1.59	3.05	2.97	2.41
New York	2.97	1.18	6.34	1.57	0.60	2.98	2.25	1.18	3.22	2.78	1.42	3.55
Munich	5.95	3.81	14.99	1.87	1.42	4.44	2.15	1.67	3.4	2.53	2.07	4.07
Orlando	3.84	2.38	7.60	1.06	0.84	1.44	1.34	1.16	1.6	1.77	1.59	2.21
Paris	4.42	3.81	4.66	2.53	1.68	5.45	2.62	2.03	2.76	3.08	2.45	3.44
Seattle	4.77	1.86	18.35	1.14	0.57	2.89	1.68	1.01	3.26	2.14	1.45	3.38

Table 6.8: Clustering Characteristics for Different Groups

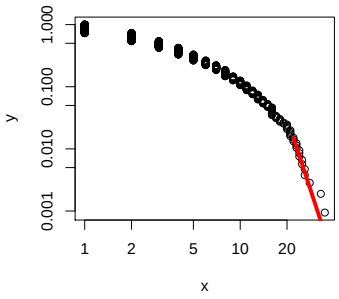
City	Mean % In Examined to Geotagged			Mean % in First Larger Cluster to Geotagged			Mean % in Second Larger Cluster to Geotagged			Mean % of Noise to Geotagged		
	Total	Res	Tour	Total	Res	Tour	Total	Res	Tour	Total	Res	Tour
Amsterdam	0.34	0.53	0.02	0.31	0.16	0.46	0.13	0.09	0.11	0.05	0.05	0.05
Athens	0.32	0.51	0.00	0.36	0.25	0.55	0.14	0.12	0.19	0.05	0.05	0.07
Copenhagen	0.24	0.40	0.00	0.36	0.26	0.55	0.14	0.11	0.22	0.05	0.06	0.05
London	0.43	0.68	0.01	0.33	0.14	0.62	0.13	0.09	0.21	0.07	0.07	0.07
Los Angeles	0.64	0.92	0.28	0.31	0.07	0.62	0.10	0.04	0.15	0.03	0.02	0.04
Munich	0.28	0.41	0.01	0.35	0.28	0.52	0.14	0.14	0.15	0.07	0.07	0.06
New York	0.74	0.94	0.30	0.15	0.03	0.30	0.08	0.02	0.12	0.01	0.01	0.03
Orlando	0.36	0.59	0.00	0.34	0.19	0.51	0.11	0.08	0.13	0.03	0.03	0.03
Paris	0.35	0.57	0.01	0.34	0.19	0.60	0.14	0.11	0.20	0.05	0.05	0.05
Seattle	0.48	0.69	0.02	0.22	0.09	0.30	0.11	0.05	0.13	0.04	0.02	0.07



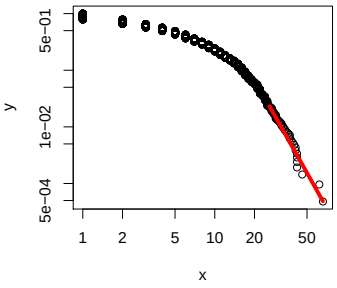
(a) Amsterdam, NL

(b) Athens, GR

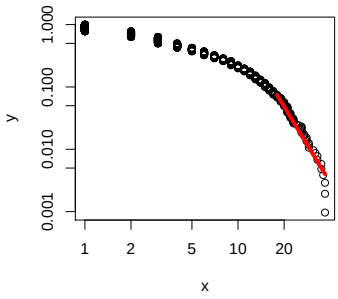
(c) Copenhagen, DK



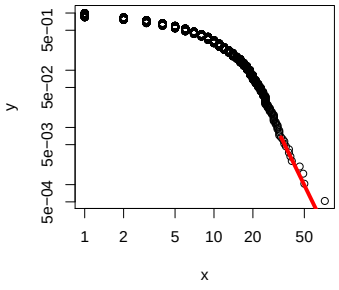
(d) London, UK



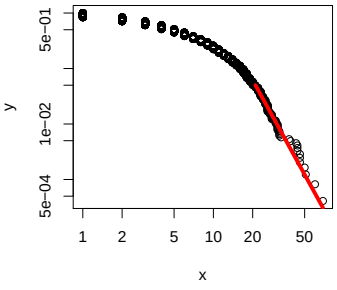
(e) Los Angeles, USA



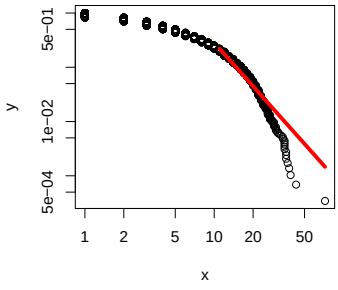
(f) Munich, DE



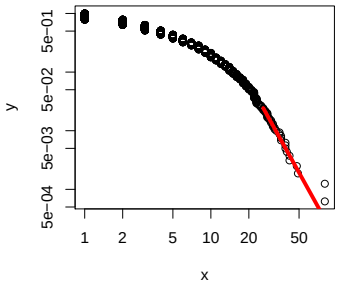
(g) New York, USA



(h) Orlando, USA



(i) Paris, FR



(j) Seattle, USA

Figure 6.18: Power-Law Plot for Users Locations Clusters

7 Conclusions and Future Work

Contents

7.1 Conclusions

With the development of disruptive technological concepts, heterogeneous data sources will continue to emerge. Although not strictly defined as transportation data, they might illustrate properties that would potentially enrich our understanding of the transportation systems. They can be used to enrich conventional methodologies and data collection frameworks in all sectors, as these emerging sources of data can extend the understanding of users mobility patterns and their choices. The improvements in terms of understanding can be identified in different transportation–related fields such as those of users activities, system observability, long-term behavioural changes and incident detection. These fields can be directly identified based on the differences observed by the comparison of the information acquired using conventional data collection methods and emerging ones.

Social Media in particular can be a game-changer in many research fields, including transportation. Its characteristics –such as the continuous stream of information; user-generated content; combination of temporal, spatial, and textual information; and existence of a social network representation–have created a stream of research that focuses on the exploitation of the information that can be directly extracted. However, the study of Social Media requires the connection of users' online and offline worlds so that their identified behaviour and patterns can be generalized and delineated. Empirical evidence suggests that SM cannot displace currently used survey methods, with the most applicable area of research being the fusion of various data sources based on identified misrepresentations and commonly known issues. The literature review presented highlights advances and methods that have guided the use of Social Media in Transportation Research. The most significant outcome is the shift from the exploration of the data properties towards the investigation of its potential to substitute or complement conventional transport data. It should be noted, however, that besides the merits of using Social Media in transportation studies, the highlighted issues concerning privacy and data availability can force the use of poorer -in terms of information- datasets that may rely in extremely expensive data collection processes, of small sample sizes. The same is evidenced with other (big) data owners that -although collecting data of high value for transportation- refuse to make it available for research purposes.

This work presented a collection of methods to analyse, enrich and, in general, use Social Media data to extract transportation–related features. The resulting methods aim at shredding light to the use of Social Media, in general, and its utilization for transportation purposes, in particular. The methods presented herein, focus on the exploration of a) data collection methods; b) descriptive measures for the investigation of the Social Media use and the comparison to conventional data collection methods; c) the Social Media use around the world and finally d) the investigation of transportation–related features extraction as well as data enrichment using data from Social Media. The various methods defined have been applied in different case studies, with the aim to be the exploration of different urban setting.

Specifically, a generic data collection methodology for Social Media and a case–specific data collection methodology for Twitter have been defined and used to collect

data from Social Media. The data collection process has been explored through an empirical analysis on commonly used spaces in transportation: spatial, temporal and activity-related. This has been performed through the comparison of different cities around the world, as well as targeting specific cases. In order to extract aggregated characteristics, classification techniques were implemented and power-law distributions were examined, without after-all identifying clear power-law properties. The analysis performed illustrates that various differences can be identified especially when comparing data from the USA and Europe. Within those geographical areas, there are again differences but not as striking as the ones examined on a continent level. The results of the analysis are promising and indicative of the transportation–related Twitter use for specific population groups (such as high income), certain temporal patterns and certain areas. The analysis results indicate the increased usage during non–working hours at leisure areas that can be found to be mostly related to leisure activities. Those findings illustrate that users post only a fragment of their activities that are most likely to be leisure–related.

Additionally, the examination of the activity space, and the temporal and spatial characteristics, illustrate that Social Media cannot be readily used as a metric to directly extract demand models, when compared to existing travel survey methods. However, there is mounting evidence that SM data can be used as an additional source of information to enrich the conventional transport–related data collection methods (especially considering that their collection can be easily automated and has low recurring costs).

Of high interest is the feature extraction methods defined in this dissertation. With regards to activity identification and data enrichment, a methodological framework was defined, for the characterization of transport–related activities for data originating from Social Media, and specifically from Twitter. This has been performed using inferred data from Foursquare essentially defining a combined dataset. The resulted combined data is then used to derive classification methodologies for activity characterization using Twitter text. The methods to infer location characterizations (and, as a consequence, activities) and to enrich data, based on the user–centric activity data enrichment, have been applied for a large dataset from London, UK. The results of the analysis indicate the capabilities of the proposed methods to derive activity patterns for individuals and to further use data from Social Media in transportation research and more specifically in activity–based models. Furthermore, the classification performance indicates that, for a sufficient number of cases, the data enrichment can indicate activities performed for all tweets recorded, based on the posted text, which –given the use of endogenous posting data– suggests that the generalization of the proposed methodology in other languages is possible.

The analysis performed clearly differentiates the use of Social Media in different areas around the world, providing a basis for the comparison of the suitability of different techniques for different geographical areas. The user-based approach that was followed, provided a better understanding of the Social Media related behaviour and has the potential for exploration of transferability of methods across different contexts. As the analysis suggest, the distinction between residents and tourists is fundamental for

the development of models, as there are major differences in terms of behaviour, for all variables examined. Additionally, the exploration of activity spaces is believed to benefit from a clustering based examination, especially when used for social media data analysis. With the proposed methodology, the definition of activity spaces can be formalized to represent different contexts of activities such as areas that are frequently visited in a period or in a particular area, avoiding points, which act as 'noise' in the samples. This is especially suitable in cases of wider in time-span datasets (such as data originating from Social Media), due to the possibility of using places visited only once. Additionally, activity space definition also benefits from the distinction between tourists and residents, who exhibit different behaviours. The exploration of the power-law distribution yielded also results concerning the various activity spaces characteristics. Although power law distribution could not be safely defined, it is believed that the exploration of the parameters of the distribution fit illustrates that there are similarities concerning activity spaces for different areas.

With regards to the activity space definition, the implications of the framework's deployment can be first understood in terms of direct merits in decoding social media data in useful, inexpensive information concerning activities that are commonly found to be under–represented in conventional travel surveys. Users activity patterns can also be derived, and the modelling of long–term activity–chains can be realized, as Social Media users share information in a much larger span of time, in comparison to conventional surveys. Second, this framework could be combined with the conventional travel surveys in order to identify the magnitude of the misrepresented activities (on both Social Media and travel surveys) as well as allow for information enrichment methodologies.

The first part of the presented framework on the user–centric activity identification, is believed to set the grounds for the deployment of Social Media into activity–based models, as it not only allows for the identification of activities from Social Media data but also enriches the fragmented information of posts via the transformation to user–centric information. The latter can allow for handling the sample bias that Social Media data bears, when making the connection of the online and offline world on a user's level (for example with a survey of Social Media users characteristics). The second part of the herein developed framework on the classification, allows for recognition of activities based solely on the text of tweets. The application of this framework builds a much larger dataset, which again might be able to reduce the bias Social Media bears. Finally, it should be noted that the framework could also be deployed on data under similar typology, from other Social Media platforms (such as Facebook, or Instagram).

7.2 Future Work

The examination of the various aspects of Social Media reveal the potential of using Social Media in the transportation sector. The methods defined and the analysis performed in the various case studies provide the basis for the further exploration of heterogeneous data sources that can potentially improve the representation of the trans-

portation system and allow us to better understand mobility and people's choices. The findings of this research provide guidance towards several aspects of future research on the use of emerging data sources and specifically Social Media in transportation.

Although there are some tendencies that can be observed from the analysis of the Social Media use, there is still a lot of room for further exploration. First of all, the examination of the representativeness of the sample is considered important, as it was found that the sample is biased towards some classes of the total population –only a proportion is using those services and only for some types of activities. In this direction, future research should concentrate on modelling users choices that are connected to sharing information, something that is yet to be explored especially for transportation. This could be performed on the basis of the inclusion of Social Media IDs in the performed travel surveys so that the combined analysis of the two datasets could yield results with regards to the connection between the online and offline work of users. The joint analysis would bring valuable information on the connection of the reported and posted activities.

Another important direction that exists as a direct outcome of this thesis is the examination of methods and models that fuse data from different data sources under the objective of increasing the knowledge of the transportation system. Further investigation is required to more conclusively pinpoint the spatial and temporal parameters that determine the relative merit of Social Media data. One way of performing this could be the exploration of the inclusion of Social Media posts as extra information on calibration methods. This could be performed either on the basis of online or offline calibration (Section 7.3). In a similar way, the fusion of Social Media data with other sources could also be useful in order to define sequences of activities for activity–based schedulers in combination with activities commonly reported on Social Media and those reported in travel surveys.

With regards to the worldwide comparison in the ten cities examined herein, although the sampling method is believed to be suitable for the analysis performed, the number of users should be increased for more conclusive analysis. Additionally, the methods used could be extended to include Latent Class clustering (e.g. as in Chaniotakis et al. [2012]), which could be used for the investigation of differences and similarities between different groups of individuals; thus could be applied for the case of social media users as well. Application of such methods could allow for exploration of different groups of Social Media users, based on their posting activity and would allow the enrichment of the information we gather from Social Media data. Finally, the investigation of distributions concerning importantly identified variables defining user-posting behaviour and its connection to mobility as well as the definition of models, would further enhance the understanding of the differences and similarities. Furthermore, it would be interesting to repeat this analysis in other geographical areas, in order to determine the extent to which these findings are transferable or scalable, or whether different patterns might emerge in different study areas.

In regard to the extraction of transportation-related features, several directions are also identified. Firstly, concerning the Areas of Interest (AoI), it is believed that the further characterization of AoI as a concept that could eventually replace POIs, and

allow for a further specification of leisure attractions, is necessary in mobility. Such a definition, would allow for the evolution of research on the way that travel destinations are chosen and could, in the long run, lead on dynamically defined areas of activities, in the course of one day or in larger time intervals. With regards to the activity space exploration, temporal and activity characteristics, location recurrence clusters and spatial settings can help define models. Results of such a sensitivity analysis is presented in Appendix A, however further investigation is required, for a higher number of users and taking into account different clusters of users.

Finally, with regards to the extraction of activities using data enrichment and text classification, it should be mentioned that although the results of the analysis seem promising, there are some drawbacks that should be discussed, concerning the wide application of the methodology in other study areas and especially upon the final text classification for the inference of activities. To begin with, the methodologies are bounded to the wide–spread existence of labelling in the area examined (such as Foursquare or generally points of interest), which might not be the case in some countries or regions. Additionally, the classification has been performed only on posts that include a Foursquare URL or have been clustered in commonly visited venues by user locations. Within this data, only a small fraction was attributed to non–activities related posts such as drinking wine and running (denoted as activities, as opposed to –for example– posting or commenting on news). For a proper representation of the activity space of individuals, based on text classification, non–activity related tweets should be detected (either using manual labelling or classification algorithms) to allow for a more complete representation of the reasons that users post on Social Media. This could be achieved by the investigation of different potential sources of information either by recognizing patterns of activities that users perform and inferring the misrepresented activities or by allowing for cross—validation, using, for example, information about from news, special events or disruptions and tie this information to tweets, based on a spatial and temporal level. Finally, although the examined algorithms seem to perform well, the performance of additional classification algorithms and the sensitivity of their parameters should be examined.

7.3 Towards Operationalization of Social Media data

Driven by the characteristics of Social Media, literature findings and the inherent uncertainty of Dynamic Demand Estimation, it is believed that it is possible to investigate ways to enrich information for Dynamic Demand Estimation using Social Media data. The transportation demand formulation from [Cascetta, 2009, page 237] is adapted in order to allow for its description as a function of sociodemographic (vector **SE**), travel related characteristics (vector **T**), the vector of parameters used to describe demand ($\boldsymbol{\beta}$) as well as a vector of related data that do not fall in the conventionally used variables, for example social media data (vector **L**).

$$D = f(\mathbf{SE}, \mathbf{T}, \mathbf{L}; \boldsymbol{\beta}) \tag{7.1}$$

Fusion methods are based on: a) probabilistic and statistical models such as Bayesian reasoning, evidence theory, robust statistics, and recursive operators; b) least-square (LS) and mean square methods such as Kalman Filter, optimization, regularization, and uncertainty ellipsoids; c) other heuristic methods such as ANNs, fuzzy logic, approximate reasoning, and computer vision. This multi-source fusion problem can be viewed as a structural inference or model selection problem, which is described by a set of measurement relationships and the description of the system dynamics. A special type of multi-source fusion is the one using non-commensurate data, referring to information of the different typology. Fusion of non-commensurate information is prohibited at its source (raw data fusion), unless it includes the transformation of information. As a consequence fusion has to take place at the feature (state vector) level, or at the decision level [Hall and Llinas, 1997].

In a structure that capture its dynamics, the transportation system can be described as a state-space model for which transition equations are used to represent the evolution of a state vector and measurement equations are used to map the states on the measurements. Let $\mathbf{X}_{h+1}$ be a state vector, representing the minimal set of data to describe the (dynamic) behaviour of the system at time interval $h+1$. A simple representation of the state equation can be expressed as a function of the previous state $\mathbf{X}_h$ using a transition matrix ($\mathbf{F}_{h+1}$) describing the dynamics of the system (Equation 7.2).

$$\mathbf{X}_{h+1} = \mathbf{F}_{h+1} \cdot \mathbf{X}_h + \boldsymbol{\epsilon}_h^s \tag{7.2}$$

where $\boldsymbol{\epsilon}_h^s \sim N(0, Q_h)$ with $\mathbf{Q}_h$ to denote the covariance matrix.

The measurement equations include the direct measurement (Equation 7.3) of the origin destination matrix (in most cases considered as historical OD measurements) and indirect measurements related to the observations that are not directly related to the unknown aspect of the system such as the traffic counts (Equation 7.4) and the social media data (Equation 7.5) (e.g. Origin Destination magnitudes or Social Media check-ins).

$$\mathbf{X}_h^o = \mathbf{X}_h + \mathbf{e}_h^z \tag{7.3}$$

$$\mathbf{M}_h = \mathbf{S}\left(\mathbf{X}_h\right) + \mathbf{e}_h^m \tag{7.4}$$

$$\mathbf{L}_h = \mathbf{K}\left(\mathbf{X}_h\right) + \mathbf{e}_h^l \tag{7.5}$$

where $\mathbf{X}_h^o$ is the direct observations measurements (if available) of the state vector (for example OD flows), e_h^z a vector of random error terms, $\mathbf{M}_h$ the indirect measurement vector referring to conventional transport related measurement, e_h^m a vector of random measurement related error terms, $\mathbf{L}_h$ the vector of indirect observations from social media data, $\mathbf{e}_h^l$ for time interval h; $\mathbf{S}$ a dynamic traffic simulator and $\mathbf{K}$ a model representing the relationship between the observed social media measurements and the state vector $\mathbf{X}_h$. The Dynamic Demand Estimation problem is an optimization problem with its objective function to be:

$$\min_{X_h} \left[\mathcal{N}_1(\mathbf{e}_h^s) + \mathcal{N}_1(\mathbf{e}_h^z) + \mathcal{N}_2(\mathbf{e}_h^m) + \mathcal{N}_3(\mathbf{e}_h^l) \right] \tag{7.6}$$

with $\mathcal{N}_i(\cdot)$ to describe a magnitude measurement function [Antoniou et al., 2007].

The solution to the Dynamic Demand Estimation has been widely investigated in the literature [Antoniou et al., 2007, Zhou and Mahmassani, 2006, Cipriani et al., 2010, Cantelmo et al., 2014]. In almost all cases, this investigation has been conducted using traffic counts or travel time observations.

The main requirement for the inclusion of Social Media data in the dynamic demand estimation problem is to model the relationship of Social Media with demand (in the form of Origin Destination flow vector). As it has been shown in this dissertation, one–way check–in can be related to a perform activity, which can define the start and the end of a trip (Figure 7.1). The trip definition could take place on the base of arrival and departure time. For example the arrival can be defined as a function of the Social Media post time $t_a = h_c - k$ where k represents a random variable $k \cdot$ and h_c the time that a check-in takes place. Similarly, departure can be defined: $t_d = h_c + q$, where q a random variable $q \cdot$.

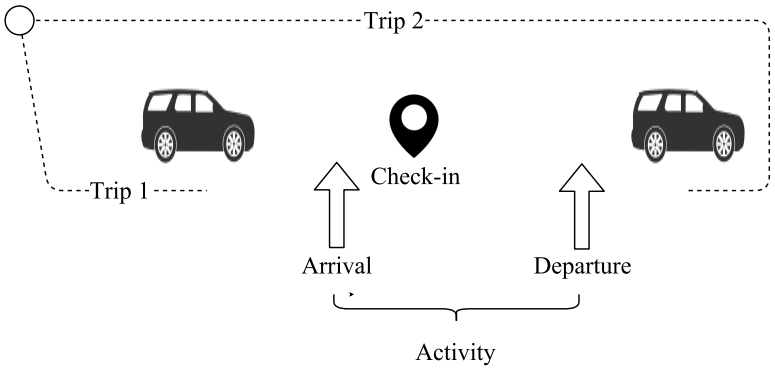


Figure 7.1: Trips and Social Media posts

Based on the above, the connection between transport demand and one-way check-ins can be defined.

$$D_{j,h} = \beta \cdot L_{j,h} + \epsilon_{DL} \tag{7.7}$$

or

$$D_{i,h} = \beta \cdot L_{i,h} + \epsilon_{DL} \tag{7.8}$$

where $D_{j,h}$ and $D_{i,h}$ represents the arrivals and departures to and from a specific area. As suggested in the pertinent literature Jin et al. [e.g. 2014], Social Media data can be related to the Origin–Destination matrices with the use of a gravity model.

$$D_{ij} = K_i \cdot D_i \cdot K_j \cdot D_j \cdot f(c_{ij}) \tag{7.9}$$

By combining Equation 7.8 and 7.9 we can derive the connection between the two quantities.

$$D_{ij} = K_i \cdot D_i \cdot K_j \cdot L_j \cdot f(c_{ij}) + \epsilon_{DL} \tag{7.10}$$

where K_i and K_j the calibrated gravity factors. The exploration of similar approaches for the better representation of demand for transportation systems could increase the actual accuracy of transport–related models and extend the scope of currently available techniques. The concept of deviations [Antoniou et al., 2007] as well as proper filtering could be used to reduce the noise of Social Media data and provide the basis for the operationalization of their use in mobility–related modelling.

Bibliography

Fabian Abel, Claudia Hauff, Geert-Jan Houben, Richard Stronkman, and Ke Tao. Twitcident: Fighting fire with information from social web streams. In *Proceedings of the 21st International Conference Companion on World Wide Web*, WWW '12 Companion, pages 305–308, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1230-1. doi: 10.1145/2187980.2188035. URL `http://doi.acm.org/10.1145/2187980.2188035`.

Nihan Akyelken, David Banister, and Moshe Givoni. The sustainability of shared mobility in london: The dilemma for governance. *Sustainability*, 10(2):420, 2018.

C. Antoniou, M. Ben-Akiva, and H. N. Koutsopoulos. Nonlinear kalman filtering algorithms for on-line calibration of dynamic traffic assignment models. *IEEE Transactions on Intelligent Transportation Systems*, 8(4):661–670, Dec 2007. ISSN 1524-9050. doi: 10.1109/TITS.2007.908569.

Theo Arentze and Harry Timmermans. Social networks, social interactions, and activity-travel behavior: a framework for microsimulation. *Environment and Planning B: Planning and Design*, 35(6):1012–1027, 2008.

Vittorio Astarita, Vincenzo Pasquale Giofrè, Giuseppe Guido, and Alessandro Vitale. A single intersection cooperative-competitive paradigm in real time traffic signal settings based on floating car data. *Energies*, 12(3):409, 2019.

Ivonne Audirac. Stated Preference for Pedestrian Proximity: An Assessment of New Urbanist Sense of Community. *Journal of Planning Education and Research*, 19(1):53–66, September 1999. URL `http://jpe.sagepub.com/content/19/1/53.abstract`.

Kay W Axhausen. Social networks, mobility biographies, and travel: survey challenges. *Environment and planning. B, Planning & design*, 35(6):981, 2008. ISSN 0265-8135.

Danya Bachir, Ghazaleh Khodabandelou, Vincent Gauthier, Mounim El Yacoubi, and Jakob Puchinger. Inferring dynamic origin-destination flows by transport mode using mobile phone data. *Transportation Research Part C: Emerging Technologies*, 101: 254–275, 2019.

M. Bagchi and P.R. White. The potential of public transport smart card data. *Transport Policy*, 12(5):464 – 474, 2005. ISSN 0967-070X. doi: https://doi.org/10.1016/j.tranpol.2005.06.008. URL `http://www.sciencedirect.com/science/article/pii/S0967070X05000855`. Road User Charging: Theory and Practices.

Donald P Ballou and Harold L Pazer. Modeling data and process quality in multi-input, multi-output information systems. *Management science*, 31(2):150–162, 1985.

Jaume Barceló, Lidin Montero, Laura Marqués, and Carlos Carmona. Travel time forecasting and dynamic origin-destination estimation for freeways based on bluetooth traffic monitoring. *Transportation research record*, 2175(1):19–27, 2010.

Carlo Batini, Cinzia Cappiello, Chiara Francalanci, and Andrea Maurino. Methodologies for data quality assessment and improvement. *ACM computing surveys (CSUR)*, 41(3):16, 2009.

Hila Becker, Mor Naaman, and Luis Gravano. Beyond trending topics: Real-world event identification on twitter. In *Fifth international AAAI conference on weblogs and social media*, 2011.

Russell Belk. You are what you can access: Sharing and collaborative consumption online. *Journal of business research*, 67(8):1595–1600, 2014.

Moshe E Ben-Akiva and Steven R Lerman. *Discrete choice analysis: theory and application to travel demand*, volume 9. MIT press, 1985.

Ashish Bhaskar and Edward Chung. Fundamental understanding on the use of bluetooth scanner as a complementary transport data. *Transportation Research Part C: Emerging Technologies*, 37:42–72, 2013.

Paul P Biemer, Robert M Groves, Lars E Lyberg, Nancy A Mathiowetz, and Seymour Sudman. *Measurement errors in surveys*, volume 173. John Wiley & Sons, 2011.

Jit Biswas, Felix Naumann, and Qiang Qiu. Assessing the completeness of sensor data. In *International Conference on Database Systems for Advanced Applications*, pages 717–732. Springer, 2006.

Isabelle Bloch. Information combination operators for data fusion: a comparative review with classification. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 26(1):52–67, 1996.

Enrique Bonsón, Lourdes Torres, Sonia Royo, and Francisco Flores. Local e-government 2.0: Social media and corporate transparency in municipalities. *Government Information Quarterly*, 29(2):123 – 132, 2012. ISSN 0740-624X. doi: http://dx.doi.org/10.1016/j.giq.2011.10.001. URL `http://www.sciencedirect.com/science/article/pii/S0740624X1200010X`.

Enrique Bonsón, David Perea, and Michaela Bednárová. Twitter as a tool for citizen engagement: An empirical study of the andalusian municipalities. *Government Information Quarterly*, 0, 2019.

Danah M. Boyd and Nicole B. Ellison. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1):210–230, 2007. ISSN 10836101. doi: 10.1111/j.1083-6101.2007.00393.x.

Susan Bregman. *Uses of social media in public transportation*. Number 99. Transportation Research Board, 2012. ISBN 0309223571.

Werner Brog and Arnim H Meyburg. Nonresponse problem in travel surveys: an empirical investigation. *Transportation Research Record*, 775:34–38, 1980.

Stephen Buckley and Deborah Lightman. Ready or not, big data is coming to a city (transportation agency) near you. In *Transportation Research Board 94th Annual Meeting*, number 15-5156 in TRB2015, 2015.

Andrew Bwambale, Charisma F Choudhury, and Stephane Hess. Modelling trip generation using mobile phone data: a latent demographics approach. *Journal of Transport Geography*, 0, 2017.

Guido Cantelmo, Francesco Viti, Chris Tampêre, Ernesto Cipriani, and Marialisa Nigro. Two-step approach for correction of seed matrix in dynamic demand estimation. *Transportation Research Record: Journal of the Transportation Research Board*, 2466 (1):125–133, 2014.

Irene Casas and Elizabeth C Delmelle. Tweeting about public transit—gleaning public perceptions from a social media microblog. *Case Studies on Transport Policy*, 5(4): 634–642, 2017.

Ennio Cascetta. *Transportation systems analysis: models and applications*, volume 29. Springer Science & Business Media, 2009. ISBN 9780387758565.

Raymond Chan and Joseph L Schofer. Role of social media in communicating transit disruptions. *Transportation Research Record*, 2415(1):145–151, 2014.

E. Chaniotakis and C. Antoniou. On the activity space derived social media: Recurrence, temporal and spatial sensitivity analysis. In *6th hEART Symposium (European Association for Research in Transportation)*, Technion, Haifa, 12-14 September, 2017, 2017.

E. Chaniotakis, C. Antoniou, J. M. S. Grau, and L. Dimitriou. Can social media data augment travel demand survey data? In *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1642–1647, Nov 2016a. doi: 10.1109/ITSC.2016.7795778.

E. Chaniotakis, C. Antoniou, and F. Pereira. Mapping social media for transportation studies. *IEEE Intelligent Systems*, 31(6):64–70, Nov 2016b. ISSN 1541-1672. doi: 10.1109/MIS.2016.98.

Emmannouil Chaniotakis, Adam Davis, Georgia Aifadopoulou, Constantinos Antoniou, and Konstantinos Goulias. A Latent Class Cluster Comparison of Travel Behavior Between Thessaloniki in Greece and San Diego in California. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing*, pages 239–248. IEEE, 2012.

Emmanouil Chaniotakis and Constantinos Antoniou. Use of geotagged social media in urban settings: Empirical evidence on its potential from twitter. In *IEEE 18th International Conference on Intelligent Transportation Systems, (ITSC)*, pages 214–219. IEEE, 2015.

Emmanouil Chaniotakis, Constantinos Antoniou, and Loukas Dimitriou. *Digital Social Networks and Travel Behaviour in Urban Environments*, chapter Social Media and Travel Behaviour., pages 0–0. CRC Press, Cham, 2010a. ISBN 9781138594630. URL `https://www.crcpress.com/Digital-Social-Networks-and-Travel-Behaviour-in-Urban-Environments/Plaut-Pinsly/p/book/9781138594630`.

Emmanouil Chaniotakis, Dimitrios Efthymiou, and Constantinos Antoniou. *Demand for Emerging Transportation Systems*, chapter Data Aspects of the Evaluation of demand for Transportation Systems, pages 0–0. Elsevier, Cham, 2010b. ISBN 9780128150184. URL `https://www.elsevier.com/books/demand-for-emerging-transportation-systems/antoniou/978-0-12-815018-4`.

Emmanouil Chaniotakis, Constantinos Antoniou, and Evangelos Mitsakis. Data for leisure travel demand from social networking services. In *4th hEART Symposium*. European Association for Research in Transportation - hEART, DTU, Denmark, 2015.

Emmanouil Chaniotakis, Constantinos Antoniou, Georgia Ayfantopoulou, and Dimitriou Loukas. Inferring activities from social media data. In *Transportation Research Board 96th Annual Meeting*, 2017a.

Emmanouil Chaniotakis, Constantinos Antoniou, and Konstantinos Goulias. Transferability and sample specification for social media data: a comparative analysis. In *Proceedings of the mobil.TUM 2017 Conference, 4-5 July, Munich Germany*, Jan 2017b.

Cyrille Médard De Chardon, Geoffrey Caruso, and Isabelle Thomas. Bicycle sharing system 'success' determinants. *Transportation Research Part A: Policy and Practice*, 100:202 – 214, 2017. ISSN 0965-8564. doi: https://doi.org/10.1016/j.tra.2017.04.020.

Po-Ta Chen, Feng Chen, and Zhen Qian. Road traffic congestion monitoring in social media with hinge-loss markov random fields. In *2014 IEEE International Conference on Data Mining*, pages 80–89. IEEE, 2014.

Yimin Chen, Xiaoping Liu, Xia Li, Xingjian Liu, Yao Yao, Guohua Hu, Xiaocong Xu, and Fengsong Pei. Delineating urban functional areas with building-level social media data: A dynamic time warping (dtw) distance based k-medoids method. *Landscape and Urban Planning*, 160:48–60, 2017.

Zesheng Cheng, Sisi Jian, Mojtaba Maghrebi, Taha Hossein Rashidi, and S Travis Waller. Is social media an appropriate data source to improve travel demand estimation models? In *Transportation Research Board*, 2018.

Zhiyuan Cheng, James Caverlee, Kyumin Lee, and Daniel Z Sui. Exploring Millions of Footprints in Location Sharing Services. *ICWSM*, 2011:81–88, 2011.

Eunjoon Cho, Seth A Myers, and Jure Leskovec. Friendship and Mobility: User Movement in Location-based Social Networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 1082–1090, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0813-7. doi: 10.1145/2020408.2020579. URL `http://doi.acm.org/10.1145/2020408.2020579`.

Yoon-Sik Cho, Greg Ver Steeg, and Aram Galstyan. Where and why users "check in". In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, AAAI'14, pages 269–275. AAAI Press, 2014. URL `http://dl.acm.org/citation.cfm?id=2893873.2893917`.

Ernesto Cipriani, Michael Florian, Michael Mahut, and Marialisa Nigro. Investigating the Efficiency of a Gradient Approximation Approach for the Solution of Dynamic Demand Estimation Problems. In *New Developments in Transport Planning*, Chapters, chapter 18. Edward Elgar Publishing, September 2010. URL `https://ideas.repec.org/h/elg/eechap/13831_18.html`.

Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

Craig Collins, Samiul Hasan, and Satish V Ukkusuri. A novel transit rider satisfaction metric: Rider sentiments measured from online social media data. *Journal of Public Transportation*, 16(2):2, 2013.

Caitlin Cottrill, Francisco Pereira, Fang Zhao, Ines Ferreira Dias, Hock Beng Lim, Moshe Ben-Akiva, and Chris Zegras. Future mobility survey. *Transportation Research Record: Journal of the Transportation Research Board*, 2354:59–67, 12 2013. doi: 10.3141/2354-07.

Caitlin Cottrill, Paul Gault, Godwin Yeboah, John D Nelson, Jillian Anable, and Thomas Budd. Tweeting transit: An examination of social media strategies for transport information management during a large event. *Transportation Research Part C: Emerging Technologies*, 77:421–432, 2017.

Judd Cramer and Alan B Krueger. Disruptive change in the taxi business: The case of uber. *American Economic Review*, 106(5):177–82, 2016.

Barry E. Cushing. A mathematical approach to the analysis and design of internal control systems. *The Accounting Review*, 49(1):24–41, 1974. ISSN 00014826. URL `http://www.jstor.org/stable/244795`.

Mazen Danaf, Bilge Atasoy, Carlos Lima de Azevedo, Jing Ding-Mastera, Maya Abou-Zeid, Nathaniel Cox, Fang Zhao, and Moshe Ben-Akiva. Context-aware stated preferences with smartphone-based travel surveys. *Journal of Choice Modelling*, 31:35 – 50, 2019. ISSN 1755-5345. doi: https://doi.org/10.1016/j.jocm.2019.03.001. URL http://www.sciencedirect.com/science/article/pii/S1755534518300381.

Ricardo A. Daziano, Mauricio Sarrias, and Benjamin Leard. Are consumers willing to pay to let cars drive for them? analyzing response to autonomous vehicles. *Transportation Research Part C: Emerging Technologies*, 78:150 – 164, 2017. ISSN 0968-090X. doi: https://doi.org/10.1016/j.trc.2017.03.003. URL http://www.sciencedirect.com/science/article/pii/S0968090X17300682.

João de Abreu e Silva and Konstadinos G. Goulias. Structural equations model of land use patterns, location choice, and travel behavior: Seattle, washington, compared with lisbon, portugal. *Transportation Research Record*, 2135(1):106–113, 2009. doi: 10.3141/2135-13. URL https://doi.org/10.3141/2135-13.

Nic DePaula, Ersin Dincelli, and Teresa M Harrison. Toward a typology of government social media communication: Democratic goals, symbolic acts and self-presentation. *Government Information Quarterly*, 35(1):98–108, 2018.

Sean T Doherty, Nathalie Noël, M Lee Gosselin, Claude Sirois, and Mami Ueno. Moving beyond observed outcomes: integrating global positioning systems and interactive computer-based travel behavior surveys. Technical report, Transportation Research Board, 2001.

David Durán Rodas, Emmanouil Chaniotakis, and Constantinos Antoniou. Built environment factors affecting bike sharing ridership: A data-driven approach for multiple cities. *Transportation Research Record (in press)*, 0(0):0, 2019.

Dimitrios Efthymiou and Constantinos Antoniou. Measuring the effects of transportation infrastructure location on real estate prices and rents: investigating the current impact of a planned metro line. *EURO Journal on Transportation and Logistics*, 3 (3-4):179–204, 2014.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, pages 226–231. AAAI Press, 1996. URL http://dl.acm.org/citation.cfm?id=3001460.3001507.

Xiang Fei and Hani S. Mahmassani. Structural analysis of near-optimal sensor locations for a stochastic large-scale network. *Transportation Research Part C: Emerging Technologies*, 19(3):440 – 453, 2011. ISSN 0968-090X. doi: https://doi.org/10.1016/j.trc.2010.07.001. URL http://www.sciencedirect.com/science/article/pii/S0968090X10001105.

E. Fishman, S. Washington, and N. Haworth. Bike share: A synthesis of the literature. *Transport Reviews*, 33(2):148–165, 2013. doi: 10.1080/01441647.2013.775612. cited By 98.

Vanessa Frias-Martinez and Enrique Frias-Martinez. Spectral clustering for sensing urban land use using Twitter activity. *Engineering Applications of Artificial Intelligence*, 35:237–245, 2014. ISSN 09521976. doi: 10.1016/j.engappai.2014.06.019. URL `http://dx.doi.org/10.1016/j.engappai.2014.06.019`.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

M. R. Friesen and R. D. McLeod. Bluetooth in intelligent transportation systems: A survey. *International Journal of Intelligent Transportation Systems Research*, 13 (3):143–153, Sep 2015. ISSN 1868-8659. doi: 10.1007/s13177-014-0092-1. URL `https://doi.org/10.1007/s13177-014-0092-1`.

Adrian Furnham. Response bias, social desirability and dissimulation. *Personality and Individual Differences*, 7(3):385 – 400, 1986. ISSN 0191-8869. doi: https://doi.org/10.1016/0191-8869(86)90014-0. URL `http://www.sciencedirect.com/science/article/pii/0191886986900140`.

Ayelet Gal-Tzur, Susan M. Grant-Muller, Tsvi Kuflik, Einat Minkov, Silvio Nocera, and Itay Shoor. The potential of social media in delivering transport policy goals. *Transport Policy*, 32:115–123, 2014a. ISSN 0967070X. doi: 10.1016/j.tranpol.2014.01.007.

Ayelet Gal-Tzur, Susan M. Grant-Muller, Einat Minkov, and Silvio Nocera. The Impact of Social Media Usage on Transport Policy: Issues, Challenges and Recommendations. *Procedia - Social and Behavioral Sciences*, 111:937–946, 2014b. ISSN 18770428. doi: 10.1016/j.sbspro.2014.01.128. URL `http://www.sciencedirect.com/science/article/pii/S1877042814001293`.

Huiji Gao, Geoffrey Barbier, and Rebecca Goolsby. Harnessing the crowdsourcing power of social media for disaster relief. *IEEE Intelligent Systems*, 26(3):10–14, 2011.

Juan Carlos García-Palomares, Javier Gutiérrez, and Carmen Mínguez. Identification of tourist hot spots based on social networks: A comparative analysis of european metropolises using photo-sharing services and gis. *Applied Geography*, 63:408–417, 2015.

M. Gentili and P.B. Mirchandani. Locating sensors on traffic networks: Models, challenges and research opportunities. *Transportation Research Part C: Emerging*

Technologies, 24:227 – 255, 2012. ISSN 0968-090X. doi: https://doi.org/10.1016/j.trc.2012.01.004. URL `http://www.sciencedirect.com/science/article/pii/S0968090X1200006X`.

Susan M. Grant-Muller, Ayelet Gal-Tzur, Einat Minkov, Silvio Nocera, Tsvi Kuflik, and Itay Shoor. Enhancing transport data collection through social media sources: methods, challenges and opportunities for textual data. *IET Intelligent Transport Systems*, 2014. ISSN 1751-9578.

Patrick Greenfield. The Cambridge Analytica files: the story so far, 2018.

Danielle Griego, Varin Buff, Eric Hayoz, Izabela Moise, and Evangelos Pournaras. Sensing and mining urban qualities in smart cities. In *2017 IEEE 31st International Conference on Advanced Information Networking and Applications (AINA)*, pages 1004–1011. IEEE, 2017.

Yiming Gu, Zhen Sean Qian, and Feng Chen. From twitter to detector: Real-time traffic incident detection using social media data. *Transportation research part C: emerging technologies*, 67:321–342, 2016. ISSN 0968-090X. doi: http://dx.doi.org/10.1016/j.trc.2016.02.011. URL `http://www.sciencedirect.com/science/article/pii/S0968090X16000644`.

David L Hall and James Llinas. An introduction to multisensor data fusion. *Proceedings of the IEEE*, 85(1):6–23, 1997.

Arturo Haro-de Rosario, Alejandro Sáez-Martín, and María del Carmen Caba-Pérez. Using social media to enhance citizen engagement with local government: Twitter or facebook? *New Media & Society*, 20(1):29–49, 2018.

Samiul Hasan and Satish V Ukkusuri. Urban activity pattern classification using topic models from online geo-location data. *Transportation Research Part C: Emerging Technologies*, 44:363–381, 2014.

Samiul Hasan and Satish V Ukkusuri. Reconstructing activity location sequences from incomplete check-in data: A semi-markov continuous-time bayesian network model. *IEEE Transactions on Intelligent Transportation Systems*, 19(3):687–698, 2018.

Bartosz Hawelka, Izabela Sitko, Euro Beinat, Stanislav Sobolevsky, Pavlos Kazakopoulos, and Carlo Ratti. Geo-located Twitter as the proxy for global mobility patterns., 2013. URL `http://arxiv.org/abs/1311.0680`.

Jingrui He, Wei Shen, Phani Divakaruni, Laura Wynter, and Rick Lawrence. Improving traffic prediction with tweet semantics. In *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.

Dirk Helbing and Evangelos Pournaras. Society: Build digital democracy. *Nature News*, 527(7576):33, 2015.

David A Hensher. Stated preference analysis of travel choices: the state of practice. *Transportation*, 21(2):107–133, 1994.

David A Hensher, John M Rose, and William H Greene. Inferring attribute non-attendance from stated choice data: implications for willingness to pay estimates and a warning for stated choice experiment design. *Transportation*, 39(2):235–245, 2012.

Stephane Hess, John M Rose, and John Polak. Non-trading, lexicographic and inconsistent behaviour in stated choice data. *Transportation Research Part D: Transport and Environment*, 15(7):405–417, 2010.

Wangsu Hu and Peter J Jin. Dynamic origin–destination estimation based on time-delay correlation analysis on location-based social network data. Technical report, 2018.

Qunying Huang and David WS Wong. Modeling and visualizing regular human mobility patterns with uncertainty: An example using twitter data. *Annals of the Association of American Geographers*, 105(6):1179–1197, 2015.

Zhiren Huang, Ximan Ling, Pu Wang, Fan Zhang, Yingping Mao, Tao Lin, and Fei-Yue Wang. Modeling real-time human mobility based on mobile phone and transportation data fusion. *Transportation Research Part C: Emerging Technologies*, 96:251–269, 2018.

Jesper Bláfoss Ingvardson, Otto Anker Nielsen, Sebastián Raveau, and Bo Friis Nielsen. Passenger arrival and waiting time distributions dependent on train service frequency and station characteristics: A smart card data analysis. *Transportation Research Part C: Emerging Technologies*, 90:292–306, 2018.

HV Jagadish, Johannes Gehrke, Alexandros Labrinidis, Yannis Papakonstantinou, Jignesh M Patel, Raghu Ramakrishnan, and Cyrus Shahabi. Big data and its technical challenges. *Communications of the ACM*, 57(7):86–94, 2014.

B. Jiang and Y. Miao. The Evolution of Natural Cities from the Perspective of Location-Based Social Media. *ArXiv e-prints*, January 2014.

Shan Jiang, Ana Alves, Filipe Rodrigues, Joseph Ferreira Jr, and Francisco C Pereira. Mining point-of-interest data from social networks for urban land use classification and disaggregation. *Computers, Environment and Urban Systems*, 53:36–46, 2015. ISSN 0198-9715. doi: 10.1016/j.compenvurbsys.2014.12.001. URL `http://dx.doi.org/10.1016/j.compenvurbsys.2014.12.001`.

Peter Jin, Meredith Cebelak, Fan Yang, Jian Zhang, C Walton, and Bin Ran. Location-based social networking data: Exploration into use of doubly constrained gravity model for origin-destination estimation. *Transportation Research Record: Journal of the Transportation Research Board*, (2430):72–82, 2014.

R. A. Johnson and D. W. Wichern, editors. *Applied Multivariate Statistical Analysis*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2002. ISBN 0-130-41146-9.

Chen Jun and Yang Dongyuan. Estimating smart card commuters origin-destination distribution based on apts data. *Journal of Transportation Systems Engineering and Information Technology*, 13(4):47 – 53, 2013. ISSN 1570-6672. doi: https://doi.org/10.1016/S1570-6672(13)60116-6. URL http://www.sciencedirect.com/science/article/pii/S1570667213601166.

Daniel Kahneman. *Thinking, fast and slow*. Macmillan, 2011.

Andreas M Kaplan and Michael Haenlein. Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, 53(1):59–68, jan 2010. ISSN 0007-6813. doi: http://dx.doi.org/10.1016/j.bushor.2009.09.003. URL http://www.sciencedirect.com/science/article/pii/S0007681309001232.

A Kheiri, F Karimipour, and M Forghani. Intra-urban movement flow estimation using location based social networking data. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 40(1):781, 2015.

Jan H. Kietzmann, Kristopher Hermkens, Ian P. McCarthy, and Bruno S. Silvestre. Social media? Get serious! Understanding the functional building blocks of social media. *Business Horizons*, 54(3):241–251, 2011. ISSN 00076813. doi: 10.1016/j.bushor.2011.01.005. URL http://dx.doi.org/10.1016/j.bushor.2011.01.005.

Rob Kitchin. *The data revolution: Big data, open data, data infrastructures and their consequences*. Sage, 2014.

Richard A Krueger and Mary Anne Casey. Designing and conducting focus group interviews, 2002.

Avinash Kumar, Miao Jiang, and Yi Fang. Where not to go?: detecting road hazards using twitter. *Proceedings of the 37th international ACM ...*, 2609550:1223–1226, 2014. doi: 10.1145/2600428.2609550. URL http://dl.acm.org/citation.cfm?id=2609550.

Kathrin Kühne, Suman K. Mitra, and Jean-Daniel M. Saphores. Without a ride in car country – a comparison of carless households in germany and california. *Transportation Research Part A: Policy and Practice*, 109:24 – 40, 2018. ISSN 0965-8564. doi: https://doi.org/10.1016/j.tra.2018.01.021. URL http://www.sciencedirect.com/science/article/pii/S0965856417305621.

Gwanhoo Lee and Young Hoon Kwak. An open government maturity model for social media-based public engagement. *Government Information Quarterly*, 29(4):492 – 503, 2012. ISSN 0740-624X. doi: http://dx.doi.org/10.1016/j.giq.2012.06.001. URL http://www.sciencedirect.com/science/article/pii/S0740624X1200086X. Social Media in Government - Selections from the 12th Annual International Conference on Digital Government Research (dg.o2011).

Jae Hyun Lee, Adam W Davis, Seo Youn Yoon, and Konstadinos G Goulias. Activity space estimation with longitudinal observations of social media data. *Transportation*, 43(6):955–977, 2016.

Jae Hyun Lee, Adam Davis, Elizabeth McBride, and Konstadinos G Goulias. Statewide comparison of origin-destination matrices between california travel model and twitter. In *Mobility Patterns, Big Data and Transport Analytics*, pages 201–228. Elsevier, 2019.

Rui Li, Shengjie Wang, and Kevin Chen-Chuan Chang. Multiple location profiling for users and relationships from social network and content. *Proceedings of the VLDB Endowment*, 5(11):1603–1614, 2012.

Dong Lin, Andrew Allan, and Jianqiang Cui. The impacts of urban spatial structure and socio-economic factors on patterns of commuting: a review. *International Journal of Urban Sciences*, 19(2):238–255, 2015. doi: 10.1080/12265934.2015.1016092. URL `https://doi.org/10.1080/12265934.2015.1016092`.

Bruce R Lindsay. Social media and disasters: Current uses, future options, and policy considerations, 2011.

Jenny H Liu, Wei Shi, OA Sam Elrahman, Xuegang Jeff Ban, and Jack M Reilly. Understanding social media program usage in public transit agencies. *International Journal of Transportation Science and Technology*, 5(2):83–92, 2016.

Jenny H Liu, Xuegang Ban, and OA Elrahman. Measuring the impacts of social media on advancing public transit. 2017a.

Xiaoping Liu, Jialv He, Yao Yao, Jinbao Zhang, Haolin Liang, Huan Wang, and Ye Hong. Classifying urban land use by integrating remote sensing and social media data. *International Journal of Geographical Information Science*, 31(8):1675–1696, 2017b.

Yu Liu, Zhengwei Sui, Chaogui Kang, and Yong Gao. Uncovering patterns of inter-urban trip and spatial interaction from social media check-in data. *PloS one*, 9(1): e86026, 2014.

Germán C Lleras, Anja Simma, M Ben-Akiva, Andreas Schafer, Kay W Axhausen, and Tomoyuki Furutani. Fundamental relationships specifying travel behavior—an international travel survey comparison. In *Proceedings of 82nd Annual Meeting of the Transportation Research Board, Washington, DC*, 2002.

Jordan J Louviere, David A Hensher, and Joffre D Swait. *Stated choice methods: analysis and applications*. Cambridge university press, 2000.

Xiaolei Ma, Yao-Jan Wu, Yinhai Wang, Feng Chen, and Jianfeng Liu. Mining smart card data for transit riders' travel patterns. *Transportation Research Part C: Emerging Technologies*, 36:1 – 12, 2013. ISSN 0968-090X. doi: https://doi.org/10.1016/

j.trc.2013.07.010. URL http://www.sciencedirect.com/science/article/pii/S0968090X13001630.

Mary Madden, Amanda Lenhart, Sandra Cortesi, Urs Gasser, Maeve Duggan, Aaron Smith, and Meredith Beaton. Teens, social media, and privacy. *Pew Research Center*, 21:2–86, 2013.

Mojtaba Maghrebi, Alireza Abbasi, and S Travis Waller. Transportation application of social media: Travel mode extraction. In *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1648–1653. IEEE, 2016.

Giacomo Manetti, Marco Bellucci, and Luca Bagnoli. Stakeholder engagement and public information through social media: a study of canadian and american public transportation agencies. *The American Review of Public Administration*, 47(8):991–1009, 2017.

Daniel McFadden. Economic choices. *American Economic Review*, 91(3):351–378, June 2001. doi: 10.1257/aer.91.3.351. URL http://www.aeaweb.org/articles?id=10.1257/aer.91.3.351.

Chuishi Meng, Yu Cui, Qing He, Lu Su, and Jing Gao. Travel purpose inference with gps trajectories, pois, and geo-tagged social media data. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 1319–1324. IEEE, 2017.

Evangellos Mitsakis and Panagiotis Iordanopoulos. Seeits: Deliverable d7.1.1: Cost-benefit analysis report for the deployment of its in greece. Technical report, 2014.

David L Morgan. *The focus group guidebook*, volume 1. Sage publications, 1997.

Mark Morrison. Aggregation biases in stated preference studies. *Australian Economic Papers*, 39(2):215–230, 2000.

James J. Murphy, P. Geoffrey Allen, Thomas H. Stevens, and Darryl Weatherhead. A meta-analysis of hypothetical bias in stated preference valuation. *Environmental and Resource Economics*, 30(3):313–325, Mar 2005. ISSN 1573-1502. doi: 10.1007/s10640-004-3332-z. URL https://doi.org/10.1007/s10640-004-3332-z.

Felix Naumann, editor. *Information Quality Criteria*, pages 29–50. Springer Berlin Heidelberg, Berlin, Heidelberg, 2002. ISBN 978-3-540-45921-7. doi: 10.1007/3-540-45921-9_3. URL https://doi.org/10.1007/3-540-45921-9_3.

Anton J. Nederhof. Methods of coping with social desirability bias: A review. *European Journal of Social Psychology*, 15(3):263–280, 1985. doi: 10.1002/ejsp.2420150303. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/ejsp.2420150303.

Ming Ni, Qing He, and Jing Gao. Forecasting the subway passenger flow under event occurrences with social media. *IEEE Transactions on Intelligent Transportation Systems*, 18(6):1623–1632, June 2016. ISSN 1524-9050. doi: 10.1109/TITS.2016.2611644.

Kamal Nigam, John Lafferty, and Andrew McCallum. Using maximum entropy for text classification. In *IJCAI-99 workshop on machine learning for information filtering*, volume 1, pages 61–67, 1999.

Anastasios Noulas, Cecilia Mascolo, and Enrique Frias-Martinez. Exploiting foursquare and cellular data to infer user activity in urban environments. In *2013 IEEE 14th International Conference on Mobile Data Management*, volume 1, pages 167–176. IEEE, 2013.

Martin T Orne. On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American psychologist*, 17(11):776, 1962.

Jorge David Gonzalez Paule, Yeran Sun, and Yashar Moshfeghi. On fine-grained geolocalisation of tweets and real-time traffic incident detection. *Information Processing & Management*, 56(3):1119 – 1132, 2019. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2018.03.011. URL `http://www.sciencedirect.com/science/article/pii/S0306457317305575`.

Delroy L. Paulhus. Chapter 2 - measurement and control of response bias. In John P. Robinson, Phillip R. Shaver, and Lawrence S. Wrightsman, editors, *Measures of Personality and Social Psychological Attitudes*, pages 17 – 59. Academic Press, 1991. ISBN 978-0-12-590241-0. doi: https://doi.org/10.1016/B978-0-12-590241-0.50006-X. URL `http://www.sciencedirect.com/science/article/pii/B978012590241050006X`.

F. C. Pereira, F. Rodrigues, E. Polisciuc, and M. Ben-Akiva. Why so many people? explaining nonhabitual transport overcrowding with internet data. *IEEE Transactions on Intelligent Transportation Systems*, 16(3):1370–1379, June 2015. ISSN 1524-9050. doi: 10.1109/TITS.2014.2368119.

Francisco C. Pereira, Filipe Rodrigues, and Moshe Ben-Akiva. Using Data From the Web to Predict Public Transport Arrivals Under Special Events Scenarios. *Journal of Intelligent Transportation Systems*, (June 2014):1–16, 2013. ISSN 1547-2450. doi: 10.1080/15472450.2013.868284. URL `http://www.tandfonline.com/doi/abs/10.1080/15472450.2013.868284`.

Pew Research Center. Smartphone ownership is growing rapidly around the world, but not always equally. Technical report, 2019.

Ioannis Politis, Patrick Langdon, Mike Bradley, Lee Skrypchuk, Alexander Mouzakitis, and P. John Clarkson. Designing autonomy in cars: A survey and two focus groups on driving habits of an inclusive user group, and group attitudes towards autonomous cars. In Giuseppe Di Bucchianico and Pete F Kercher, editors, *Advances in Design for Inclusion*, pages 161–173, Cham, 2018. Springer International Publishing. ISBN 978-3-319-60597-5.

Stephen R. Porter, Michael E. Whitcomb, and William H. Weitzer. Multiple surveys of students and survey fatigue. *New Directions for Institutional Research*, 2004(121): 63–73, 1 2004. ISSN 1536-075X. doi: 10.1002/ir.101. URL `https://doi.org/10.1002/ir.101`.

Adrian C. Prelipcean, Gyozo Gidofalvi, and Yusak O. Susilo. Meili: A travel diary collection, annotation and automation system. *Computers, Environment and Urban Systems*, 70:24 – 34, 2018. ISSN 0198-9715. doi: https://doi.org/10.1016/j.compenvurbsys.2018.01.011. URL `http://www.sciencedirect.com/science/article/pii/S0198971517305240`.

John Preston and Fiona Rajé. Accessibility, mobility and transport-related social exclusion. *Journal of transport geography*, 15(3):151–160, 2007.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL `https://www.R-project.org/`.

Taha H Rashidi, Alireza Abbasi, Mojtaba Maghrebi, Samiul Hasan, and Travis S Waller. Exploring the capacity of social media data for modelling travel behaviour: Opportunities and challenges. *Transportation Research Part C: Emerging Technologies*, 75:197–211, 2017. ISSN 0968-090X. doi: https://doi.org/10.1016/j.trc.2016.12.008. URL `http://www.sciencedirect.com/science/article/pii/S0968090X16302625`.

Jonathan Reades, Francesco Calabrese, Andres Sevtsuk, and Carlo Ratti. Cellular census: Explorations in urban data collection. *Pervasive Computing, IEEE*, 6(3): 30–38, 2007. ISSN 1536-1268.

Francisco Rebelo, Carlos Soares, and Rosaldo JF Rossetti. Twitterjam: Identification of mobility patterns in urban centers based on tweets. In *Smart Cities Conference (ISC2), 2015 IEEE First International*, pages 1–6. IEEE, 2015.

Thomas C. Redman. *Data Quality for the Information Age*. Artech House, Inc., Norwood, MA, USA, 1st edition, 1997. ISBN 0890068836.

A. J. Richardson. Behavioral mechanisms of nonresponse in mail-back travel surveys. *Transportation Research Record*, 1855(1):191–199, 2003. doi: 10.3141/1855-24. URL `https://doi.org/10.3141/1855-24`.

Anthony J Richardson, Elizabeth S Ampt, and Arnim H Meyburg. *Survey methods for transport planning*. Eucalyptus Press Melbourne, 1995.

John M Rose, Michiel CJ Bliemer, David A Hensher, and Andrew T Collins. Designing efficient stated choice experiments in the presence of reference alternatives. *Transportation Research Part B: Methodological*, 42(4):395–406, 2008.

Patricia Sánchez Abril, Avner Levin, and Alissa Del Riego. Blurred boundaries: Social media privacy and the twenty-first-century employee. *American Business Law Journal*, 49(1):63–124, 2012. ISSN 0002-7766.

Martin Schaefer and Tara Woodyer. Assessing absolute and relative accuracy of recreation-grade and mobile phone GNSS devices: a method for informing device choice. *Area*, 47(2):185–196, 2015. ISSN 1475-4762. doi: 10.1111/area.12172. URL `http://dx.doi.org/10.1111/area.12172`.

Andreas Schafer. Regularities in travel demand: an international perspective. 2000.

Stefan Schmöller, Simone Weikl, Johannes Müller, and Klaus Bogenberger. Empirical analysis of free-floating carsharing usage: The munich and berlin case. *Transportation Research Part C: Emerging Technologies*, 56:34 – 51, 2015. ISSN 0968-090X. doi: https://doi.org/10.1016/j.trc.2015.03.008. URL `http://www.sciencedirect.com/science/article/pii/S0968090X1500087X`.

Axel Schulz, Petar Ristoski, and Heiko Paulheim. I see a car crash: Real-time detection of small scale incidents in microblogs. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7955 LNCS:22–33, 2013a. ISSN 03029743. doi: 10.1007/978-3-642-41242-4_3.

Axel Schulz, Petar Ristoski, and Heiko Paulheim. I see a car crash: Real-time detection of small scale incidents in microblogs. In *Extended semantic web conference*, pages 22–33. Springer, 2013b.

Tim Schwanen, Frans M Dieleman, and Martin Dijst. Travel behaviour in dutch monocentric and policentric urban systems. *Journal of Transport Geography*, 9(3):173 – 186, 2001. ISSN 0966-6923. doi: https://doi.org/10.1016/S0966-6923(01)00009-6. URL `http://www.sciencedirect.com/science/article/pii/S0966692301000096`. Mobility and Spatial Dynamics.

Lisa Schweitzer. Planning and social media: a case study of public transit and stigma on twitter. *Journal of the American Planning Association*, 80(3):218–238, 2014. doi: 10.1080/01944363.2014.980439. URL `http://dx.doi.org/10.1080/01944363.2014.980439`.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Li Shen and Peter R Stopher. Review of gps travel survey and gps data-processing methods. *Transport Reviews*, 34(3):316–334, 2014.

A Simma and K.W Axhausen. Structures of commitment in mode use: a comparison of switzerland, germany and great britain. *Transport Policy*, 8(4):279 – 288, 2001. ISSN 0967-070X. doi: https://doi.org/10.1016/S0967-070X(01)00023-3. URL `http://www.sciencedirect.com/science/article/pii/S0967070X01000233`.

Richard O Sinnott and Shuangchao Yin. Accident black spot identification and verification through social media. In *2015 IEEE International Conference on Data Science and Data Intensive Systems*, pages 17–24. IEEE, 2015.

Matthew Smith, Christian Szongott, Benjamin Henne, and Gabriele Von Voigt. Big data privacy issues in public social media. In *Digital Ecosystems Technologies (DEST), 2012 6th IEEE International Conference on*, pages 1–6. IEEE, 2012. ISBN 1467317039.

Dominic Stead and Stephen Marshall. The relationships between urban form and travel patterns. an international review and evaluation. *European Journal of Transport and Infrastructure Research*, 1(2):113–141, 2001.

Jim Sterne. *Social media metrics: How to measure and optimize your marketing investment*. John Wiley & Sons, 2010.

Patrizia Sulis, Ed Manley, Chen Zhong, and Michael Batty. Using mobility data as proxy for measuring urban vitality. *Journal of Spatial Information Science*, 16:137–162, 2018.

Yeran Sun, Hongchao Fan, Ming Li, and Alexander Zipf. Identifying the city center using human travel flows generated from location-based social networking data. *Environment and Planning B: Planning and Design*, 43(3):480–498, 2016.

Jeannette N Sutton, Leysia Palen, and Irina Shklovski. *Backchannels on the front lines: Emergency uses of social media in the 2007 Southern California Wildfires*. University of Colorado, 2008.

O. Z. Tamin and L. G. Willumsen. Transport demand model estimation from traffic counts. *Transportation*, 16(1):3–26, Mar 1989. ISSN 1572-9435. doi: 10.1007/BF00223044. URL `https://doi.org/10.1007/BF00223044`.

Lei Tang and Piyushimita Vonu Thakuriah. Ridership effects of real-time bus information system: A case study in the city of chicago. *Transportation Research Part C: Emerging Technologies*, 22:146–161, 2012.

Richard H Thaler and Cass R Sunstein. *Nudge: Improving decisions about health, wealth, and happiness*. Penguin, 2009.

Harry Timmermans, Peter van der Waerden, Mario Alves, John Polak, Scott Ellis, Andrew S Harvey, Shigeyuki Kurose, and Rianne Zandee. Spatial context and the complexity of daily travel patterns: an international comparison. *Journal of Transport Geography*, 11(1):37–46, 2003.

Sandar Tin Tin, Alistair Woodward, and Shanthi Ameratunga. Completeness and accuracy of crash outcome data in a cohort of cyclists: a validation study. *BMC public health*, 13(1):420, 2013.

Kenneth E Train. *Discrete choice methods with simulation*. Cambridge university press, 2009.

Andrea Trentini and Federico Losacco. Analyzing carsharing "public" (scraped) data to study urban traffic patterns. *Procedia Environmental Sciences*, 37:594 – 603, 2017. ISSN 1878-0296. doi: https://doi.org/10.1016/j.proenv.2017.03.046. URL `http://www.sciencedirect.com/science/article/pii/S1878029617300464`. Green Urbanism (GU).

Mariko Utsunomiya, John Attanucci, and Nigel Wilson. Potential uses of transit smart card registration and transaction data to improve transit planning. *Transportation Research Record: Journal of the Transportation Research Board*, (1971):119–126, 2006.

Michael AB van Eggermond, Haohui Chen, Alexander Erath, and Manuel Cebrian. Investigating the potential of social network data for transport demand models. *arXiv preprint arXiv:1706.10035*, 2017.

Joan L. Walker, Yanqiao Wang, Mikkel Thorhauge, and Moshe Ben-Akiva. D-efficient or deficient? a robustness analysis of stated choice experimental designs. *Theory and Decision*, 84(2):215–238, Mar 2018. ISSN 1573-7187. doi: 10.1007/s11238-017-9647-3. URL `https://doi.org/10.1007/s11238-017-9647-3`.

Richard Y Wang and Diane M Strong. Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*, 12(4):5–33, 1996.

Zhenzhen Wang, Sylvia Y He, and Yee Leung. Applying mobile phone data to travel behaviour research: A literature review. *Travel Behaviour and Society*, 11:141–155, 2018.

N. Wanichayapong, W. Pruthipunyaskul, W. Pattara-atikom, and P. Chaovalit. Social-based traffic information extraction and classification. In *ITS Telecommunications (ITST), 2011 11th International Conference on*, pages 107–112, Aug 2011. doi: 10.1109/ITST.2011.6060036.

Simon P Washington, Matthew G Karlaftis, and Fred Mannering. *Statistical and econometric methods for transportation data analysis, 2nd edition*. Chapman and Hall/CRC, 2010.

Zheng Xiang and Ulrike Gretzel. Role of social media in online travel information search. *Tourism Management*, 31(2):179–188, 2010. ISSN 02615177. doi: 10.1016/j.tourman.2009.02.016. URL `http://dx.doi.org/10.1016/j.tourman.2009.02.016`.

Yusuke Yamamoto. Java library for the twitters, 2007. URL `http://www.twitter4j.org/`.

Chao Yang, Meng Xiao, Xuan Ding, Wenwen Tian, Yong Zhai, Jie Chen, Lei Liu, and Xinyue Ye. Exploring human mobility patterns using geo-tagged social media data at the group level. *Journal of Spatial Science*, pages 1–18, 2018.

Fan Yang, Peter J Jin, Xia Wan, Rui Li, and Bin Ran. Dynamic origin-destination travel demand estimation using location based social networking data. In *Transportation Research Board 93rd Annual Meeting*, number 14-5509, 2014.

Hong Yao, Muzhou Xiong, Deze Zeng, and Junfang Gong. Mining multiple spatial–temporal paths from social media data. *Future Generation Computer Systems*, 87: 782–791, 2018.

MD Yap, S Nijënstein, and N Van Oort. Improving predictions of public transport usage during disturbances based on smart card data. *Transport Policy*, 61:84–95, 2018.

Dave Yates and Scott Paquette. Emergency knowledge management and social media technologies: A case study of the 2010 haitian earthquake. *International journal of information management*, 31(1):6–13, 2011.

Junjun Yin, Aiman Soliman, Dandong Yin, and Shaowen Wang. Depicting urban boundaries from a mobility network of spatial interactions: a case study of great britain with geo-located twitter data. *International Journal of Geographical Information Science*, 31(7):1293–1313, 2017.

Xianyuan Zhan, Satish V Ukkusuri, and Feng Zhu. Inferring urban land use using large-scale social media check-in data. *Networks and Spatial Economics*, 14(3-4): 647–667, 2014.

Zhenhua Zhang, Ming Ni, Qing He, Jing Gao, Jizhan Gou, and Xiaoling Li. Exploratory study on correlation between twitter concentration and traffic surges. *Transportation Research Record*, 2553(1):90–98, 2016.

Zhenhua Zhang, Qing He, Jing Gao, and Ming Ni. A deep learning approach for detecting traffic accidents from social media data. *Transportation research part C: emerging technologies*, 86:580–596, 2018. ISSN 0968-090X. doi: https://doi.org/10.1016/j.trc.2017.11.027. URL `http://www.sciencedirect.com/science/article/pii/S0968090X1730356X`.

Jinbao Zhao, Wei Deng, and Yan Song. Ridership and effectiveness of bikesharing: The effects of urban features and system characteristics on daily use and turnover rate of public bikes in china. *Transport Policy*, 35:253 – 264, 2014. ISSN 0967-070X. doi: https://doi.org/10.1016/j.tranpol.2014.06.008.

Liang Zhao, Feng Chen, Chang-Tien Lu, and Naren Ramakrishnan. Spatiotemporal event forecasting in social media. In *Proceedings of the 2015 SIAM international conference on data mining*, pages 963–971. SIAM, 2015.

Zhan Zhao, Haris N. Koutsopoulos, and Jinhua Zhao. Individual mobility prediction using transit smart card data. *Transportation Research Part C: Emerging Technologies*, 89:19 – 34, 2018. ISSN 0968-090X. doi: https://doi.org/10.1016/j.trc.2018.01.022. URL `http://www.sciencedirect.com/science/article/pii/S0968090X18300676`.

Elena Zheleva and Lise Getoor. To Join or Not to Join: The Illusion of Privacy in Social Networks with Mixed Public and Private User Profiles. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 531–540, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-487-4. doi: 10.1145/1526709.1526781. URL `http://doi.acm.org/10.1145/1526709.1526781`.

Xinhu Zheng, Wei Chen, Pu Wang, Dayong Shen, Songhang Chen, Xiao Wang, Qingpeng Zhang, and Liuqing Yang. Big data for social transportation. *IEEE Transactions on Intelligent Transportation Systems*, 17(3):620–630, 2016.

Xingang Zhou, Anthony GO Yeh, and Yang Yue. Spatial variation of self-containment and jobs-housing balance in shenzhen using cellphone big data. *Journal of Transport Geography*, 68:102–108, 2018.

Xuesong Zhou and George F. List. An information-theoretic sensor location model for traffic origin-destination demand estimation applications. *Transportation Science*, 44 (2):254–273, 2010. doi: 10.1287/trsc.1100.0319. URL `https://doi.org/10.1287/trsc.1100.0319`.

Xuesong Zhou and H. S. Mahmassani. Dynamic origin-destination demand estimation using automatic vehicle identification data. *IEEE Transactions on Intelligent Transportation Systems*, 7(1):105–114, March 2006. ISSN 1524-9050. doi: 10.1109/TITS.2006.869629.

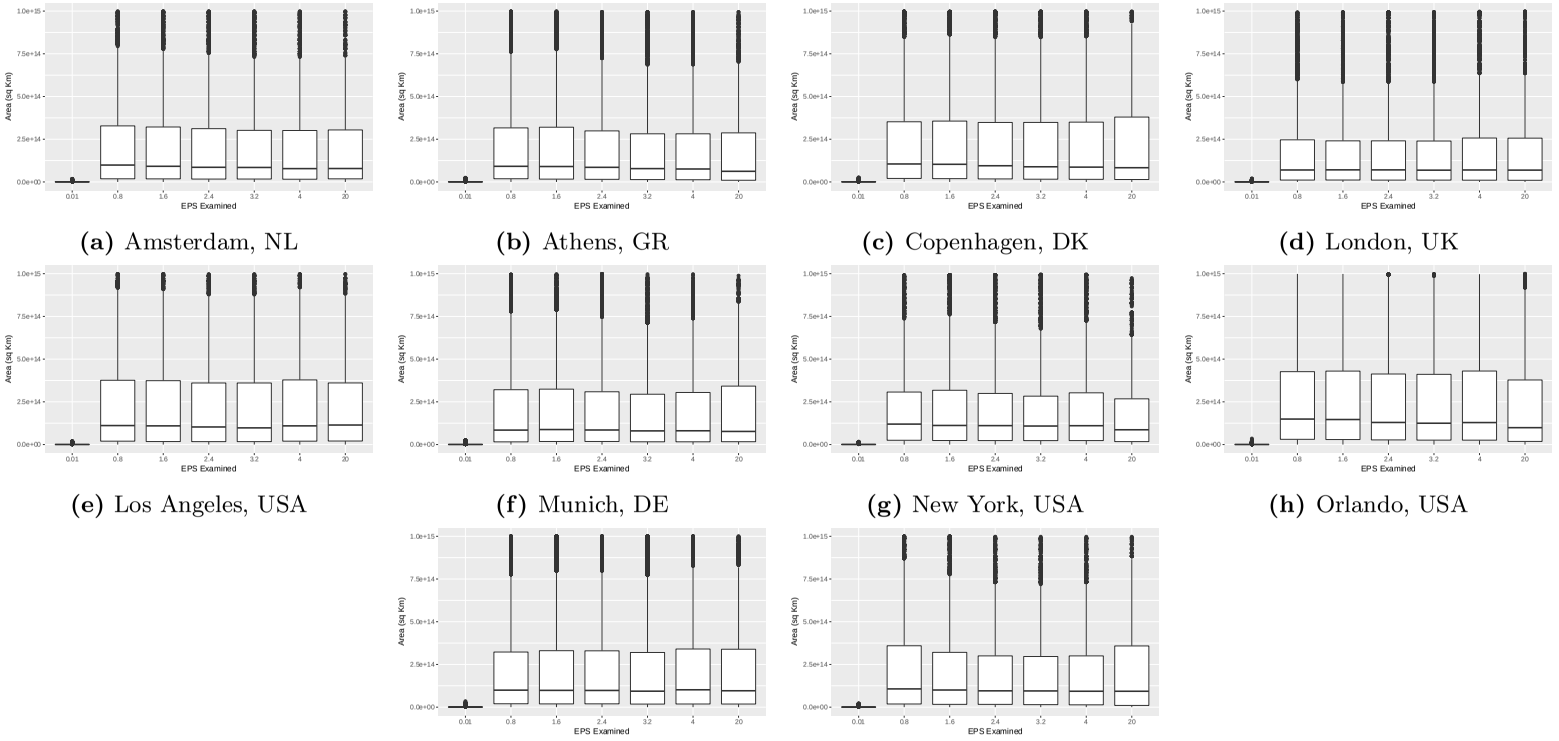
A Appendix on Sensitivity Analysis for activity space exploration

Contents

Components of this chapter are presented in:

E. Chaniotakis and C. Antoniou. On the activity space derived social media: Recurrence, temporal and spatial sensitivity analysis. In *6th hEART Symposium (European Association for Research in Transportation)*, Technion, Haifa, 12-14 September, 2017, 2017

A.1 Activity Space Area, Spatial Sensitivity



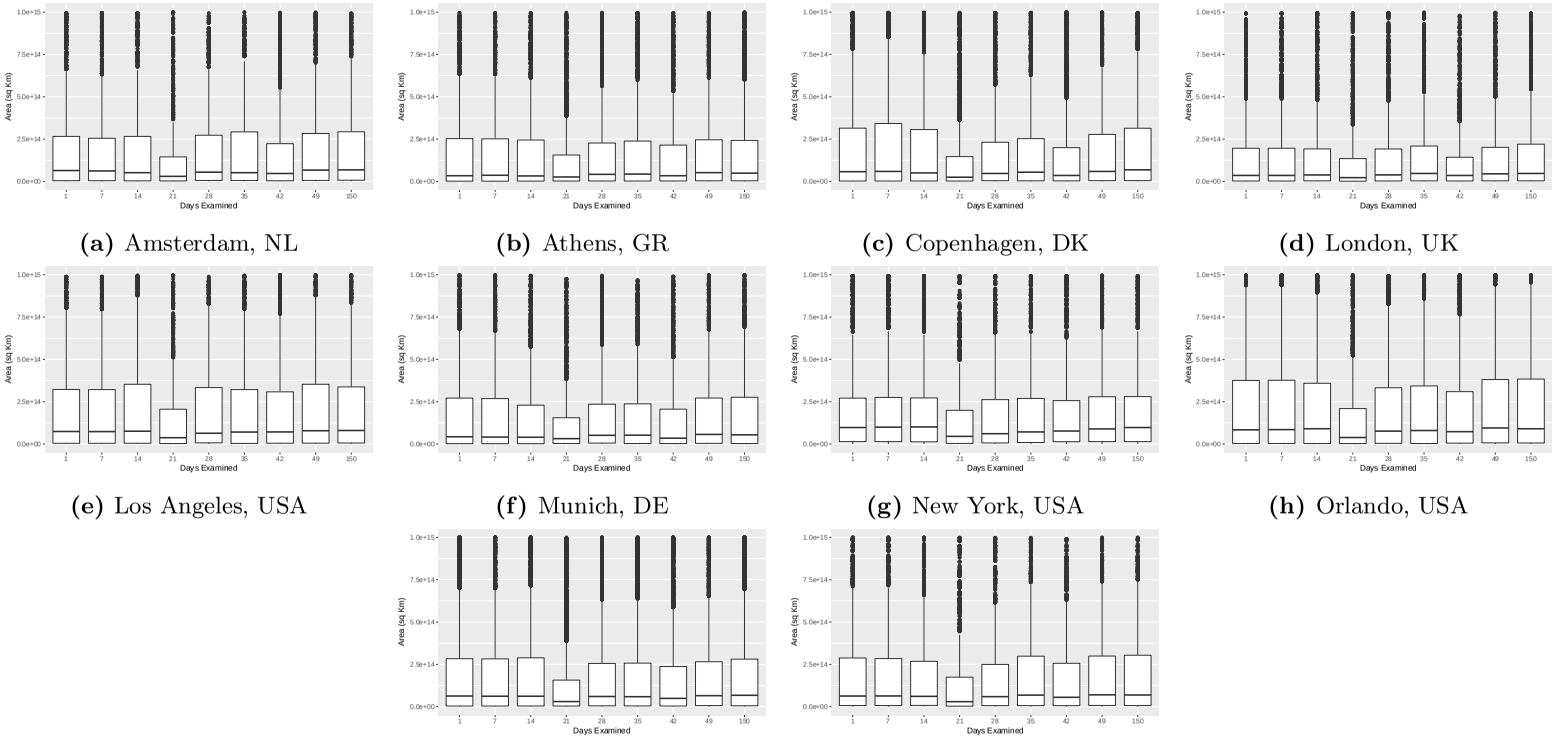
(a) Amsterdam, NL **(b)** Athens, GR **(c)** Copenhagen, DK **(d)** London, UK

(e) Los Angeles, USA **(f)** Munich, DE **(g)** New York, USA **(h)** Orlando, USA

(i) Paris, FR **(j)** Seattle, USA

Figure A.1: Activity Space Area, Spatial Sensitivity

A.2 Activity Space Area, Temporal Sensitivity



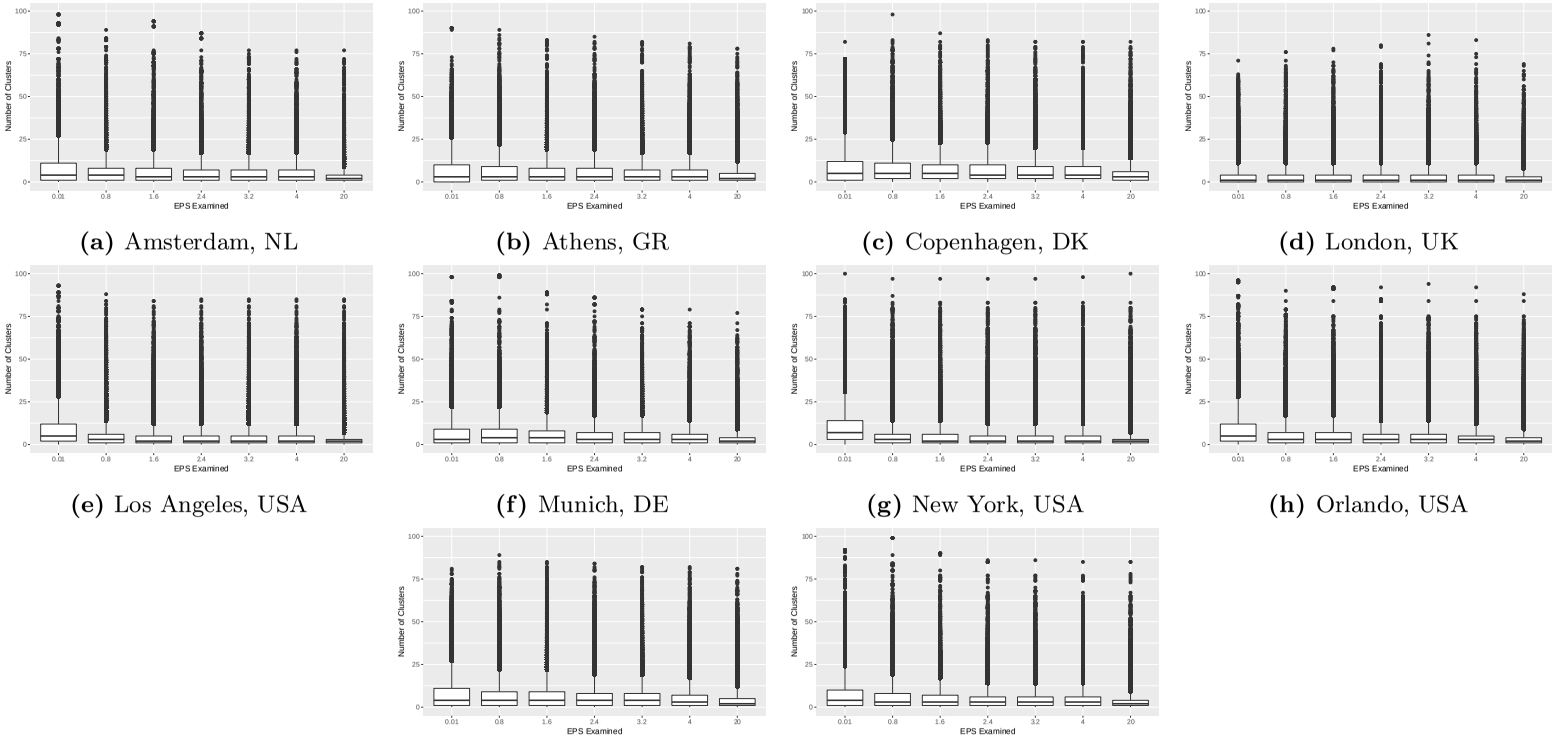
(a) Amsterdam, NL **(b)** Athens, GR **(c)** Copenhagen, DK **(d)** London, UK

(e) Los Angeles, USA **(f)** Munich, DE **(g)** New York, USA **(h)** Orlando, USA

(i) Paris, FR **(j)** Seattle, USA

Figure A.2: Activity Space Area, Temporal Sensitivity

A.3 Number of Clusters, Spatial Sensitivity



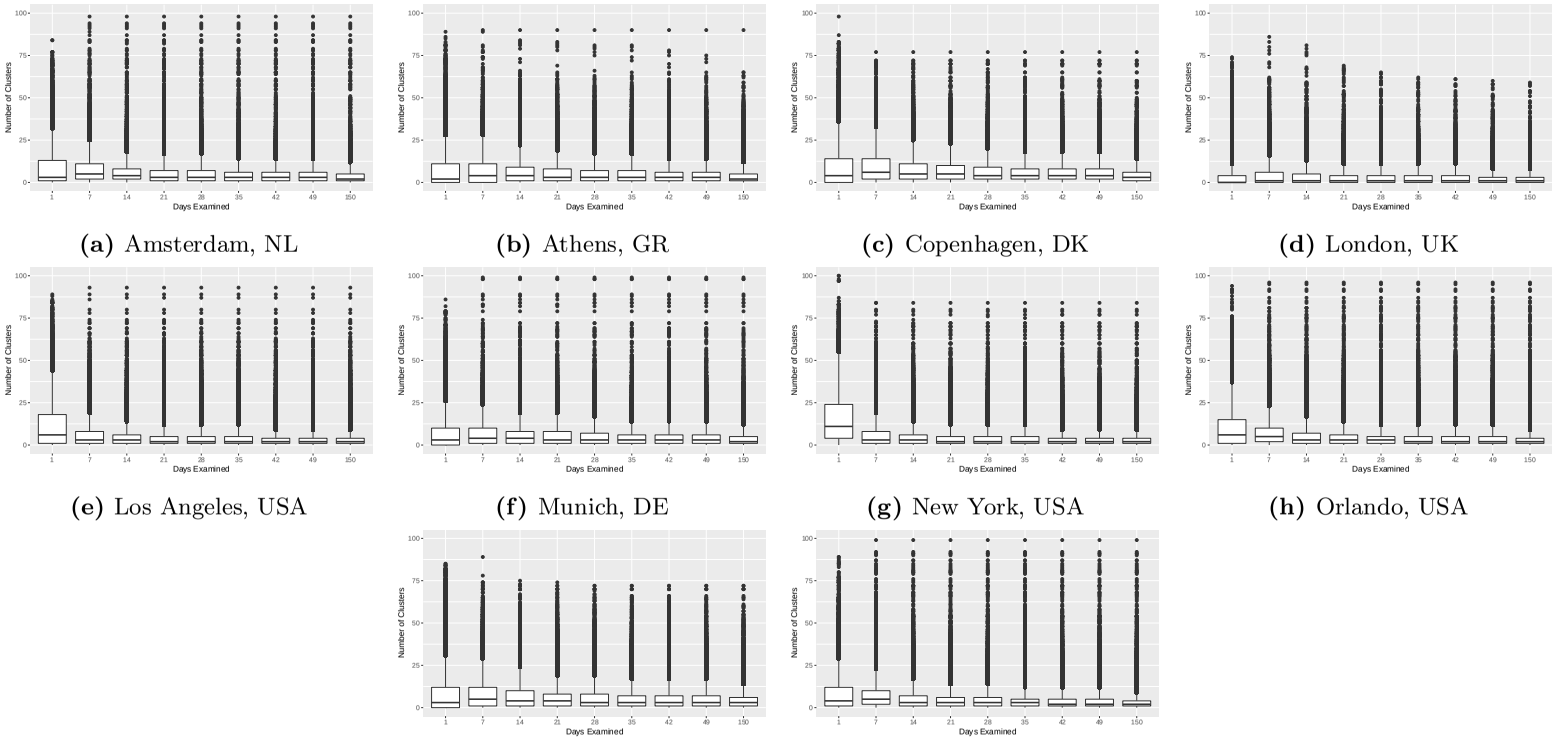
(a) Amsterdam, NL **(b)** Athens, GR **(c)** Copenhagen, DK **(d)** London, UK

(e) Los Angeles, USA **(f)** Munich, DE **(g)** New York, USA **(h)** Orlando, USA

(i) Paris, FR **(j)** Seattle, USA

Figure A.3: Number of Clusters, Spatial Sensitivity

A.4 Number of Clusters, Temporal Sensitivity



(a) Amsterdam, NL **(b)** Athens, GR **(c)** Copenhagen, DK **(d)** London, UK

(e) Los Angeles, USA **(f)** Munich, DE **(g)** New York, USA **(h)** Orlando, USA

(i) Paris, FR **(j)** Seattle, USA

Figure A.4: Number of Clusters, Temporal Sensitivity