NOVA IMS Information Management School

DOCTORAL PROGRAMME

Information Management

Specialization in Geographic Information Systems

A statistical approach for studying urban human dynamics

Luis Fernando Santa Guzmán

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor in Information Management

August, 2018

NOVA Information Management School
Universidade NOVA de Lisboa

GEO-C
Joint Doctorate in Geoinformatics: Enabling Open Cities

Prof. Doctor Roberto Henriques, Supervisor

Prof. Doctor Joaquín Torres-Sospedra, Cosupervisor

Prof. Doctor Edzer Pebesma, Cosupervisor

A statistical approach for studying urban human dynamics

To my mom...

Acknowledgements

With these words, I want to express my sincere appreciation to all those who have accompanied me during this period. I am infinitely grateful for the shared moments, the calls of support, and above all for giving me their words of encouragement when I felt that I could not follow. In random order: *Adrian, mi Reyes, Anita, Albert (mi escudero), Marekcito (don Cabrito), Marthica, Dianita, Blanquis, mi Paez, Calderoncho, Diego, John, Juli, Felix, Dave...*

This dissertation is funded by the European Commission within the Marie Skłodowska–Curie Actions (ITN–EJD). Grant Agreement Number 642332 – GEO–C – H2020–MSCA–ITN–2014.

“What we think, we become”

Abstract

This doctoral dissertation proposed several statistical approaches to analyse urban dynamics with aiming to provide tools for decision making processes and urban studies. It assumed that human activity and human mobility compose urban dynamics. Initially, it studied geolocated social media data and considered them as a proxy for where and when people carry out what it is defined as the human activity. It employed techniques associated with generalised linear models, functional data analysis, hierarchical clustering, and epidemic data, to explain the spatio-temporal distribution of the places where people interact with their social networks. Afterwards, to understand the mobility in urban environments, data coming from an underground railway system were used. The information was considered repeated daily measurements to capture the regularity of human behaviour. By implementing methods from functional principal components data analysis and hierarchical clustering, it was possible to describe the system and identify human mobility patterns.

Resumo

Esta tese de doutorado propôs várias abordagens estatísticas para analisar a dinâmica urbana com o objetivo de fornecer ferramentas para processos decisórios e estudos urbanos. Assumiu que a atividade humana e a mobilidade humana compõem a dinâmica urbana. Inicialmente, ele estudou dados de mídias sociais geolocalizadas e os considerou como uma proxy para onde e quando as pessoas realizam o que é definido como a atividade humana. Empregou técnicas associadas a modelos lineares generalizados, análise de dados funcionais, agrupamento hierárquico e dados epidemiológicos, para explicar a distribuição espaço-temporal dos lugares onde as pessoas interagem com suas redes sociais. Posteriormente, para entender a mobilidade em ambientes urbanos, foram utilizados dados provenientes de um sistema ferroviário subterrâneo. As informações foram consideradas medidas diárias repetidas para capturar a regularidade do comportamento humano. Através da implementação de métodos a partir de análise de dados de componentes principais funcionais e clustering hierárquico, foi possível descrever o sistema e identificar padrões de mobilidade humana.

Contents

List of Figures

List of Tables

Acronyms

API Application programming interface

BIC Bayesian information criterion

CDR Call Detail Records

CRS Coordinate reference system

DBSCAN Density-based spatial clustering of Applications with Noise

ERGM Exponential Random Graph Models

ESF Eigenvector Spatial Filtering

FDA Functional Data Analysis

FPCA Functional principal component analysis

GAM Generalised additive models

GLM Generalised linear models

GMM Gaussian mixture models

GNSS Global Navigation Satellite System

HMM Hidden Markov Models

ICA Independent Component Analysis

ICT Information and Communication Technologies

ITS Intelligent Transportation Systems

IWLS Iteratively weighted least squares

KDE Kernel density estimation

LBSN Location–Based Social Networks

LDA Latent Dirichlet allocation

LISA Local indicators of spatial association

MAUP Modifiable area unit problem

MCMC Markov chain Monte Carlo

OLS Density-based spatial clustering of Applications with Noise

PCA Principal components analysis

SAR Spatial Autoregressive Models

SEM Spatial Error Models

SID Spatial Interaction Data

SNA Social Networks Analysis

SOM Self-organising maps

STSS Space-time scan statistics

Introduction

1

1.1 Overview

Geospatial research comprises substantial efforts in studying the dynamics of cities (França et al., 2015; Celikten et al., 2017; Jiang et al., 2012). In general terms, humans exhibit regularity in their spatial, temporal, and social behaviours (Simini et al., 2012; Song et al., 2010b; González et al., 2008; Brockmann et al., 2006); nevertheless, large-scale urban social systems are complex and challenging to model or represent (Batty, 2009; Jackson, 1985). The accelerated urbanisation process of the current society and the forecasts for a significant increase in urban populations are expected to enhance this complexity (United Nations, 2014). To predict these systems requires a mathematical description of the patterns found in city data, forming the basis of the models that can be used to anticipate trends, assess risks, and manage future events (Vespignani, 2009). The lack of data in this context has historically been a substantial problem (Thériault and Des Rosiers, 2013); however, the increase in the availability of crowdsourced data over the last decade, gives a rich and real-time data source of detailed images of urban systems (Jiang et al., 2012; Vespignani, 2009).

Previous studies about human dynamics have focused mainly on two directions: (1) in the branch of complex systems in statistical physics highlighting specific aspects such as dimensions and mechanistic models and (2) on the use of survey sampling techniques to record data of the users' behaviour, i.e., origins and destinations (Hyman, 1969; Beckmann, 1967). These studies have notably omitted larger explorations and insights into new methods for discovering patterns using data coming from new and massive sources in the context of mobile and big data era (Shaw et al., 2016). Currently, ubiquitous computing has permitted collecting a large amount of data shared by people about themselves (Kaplan and Haenlein, 2010) and their interaction with the physical world (Nummi, 2017). Those datasets are far from conventional in the sense of tabular or structured data, and data processing has not analysed a significant amount of them because of the computational expense and the need for specific data analysis techniques (Gandomi and Haider, 2015). In addition, such information remains sparse in the geographical space, is incomplete in a time interval (Huang, 2016; Gao and Liu, 2014; Ferrari et al., 2011), and might not be representative (Toole et al., 2015); still, this informations is considered a complementary alternative to the gathered information through survey sampling techniques for analysing human dynamics since it captures people's perceptions and spatio-temporal changes more accurately (França et al., 2015; Frias-Martinez et al., 2012; Wakamiya et al., 2011).

In recent years, the concept of smart cities has been extensively studied to address the development of methodologies based on the use of Information and Communication Technologies (ICT) to improve the citizens' quality of life in a sustainable development framework (Steenbruggen et al., 2015; Pan et al., 2013a; Bakci et al., 2012; Chourabi et al., 2012). However, it is common that citizens have a negative perception regarding the dynamics of the cities

(Shittu et al., 2015; Enemark and Kneeshaw, 2013; Pressl and Köllinger, 2012). In this sense, Buscher et al. (2014) mention that a constraint for the cities is related to the population growth because it raises the demands for efficient flows of people in an environment with limited physical infrastructure and requires for adequate management of systems to avoid congested and unpleasant situations for the users. Thus, as a result of the emergence of the Internet, the urban sensors, the smart devices, the wireless networks, and the development of online social networks have made possible the storage of significant amounts of data broadening the spectrum of modelling possibilities for understanding urban issues.

Therefore, the use of such data is allowing the development of particular analytical methods, which characterise the structure of the cities by identifying, describing, and predicting similar behaviours on the conducts of the people, in a branch of knowledge called *urban analytics* or *urban informatics* (Zheng et al., 2014). These methods seek to explain when, where, and why humans develop their activities in the cities. Thus, provide meaningful insights into the spatio-temporal patterns of human activity and mobility.

In this sense, statistical modelling—and mainly spatio-temporal statistics—is an alternative approach to study urban dynamics because it provides, by estimating the parameters of the models, a way to explain the processes that generate the data Diggle (2013) and can be useful in monitoring, comparing, and simulating urban environments more reliably. From a statistical point of view, the selection of data analysis methods requires to define among others: (1) the nature and the source of the studied information, (2) the characteristics of the analysed phenomenon, (3) the sampling mechanism, and (4) the scope of the expected results.

Concerning this, as mentioned earlier, humans exhibit a high degree of spatio-temporal regularity. However, the exact place and time of where and when people carry out their activities can neither be fixed nor established by some sampling mechanism. Likewise, urban dynamics can present sporadic, sudden happenings, such as massive events and traffic jams, etc. On the other hand, the considerable advances of computational and analytical techniques have allowed many processes to be continuously monitored, and their immediate consequence is the augment of the amount of data to be analysed that demands for developing new statistical methods (Chen and Müller, 2012; Martínez-Camblor and Corral, 2011).

For example, the analysis of geolocated social media data encompasses the study of the number of events per area per hour. Those counts show strong temporal trends due to the regularity of human behaviour. Likewise, in the context of public health, where is registered the number of cases of a particular disease, those time series exhibit seasonality and occasional outbreaks Paul et al. (2008). These data are called epidemic data and are conceived as realisations of spatio-temporal processes with autoregressive behaviour which do not come from planned experiments. Its observations, number of events, are not independent, and phenomena are only partially observed Meyer et al. (2017). Thus, there is a high similarity between the number of geolocated social media events and the counts of cases in public health studies. Hence, an approach based on the statistical modelling of epidemic data can accommodate the presence of abnormal events in urban dynamics and even be capable of predicting them. Besides, those models include autoregressive trends and spatio-temporal structures in the estimation of the parameters that describe human social conducts more accurately.

On the other hand, daily flows of people in urban public transport systems is a continuous process that under the scope of Intelligent Transportation Systems (ITS) is observed through sensors at entries and exits of undergrounds, buses, and botes, etc. These systems usually register the time when a user enter and exit to the network (Zhang et al., 2011). Such data are repeatedly observed for each sensor along a set of locations. In this regard, Functional Data Analysis (FDA) methodology that aims to study random functions of time-dynamic processes (Ramsay and Silverman, 2005; Chen and Müller, 2012; Kokoszka and Reimherr, 2017) can be useful to describe and forecast mobility behaviours.

1.2 Problem statement

Urban dynamics issues are becoming one of the most frequent in the cities. Different alternatives have appeared to manage aspects related to human activity and mobility, such as the investment in improving the physical infrastructure, the instruments to define policies regarding the provision and administration of public services, and urban planning processes. These three latter options are closely related to the development of models to study spatio-temporal patterns of human conducts because that analysis can be meaningful to identify how urban spaces are used, therefore, helping in decision-making processes. In this sense, previous works about urban dynamics were mainly based on the use of survey sampling techniques to gather data about people's behaviour. Nowadays, it is more common to have a significant amount of data regarding how people interact with their cities coming from new sources such as sensors, smart devices, and social media. This recently collected information has allowed developing new approaches in data analysis. However, these data have not been explored yet in-depth, and there is still room for proposing, implementing and evaluating alternatives of data analysis to identify, describe, explain, and predict patterns of human activity and mobility using statistical methods.

1.3 Scope, objectives and research questions

1.3.1 Scope

This dissertation aims to develop methods of data analysis to understand the dynamics of the cities and provide insightsinto human activity and mobility, for urban planning and decision-making processes in the scope of smart cities. Assuming that ICT allow collecting data about city environments and human behaviour by accessing ubiquitous devices such as mobile phones and sensors.

Specifically, the scope of this research is framed at the city level, i.e. the information used, and its scale is limited to the urban environments analysed. However, the methodological approaches proposed can be replicated in cities of several sizes and provide insights to compare places with different characteristics.

1.3.2 Objectives

Based on the previously presented scope, the current research has three primary objectives, as follows:

- **Objective 1** to use models to describe, identify, quantify, and predict the impact of the factors associated with the regularity of the human conducts in the number of human activities in urban spaces.
- **Objective 2** To evaluate the goodness of fit of including spatio-temporal correlation structures in statistical methods to analyse human activity and mobility.
- **Objective 3**To develop alternatives to characterise the spatio-temporal flows of users in origin-destination systems.

1.3.3 Research Questions

To address the prior objectives, the main overarching research question proposed as part of this dissertation is:

What are the new analytical strategies which became applicable with recently increased available data sources to study different spatio-temporal aspect of urban inhabitants behaviour?

As a way to answer the above main question, three sub-guiding research questions were proposed as follows:

- **RQ 1** To what extent factors as the hour of the day, the day of the week, and autoregressive trends are related to the human activities in urban environments?
- **RQ 2** Which type of spatio-temporal structures can improve the goodness of fit in models for studying human activity and mobility?
- **RQ 3** Is it possible to identify spatio-temporal communities of stations in origin-destination systems associated with underground railways?

1.4 Methods overview

In general terms, to answer the research questions, a process that involved four stages was developed (*see* Figure 1.1).

Data collection. It implied to harvest information from two sources: (1) social media and smart cards with entries and exits registries into an underground railway system. For example, social media data was downloaded from Lisbon, London, and Manhattan using the R software and connecting the Twitter Application programming interface (API) . Additionally, the registries of entries and exits of smart cards for the Lisbon underground railway in may 2015 was acquired through private agreements for collaboration with the company in charge of the management of the system.

Data pre-processing. This stage required the implementation of particular procedures to identify and remove erroneous values from the data collected on the first stage, as well as to aggregate data, and bring it in the appropriated standards for following stage, data analysis.

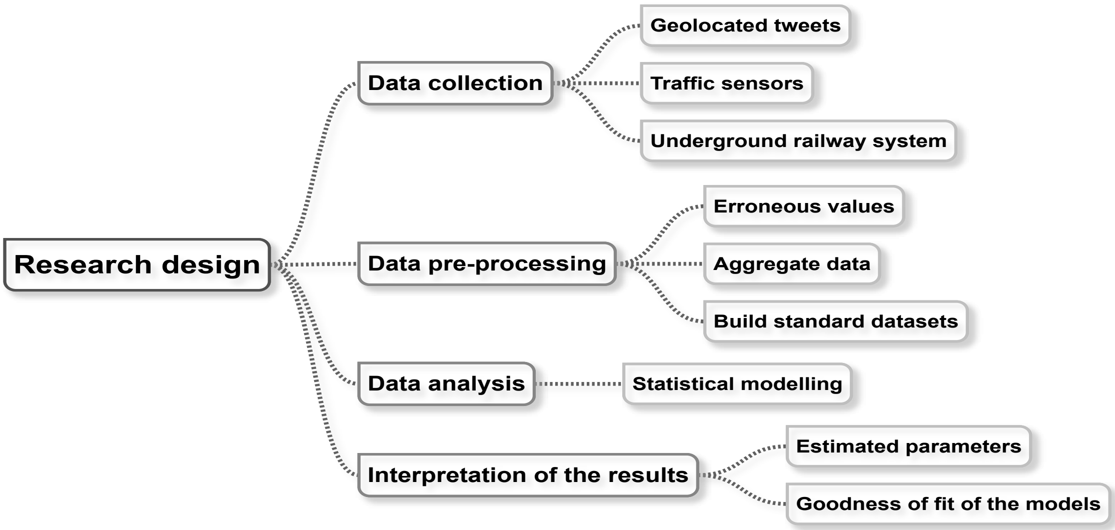


Figure 1.1: Schema of the research design.

Data analysis and interpretation of results. This thesis recurred to statistical modelling since its methods provide elements, parameters of the models, to understand and explain underlying processes that generate the data and can replicate or simulate complex systems. These methods can be useful in monitoring urban dynamics more reliably. Particularly, the following modelling approaches were considered: generalised linear models (GLM), functional principal components analysis (FPCA), statistical analysis of spatial point patterns, hierarchical clustering, spatio-temporal graph theory, and infectious disease surveillance models (*see* Figure 1.2). In most of the cases, the observations were considered as daily repeated measurements and opted for non-parametric estimation to avoid strong distributional assumptions when possible.

For social media data, geolocated tweets were downloaded as a proxy for human activity in urban environments. The analysis was divided into two parts. First, by estimating regression models under the scope of the GLMs to explain the number of geolocated tweets per hour in a city as a function of the hour-of-the-day, the day-of-the-week, and autoregressive trends. Second, by clustering hours of the day with similar patterns of spatial arrangement of the places where people interact with their social networks.

Furthermore, an endemic-epidemic model, was estimated assuming a negative binomial distribution for the counts and including seasonal effects as *normal* or *endemic* human behaviour

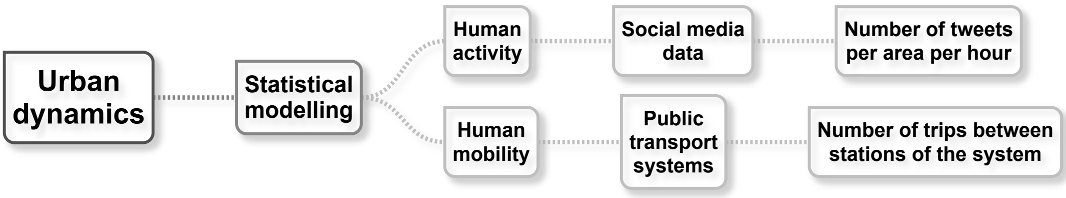


Figure 1.2: Schema of statistical modelling.

and spatio-temporal autoregressive parameters for *epidemic* or *abnormal* situations as crowded events.

Finally, for modelling the spatio-temporal directed graph that represents the flows origin-destination within an underground railway system, daily time series for every pair of stations, with the number of trips every 15 minutes starting in one of them and finishing in the other, were considered. In this case, the method involved a mix between two steps FPCA and hierarchical clustering to summarise daily behaviours and to describe the activity over the entire graph.

1.5 Contributions

The overall results and insights from this research can help on providing efficiently performed and replicable methods for analysing significant amounts of urban data in smart cities. It advances on this by using more advanced statistical techniques which identify several spatio-temporal characteristics of the dynamics of urban systems on a more straightforward manner than alternatives developed in previous studies. These methods include and statistically test the effect of considering spatio-temporal autocorrelation structures in models and predictions.

The proposed methods allow to predict, monitor and simulate human activity and mobility in cities in a more accurate way by introducing associated effects with the regularity of the human behaviour into the previously existing spatio-temporal models. All of the suggested approaches admit the inclusion of data as soon as additional information is available, to potentially improve the goodness of fit of models, anticipate changes in human behaviour in near real-time, and to refine the precision in pattern discovery.

1.6 Thesis Outline

This doctoral dissertation is organised into five chapters, as follows:

- The current chapter 1 gives a general overview of the origin of the research, states the problem, defines the objectives that are being persuit, and summarises the approach and methods.

- Chapter 2 presents a literature review regarding human mobility and urban human mobility. It contains a summary of the main findings in this field from the perspective of the statistical physics and the mechanistic models. Finally, it concludes with an exposition of the recent trends in the analysis of mobility in the urban scenario and the role of the data analysis to tackle its study.

- Chapter 3 is devoted to provide the statistical framework of the utilised methods for analysing urban human dynamics. It included adaptations done on previous statistical modelling methods as well as the novel approaches developed as part of the current dissertation.

- Chapter 4 introduces and details a statistical approach for the study of the spatio-temporal distribution of geolocated tweets in the cities. This part gives and justifies the statistical details of the selected methods, as well as, evaluates the proposal in three different urban scenarios.

- Chapter 5 is dedicated to present the methods and insights gained from analysing human-generated social media data in cities by using tools of epidemic data and assuming a model-based approach developed under the scope of the diseases surveillance systems. In this regard, spatio-temporal models were estimated that describe simultaneously normal and unusual events.

- Chapter 6 presents the results of the analysis of Lisbon underground railway system.

- Finally, chapter 7 outlines the main findings of this dissertation and answers the research questions. Additionally, it discusses aspects related to further research.

Urban Human Mobility 2

Urban human mobility has always been a highly relevant issue for human settlements, particularly for middle and large-scale ones. It serves to study and understand these in diverse areas and scales. However, it becomes even more crucial for urban studies nowadays to constantly track and address the possibility of city transportation systems becoming insufficient. Two reasons may be responsible for this: 1) the fast growth of population and; 2) the complexity of building transportation infrastructure to address it at the same rate. By 2014, around 54% of people live in urban areas, an increase of 12% is expected by 2050 according to forecasts (United Nations, 2014), and the previous situation is predicted to triple the travelled distances in cities in the upcoming 40 years (Van Audenhove et al., 2014). These elements and conditions, as stated by Buscher et al. (2014) can produce a constraint for cities and their transportation systems due to an increase in population demand for efficient flows of people in an environment with limited physical infrastructure capacity. Cities that have poor results in traffic congestion such as Brussels, Los Angeles, Milan, London Paris, and Mexico City, among others, can benefit from the results of a detailed urban human mobility research and its use in policy-making and planning strategies (Kirkpatrick, 2015; Cox, 2014; Gorzelany, 2013).

2.1 Human mobility

Mobility has been a central topic in many investigations since there has been scientific interest in understanding how objects, animals, and people move. The initial works in this matter were related to Robert Brown's findings on the movement of particles through a fluid, which has led to the use of stochastic models such as Brownian motion, random walks, and Lévy flights to describe such displacements (Giannotti et al., 2013). However, the social nature of humans directs the analysis of mobility for understanding social conducts such as grouping, access to goods and services, and exchange of information (Toole et al., 2015). *Human mobility* is defined as "***when** and **where** a user (**who**) has been to **for what**.*" It reflects the mobile aspect of people behaviour in the real world and is commonly treated as a stochastic process (Gao and Liu, 2015).

The initial modern attempts to study human movement using empirical observation appeared in the work of Ravenstein (1885), which through an analysis of census data of the United Kingdom, found significant regularities in the populations' motions. Recently, studies related to human mobility have raised a particular interest due to the availability of data and to the relevance of the topic in several domains (*see* Table 2.1). Although there is not a unified schema for conducting studies on human mobility, Karamshuk et al. (2011) and Pan et al. (2013a) suggest that such a process should involve among others three stages. First, collecting real-life data from traces of moving objects. Second, developing methods of analysis to get knowledge about mobility such as statistical properties, patterns, and models. Third, creating applications that allow using the knowledge acquired in several fields.

Table 2.1: Applications of previous studies in human mobility

Field	Authors
Management and planning of urban and transport facilities and services	Cats et al. (2015); Chua et al. (2015); De Domenico et al. (2015); Rebelo et al. (2015); Anbaroglu et al. (2014); Chen et al. (2014); Louail et al. (2014); Nanni et al. (2014); Ren et al. (2014); Sun et al. (2014); Liang et al. (2013); Castro et al. (2012); Hasan et al. (2012); Yuan et al. (2012); Yuan and Raubal (2012); Noulas et al. (2011); Phithakkitnukoon et al. (2010)
Predict and prevent disease outbreaks	Wesolowski et al. (2012); Colizza et al. (2006)
Behaviour modelling	Bettencourt (2013); Pan et al. (2013b); Bagrow and Lin (2012)
Migration trends	Hawelka et al. (2014); Vaca-Ruiz et al. (2014); Brockmann (2012); Murgante and Borruso (2012); Simini et al. (2012); Rae (2009)
Management and optimization of networks	Coscia et al. (2014); Karamshuk et al. (2014); Pirozmand et al. (2014); Zhao et al. (2014); Szell et al. (2012); Karamshuk et al. (2011)
Disasters, catastrophes and preparation of big events	Wachowicz and Liu (2016); Pinheiro (2014); Sagl et al. (2012)

From the perspective of opportunistic networks,(Karamshuk et al., 2011) proposes a framework for studying human mobility, from data to models and it includes the collection and analysis of the traces. Based on results from previous studies, these authors stated that findings could be classified into three axes: spatial, temporal, and social whose components and statistical properties are summarised. Besides, they infer that the predictability or regularity of the movements of the individuals does not describe all their aspects, so they propose a more general concept called "*human mobility patterns*." Pirozmand et al. (2014) update the content put forward by Karamshuk et al. (2011). They give more detail in the description of each axis and its components and include a brief exposition about the prediction of human mobility. The paper exposes an alternative proposal for classifying mobility models. Diab and Mitschele-Thiel (2014) make a detailed presentation, discussion, and qualitative comparison of mathematical models to identify patterns of movement in fields such as mobile communication and urban planning, among others.

Statistical physics have also devoted efforts to describe the general aspects of people's movements. In this sense, Giannotti et al. (2013) expose individual human mobility models from an approach from the theory of stochastic processes and complex networks. They envisioned the convergence of data mining research and network science research to increase the accuracy of results in mobility studies.

Toole et al. (2015) make a presentation of concepts, data sources, models, and applications about the movement of individuals. The review highlights the social nature of travel undertaken by people and how this defines the functioning of societies. The paper mentions the changes produced in modelling by the availability of large, high-resolution data sets collected due to the emergence of ubiquitous computing. Besides, they emphasise that despite the volume of information, this is not enough to decrease the bias on data because they might not be representative of the population, being necessary to combine modern techniques of analysis with sampling and robust statistical methods to improve the understanding of the phenomenon. In that same sense, Noulas (2013) makes a complete revision from a historical perspective of human movement studies until the modelling in urban spaces. It details changes in mobility analyses due to location-based social networks and ubiquitous devices.

2.1.1 Dimensions

Three dimensions, spatial, temporal, and social or connectivity characterises human mobility. *Spatial* or *geographical features* refer to location information of mobile users and their trajectories in physical space and rely on distance travelled. *Temporal properties* explore time-varying structures of human mobility such as the times a user visits some specific locations. Meanwhile, *connectivity information* examines contact and interaction patterns of people, which are related to social relationships and similarities between them (Karamshuk et al., 2014; Pirozmand et al., 2014; Karamshuk et al., 2011). Previous studies show that human mobility exhibit a high spatio-temporal regularity. For example, probability distributions such as power-law and truncated power-law with exponential cutoff are associated with the spatial and temporal dimensions, respectively. This latter means that people tend to travel the same distance (relatively 1.5 km) in approximately similar periods of time (24h, 48h, 72h). Even, it has been identified that people do not travel long distances (Simini et al., 2012; Song et al., 2010b; González et al., 2008; Brockmann et al., 2006). Each individual is characterised by having a significant probability to return to a few highly frequented places (Gao and Liu, 2015; Nanni et al., 2014; Pinheiro, 2014; Lu et al., 2013; Wang et al., 2011; Song et al., 2010a,b; González et al., 2008). Moreover, there is a strong correlation in daily activity patterns among people who share a common work area's profile (Phithakkitnukoon et al., 2010). A detailed description of the components of each dimension is presented in Table 2.2.

2.1.2 Aggregation levels

Mobility studies are usually conducted in two aggregation levels, *small* or *individual* and *large*, *collective*, or *aggregate* (Sun et al., 2010). The analysis of *small-scale human mobility*, namely, the movement trajectory of only one person attempts to explain the underlying patterns of individuals using new high-resolution data with information of times, places, and semantic attributes about how and why human beings travel between them. This analysis seeks to provide insights into the nature of people behaviour by developing models in statistical physics, such as random walks (Toole et al., 2015; Giannotti et al., 2013). On the other hand, *large-scale human mobility*, i.e., the overall movement behaviour of large crowds, also called "*urban dynamics*" is used for providing services to all citizens, such as public transportation, as well as planning city

Table 2.2: Characteristics of human mobility

Dimension	***Component***
Spatial	***Travel distance*** or ***jump size*** (Δr) is an important feature of human walks to characterize the spatial dimension of people. It has been described using: a) power-law (Brockmann et al., 2006), $P(\Delta r) \sim \Delta r^{-(1+\beta)}$ where $\beta < 2$ and b) truncated power-law with an exponential cutoff (González et al., 2008), $P(\Delta r) = (\Delta r + \Delta r_0)^{-\beta} \exp(-\Delta r/\kappa)$ where $\beta = 1.65 \mp 0.15$, $\Delta r_0 = 1.5$km, and κ a cutoff value varying in different experiments.
	Radius of gyration (r_g) is a measurement of the characteristic distance travelled by an individual during an observation period t. González et al. (2008) determined of the r_g distribution like a truncated power-law with an exponential cutoff, i.e., $P(r_g) = \left(r_g + r_g^0\right)^{-\beta} \exp(-r_g/\kappa)$ with $r_g^0 = 5.8$km, $\beta = 1.65 \mp 0.15$ and $\kappa = 350$km
Temporal	***Return time*** (t) is the period of time in which a random walker returns regularly to the same location visited previously. González et al. (2008) have found that return probability has peaks at 24h, 48h and 72h.
	Pause time (Δt) indicates the time period that a person stays in a specific position. The probability pause time distribution has been found: a) as a fat-tailed (González et al., 2008; Brockmann et al., 2006): $P(\Delta t) \sim (\Delta t)^{-(1+\beta)}$ with $0 < \beta \leq 1$ or b) according to Song et al. (2010a) like $P(\Delta t) \sim \lvert\Delta t\rvert^{-(1+\beta)}$ with $\beta = 0.8 \mp 0.1$ and a cutoff of $\Delta t = 17$h
Social	***Contact time*** is the time intervals during two people are in the radio range of one another.
	Inter-contact time is the amount of time passed among two consecutive contact periods for a given couple of people.

spaces; its study relies on *origin-destination matrices* that have the number of users travelling through different locations and times by several means (Toole et al., 2015; Sun et al., 2010).

2.1.3 Spatial scales

Being mobility is an inherently spatial concept, as well geography and other spatial sciences, the notion of scale plays an essential role in its analysis (Saberi et al., 2016) since the results of the models are highly dependent on it. For example, multiple applications have studied human mobility characteristics at different spatial scales and using several data sources; this has shown that in the particular case of the patterns of the travel distance, there are differences which are explained by the mode of travel and the scale of the data (Lloyd, 2014). Studies of human mobility highlight four spatial scales: a) global or worldwide, b) national, c) regional, and d) urban.

2.1.4 Models

Moreover, the study of human mobility can be addressed from two perspectives. First, based on developing theoretical or *mechanistic models* to explain underlying behaviours of the movement, as well as natural laws that govern it. Second, through generating of techniques to learn from data or *empirical models*, with the aim to describe patterns of displacement defined by objects or individuals.

2.1.5 A theoretical framework

The analysis of human mobility encompasses four main elements which are present in almost every research that aims to understand the behaviour of movement of people. It includes *dimensions*, *aggregation levels*, *spatial scales*, and *models*. Figure 2.1 shows a theoretical framework that explains human mobility as a transversal concept in which these four aspects converge. This structure also highlights the multidisciplinary nature of the study of human displacement, which includes among others, geography, sociology, data science, and physics. For instance, modelling the characteristics (spatial, temporal, social) of movements requires the definition of spatial resolution, as well as the scope regarding aggregation.

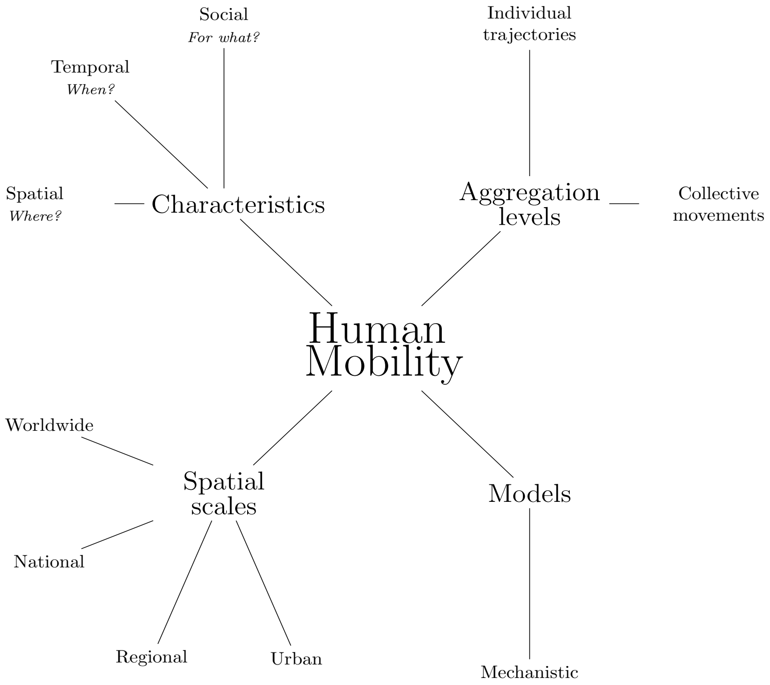


Figure 2.1: Theoretical framework

2.2 Urban human mobility and data sources

Although first studies of human mobility were related to analysing migration trends based on information coming from population censuses, these were later replaced by survey sampling techniques about the origin and destination of the citizens, especially users of transport systems. This type of data had the inconvenient of not accurately representing the changes produced by urban dynamics. Thus, recent attempts have used as a proxy of the human movement, among others, banknotes (Brockmann et al., 2006) and Call Detail Records (CDR) (González et al., 2008) trying to catch in real time the underlying nature of displacements.

Nowadays, data-acquisition methods continuously record all sort of events occurring in real life. Improvements in both hardware and software technologies allow to collect significant amounts of data, producing, every day, more complete, accurate and detailed pictures of human activity. These developments have increased the relevance of information in modern society, which in turn has led to an even higher rate of data. It is a shared idea that information is a valuable resource for any organisation.

Social media is conceived as Internet applications that allow creating, obtaining, and exchanging content created by users which can be accessed everywhere (Kaplan and Haenlein, 2010). They take in information about events and facts that occur in the real world (Ferrari et al., 2011). Thus, social media data reflects human behaviour, prompting new alternatives to understand individuals, groups, and society (Batrinca and Treleaven, 2014). When information of the geographical location is added to the social media data, this offers a source of social data that contains information about people's attitude, mobility, and feelings about places (Nummi, 2017).

In this sense, social and human researchers consider LBSN data is a crowd-data source useful for studying and understanding cities due to the frequent interaction with the ubiquitous devices by the dwellers (Silva et al., 2013; Frias-Martinez et al., 2012). Despite that this information is sparse in the geographical space, incomplete in a time interval (Ferrari et al., 2011), and might not be representative (Toole et al., 2015). It is considered a better alternative for analysing city dynamics, human activity, and urban planning, than survey sample techniques through using questionnaires because it catches people's perceptions and spatio-temporal changes more accurately in real-time (França et al., 2015; Frias-Martinez et al., 2012; Wakamiya et al., 2011).

However, social media data is not the only data source available to understand city behavioural patterns and dynamics. Some examples of these are census data, remote sensing data, traffic cameras, GNSS data, WIFI and mobile network data, e-transactions, smart card technologies, among others. Integration of all this data is both, a challenge and opportunity for improving our understanding of urban dynamics. Some of the most interesting aspects are the different spatial and temporal resolution. Typically, all these data sources share a common characteristic: they are georeferenced. Table 2.3 shows a summary of main data sources that have been used on human mobility in different contexts.

2.3 Data-driven approach

Recent attempts at studying urban human mobility have been focused on alternatives for identifying human mobility patterns. The type of data available points out the modelling approach of urban mobility, i.e., a data-driven approach. In this way, human mobility patterns are an empirical characterisation of objects' collective behaviour through data. In general terms, it is possible to classify the in three ways, thus:

Table 2.3: Data sources in human mobility studies

Data Source	Studies
Surveys	
Online	Cottrill et al. (2013)
Paper-based	Yan et al. (2013); Maat et al. (2005); Schlich and Axhausen (2003); Ewing and Cervero (2001); Vilhelmson (1999); Hanson and Huff (1988)
LBSN	
Altergeo	Karamshuk et al. (2014)
Brightkite	Chen et al. (2015); Cho et al. (2011)
Facebook	Cranshaw et al. (2010)
Foursquare	Espín-Noboa et al. (2016); Forghani and Karimipour (2014); Karamshuk et al. (2014); Nin et al. (2014); Cebelak (2013); Noulas et al. (2011)
Gowalla	Chen et al. (2015); Karamshuk et al. (2014); Nguyen and Szymanski (2012); Cho et al. (2011); Scellato et al. (2011)
Twitter	Prasetyo et al. (2016); Wachowicz and Liu (2016); Chua et al. (2015); Llorente et al. (2015); Rebelo et al. (2015); Gabrielli et al. (2014); Hawelka et al. (2014); Nin et al. (2014); Ferrari et al. (2011); Wakamiya et al. (2011)
Yahoo Meme	Vaca-Ruiz et al. (2014)
Trajectories	
GNSS on cabs	Espín-Noboa et al. (2016); Sun et al. (2014); Castro et al. (2012); Yuan et al. (2012)
GNSS on vehicles	Pappalardo et al. (2015); Coscia et al. (2014)
CDR	Wachowicz and Liu (2016); De Domenico et al. (2015); Gao (2015); Hawelka et al. (2015); Herrera-Yagüe et al. (2015); Louail et al. (2015); Pappalardo et al. (2015); Steenbruggen et al. (2015); Amini et al. (2014); Louail et al. (2014); Nanni et al. (2014); Palchykov et al. (2014); de Montjoye et al. (2013); Lu et al. (2013); Bagrow and Lin (2012); Ranjan et al. (2012); Sagl et al. (2012); Wesolowski et al. (2012); Yuan and Raubal (2012); Cho et al. (2011); Wang et al. (2011); Phithakkitnukoon et al. (2010); Sohn et al. (2006)
Ships	Demšar and Virrantaus (2010)
Others	
Banknotes	(Brockmann et al., 2006)
Census	Rae (2009)
Credit-card transactions	Lenormand et al. (2015)
Highway newtorks	Ren et al. (2014)
Public Transport Cards	Cats et al. (2015); Chen et al. (2014); Hasan et al. (2012)
Surveillance cameras	Anbaroglu et al. (2014)

1. The study of the trajectories of people or means of transport, such as vehicles and boats by using space-time locations registered into CDR from cell phones or Global Navigation Satellite System (GNSS) receivers. For each unit (person or vehicle), the collection of positions form a trajectory; then the primary goal is using classification methods for clustering similar paths (De Domenico et al., 2015; Gao, 2015; Hawelka et al., 2015; Louail et al., 2014; Nanni et al., 2014; Palchykov et al., 2014; Pinheiro, 2014; Lu et al., 2013; Sagl et al., 2012; Yuan and Raubal, 2012; Wang et al., 2011; Demšar and Virrantaus, 2010; Phithakkitnukoon et al., 2010).

2. Geospatial data mining techniques related to the extraction of urban patterns from check-ins and/or content data in Location–Based Social Networks (LBSN), such as Brightkite, Foursquare, Gowalla, and Twitter (Chen et al., 2015; Gao and Liu, 2015; Wu et al.,

2015; Forghani and Karimipour, 2014; Gabrielli et al., 2014; Noulas, 2013; Nguyen and Szymanski, 2012; Ferrari et al., 2011; Noulas et al., 2011; Wakamiya et al., 2011).

Social science is related to the analysis of agents, which can be individuals or organisations and the relationships between them. In this sense, social science data represents cultural values and symbols of agents. This implies that data refers to meanings, motives, definitions, and typifications. There are distinguished three types of data: 1) attribute data, 2) relational data and 3) ideational data. Attribute data refer to attitudes, opinions and behaviour of the agents. Relational data concerns with the ties and connections that link one agent to another. Ideational data expresses definitions, causes, and symbolizations associated in actions of agents (Scott, 2012).

Hence, social networks are framed in the study of relational data. A social network is a social structure consisting of *nodes* (individuals or organisations), sometimes also called *actors* and *relationships* between them (friendship, work, collaboration or siblings), which are also called links (Scott, 2012; Snijders, 2011; Otte and Rousseau, 2002). In Otte and Rousseau (2002) SNA is defined as "*a strategy for investigating social structures through the use of network and graph theories*". Also, Marshall and Staeheli (2015) say the SNA is used to uncover structural patterns of social relations. SNA is a concept applied in many fields, such as marketing, geography, and transport networks. Furthermore, with a mathematical basis in graph theory, SNA is considered a multidisciplinary method becoming a hybrid of information sciences, computer science, geography, and statistics (Robins, 2013).

In addition to spatial and temporal components in social networks, now actors are able to share content data, which is referred to texts, pictures, and videos, among others, this is called LBSN. In this sense, Gao and Liu (2014) indicate that aspects related to human mobility can be seen in a "W^4" (who, when, where, and what) information layout.

Human activity understanding embraces activity recognition and activity pattern discovery. While the first one is related to the accurate detection of human activities based on a predefined activity model, the second one is more about uncovering hidden patterns from low-level sensor data without any predefined models or assumptions (Kim et al., 2010). On the other hand, Goodchild (2007) establishes humans can act as sensors of activities that occur in real life, and this allows for generating content with some associated geographical aspect. Thus, social media data and mainly location-based social networks (LBSN) have become an information source for studying and identifying human activity patterns. The analysis of LBSN data has been an active area of research in urban studies over the last decade which has allowed for developing applications in urban planning (Frias-Martinez et al., 2012; Frias-Martinez and Frias-Martinez, 2014; García-Palomares et al., 2018; Soliman et al., 2017; Resch et al., 2016), human activity (França et al., 2015; Celikten et al., 2017; Ferrari et al., 2011; Wakamiya et al., 2011; Hasan et al., 2013; Huang and Li, 2016), population dynamics (Thakur et al., 2018; Steiger et al., 2015; Patel et al., 2016; Huang and Wong, 2016), and event detection and disaster management (Cheng and Wicks, 2014; Tasse and Hong, 2014; Huang et al., 2018; Resch et al., 2017; de Albuquerque et al., 2015; Shi et al., 2016), among others, as well as, implementing several analytical techniques.

Data analysis on LBSN can be divided into two main approaches, data mining and statistical methods that are looking for identifying groups of spatio-temporal similar locations. From data mining, it stands out: self-organising maps (SOM) (Frias-Martinez et al., 2012; Frias-Martinez and Frias-Martinez, 2014), hierarchical SOM (Steiger et al., 2016), independent component analysis (ICA) (Ferrari et al., 2011), density-based spatial clustering of Applications with Noise (DBSCAN) and ST-DBSCAN (Huang et al., 2018; Shi et al., 2016), and random forests (Patel et al., 2016). Meanwhile, from the branch of statistical analysis, it emerges ordinary least-squares (OLS) (García-Palomares et al., 2018), generalised additive models (GAM) (de Albuquerque et al., 2015), local indicators of spatial association (LISA)(Steiger et al., 2015), space-time scan statistics (STSS) (Cheng and Wicks, 2014), Gaussian mixture models (GMM) (Bakerman et al., 2018), and kernel density estimation (KDE) (Hasan et al., 2013; França et al., 2015). Generally, both alternatives combine the discovery of patterns for the spatio-temporal locations, as well as, for the content data. This latter implies the use of probabilistic topic models such as latent Dirichlet allocation (LDA) algorithms.

Although the aforementioned studies provided promising results, there are still some limitations on them. For example, (García-Palomares et al., 2018) normalised the number of geolocated tweets to use models under the statistical assumptions of the OLS; however, regression for count data captures the nature of the variable in the study directly. On the other hand, techniques such as DBSCAN and ST-DBSCAN do not include spatial, temporal, or spatio-temporal autocorrelation structures in the clustering process, whilst, the analysis of point patterns estimates such as structures with the target of study the distribution of the events. Also, ST-DBSCAN depends on the selection of distance and tolerance parameters (Huang et al., 2018). Finally, in the case of LISA statistics, information is spatially aggregated in pixels without considering the temporal variations of the social media activity that causes modifiable area unit problem (MAUP) (Soliman et al., 2017); then, for avoiding this issue, a better alternative is analysing the locations and timestamps as points rather than spatial aggregations.

3. Spatial interaction data (SID) analysis of origin-destination systems such as flows of people in transport means, such as subway, buses, or bicycles (Cats et al., 2015; Anbaroglu et al., 2014; Chen et al., 2014; Ren et al., 2014; Hasan et al., 2012).

   SID analysis is the field of the Spatial Statistics, which is responsible for analysing the objects' flows within an origin-destination system (Bailey and Gatrell, 1995; Thompson, 1974). The origin and destination are associated with a spatial location; meanwhile, the flow is the strength of relationship between them. This idea is similar to Newton's universal gravitation law, where two bodies exert a reciprocal action in space. Although, initial models almost took in a literal way the classic expression of Newton, this has been changing due to they do not represent adequately the reality, and their assumptions are invalid (Roy and Thill, 2004).

Currently, there are two methodological approaches in models for studying SID (Patuelli and Arbia, 2013). The first rests upon the independence's hypothesis on observed flows, i.e.; they are considered a set of independent random variables with a specified probability distribution. The second is the assumption of spatial dependence; the movements are not independent in space, which means that measuring a characteristic attributable to an entity in space, depends on characteristics of other entities and the spatial relationships that exist among them. Both forms have been useful throughout the history of spatial analysis and assuming a position or another depends on specific factors, such as, scale and characteristics of the base information (Arbia and Petrarca, 2013).

The classic model of SID assumes two fundamental hypothesis. Flows are independent random variables that follow a specific distribution, and the ability to involve spatial effects is determined by a distance or decay function. The interest lies in modelling the mean or expected value of flow from an origin i to a destination j, for which there are several specifications (Griffith and Fischer, 2013; Fischer and Wang, 2011; Roy and Thill, 2004).

The models seek to incorporate variables regarding the ability of a place to generate the outflow and the attractiveness of a destination site for that the flow gets there. These variables must be supported by the nature of the phenomenon under study. Additionally, the models include an impedance's effect in the flow. It is associated with the geographical distance between the origin and destination. Nonetheless, this impedance may be a different variable associated with other distances, whether social, economic or temporal.

However, specifying the expected value of the flows opens other consideration related to determine their stochastic nature (Griffith and Fischer, 2013). The Poisson distribution has traditionally been used to model migration, another distribution that can be associated is the negative binomial of spatial interaction (Fischer and Wang, 2011); the latter is a derived development, where, randomness assumptions, are made on the mean's specification.

According to Griffith (1992) the spatial dependence effect can be interpreted in different ways, among which are: autocorrelation, pattern mapping, absent or unspecified variables, redundant information. Likewise, there are several ways to incorporate, analyse, model and visualise such effect (Griffith and Chun, 2013). For modelling the mean of flows, under spatial dependence, there have used mainly three methodologies. The first is based on spatial econometrics through SAR and SEM, which assume normality in the logarithm of the count of movements (Fischer and Getis, 2010; Fischer et al., 2010; LeSage and Pace, 2008). The second is ESF, using as explanatory variables the eigenvectors of origin-destination matrix on interaction models (Fischer and Griffith, 2008). Finally, a Bayesian Statistics technique that combines data augmentation and MCMC (LeSage and Pace, 2009; LeSage et al., 2007; Frühwirth-Schnatter and Wagner, 2004).

Statistical Framework 3

3.1 Regression models for count data

The regression techniques are used to explain the variations in the mean μ of a variable (called *response variable* and usually denoted by y) associated with a set of factors (called *explanatory variables* $x_1, \ldots, x_p$) and to quantify the magnitude of their effect through a collection of values called *parameters of the model* $\beta_0, \beta_1, \ldots, \beta_p$ (Montgomery et al., 2012). Classical regression models rely on the assumption that the response variable follows a Gaussian distribution (Myers et al., 2012). However, in the case where the response variable y_i, $i = 0, \ldots, n$ represents a count of the number of events per unit of time, area, or volume, i.e. a non-negative, discrete variable, statistical modelling falls under the scope of generalised linear models (GLMs). Those models are a general framework that allows for the modelling of responses whose probability distribution belongs to the exponential family of distributions, such as binomial, Poisson, gamma, and negative binomial, among others (Nelder and Wedderburn, 1972).

To formulate a GLM requires the specification of three elements (Dobson and Barnett, 2008). First, a *random component* referred to the probability distribution of the response variable. Second, a *systematic part* or *linear predictor* η that expresses the parameters of the model as a linear function of the explanatory variables, $\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$. Finally, a monotone, differentiable *link function* g that relates the mean of the response variable with the systematic part, $g(\mu) = \eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$.

In the context of count data, it is common to assume that the random component follows a Poisson distribution when the mean and the variance of the response are equal or a negative binomial distribution when the variance is greater than the mean of the response, which is called overdispersion effect (Hilbe, 1993). In addition, it is customary to use the natural logarithm as the link function to ensure that the predictions of the mean of the response are non-negative (McCullagh and Nelder, 1989). The estimation of the parameters of the model is made by maximum likelihood through the iteratively weighted least squares (IWLS) algorithm (Hardin and Hilbe, 2012).

3.2 Multitype spatial point patterns

Under the scope of spatial statistics, the analysis of spatial point patterns is the branch where the locations of the phenomenon of the study, called *events*, are not fixed, and themselves are the variable of interest (Cressie, 1993). Thus, a set $\mathbf{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_k\} : s_i = \begin{pmatrix} x_i & y_i \end{pmatrix}^\intercal$, $i = 1, \ldots, k$, where $\mathbf{s}_i \in W \subset \mathbb{R}^2$ denotes the location of the i-th event, is called a *spatial point pattern* (O'Sullivan and Unwin, 2014). From the statistical point of view, two types of measures summarise the pattern: (1) *first-order statistics* characterise the mean of the process

$\lambda(\mathbf{s})$, i.e., the number of events per unit of area and (2) *second-order statistics* outline the spatial autocorrelation between the events $\lambda(\mathbf{s}_i, \mathbf{s}_j)$ (Dale and Fortin, 2014). One of the primary objectives of the analysis relies on identifying whether the events exhibit spatial clustering or spatial regularity by using the second-order summary statistics (Baddeley et al., 2015).

Second-order summary statistics are functions that express the degree of spatial relationship between the events of the pattern for several spatial scales (Diggle, 2013). Conventionally, Ripley's K-function ($K(r)$, $r \geq 0$), Besag's L-function ($L(r) = \sqrt{K(r)/\pi}$, $r \geq 0$), and the pair correlation function g ($g(r) = K'(r)/2\pi r, \, , r \geq 0$) are the essential elements for the analysis of point patterns. Thus, the evaluation of the characteristics of the process, such as complete spatial randomness (CSR), clustering, or regularity, is based on the empirical estimation and the values that these functions take. For instance, under CSR, $K(r) = \pi r^2$, $L(r) = r$, and $g(r) = 1$ (Illian et al., 2008).

In many applications, the aim lies in analysing the distribution of various types of points that come from the same origin or are of the same nature. The context might be the research of species in ecology, the characterisation of different classes of crimes in a city, or analysis of case-control studies in epidemiology. In this context, each event is labelled with a mark ζ_j, $j = 1, \ldots, l$ to identify its type and then the set $\mathbf{S} = \{\mathbf{s}_{i_j}, \zeta_j\}$ is called a *multitype* point pattern. Thus, multivariate statistical methods play an essential role in the data analysis since they provide elements for identifying groups of events with similar spatial distribution through the use of clustering algorithms in second-order summary statistics (Baddeley et al., 2015).

The analysis implies the estimation of summary statistics for the formed pattern by each type of mark in several distances, e.g., r_q, $q = 1, \ldots, m$. This estimation produces numerical realisations of l non-observable functions. Although it would be possible to conduct this work by using classical multivariate analysis, the summary statistics are functions instead of single values (Illian et al., 2008). Then, techniques such as the FDA provides tools for understanding the spatial behaviour of the pattern since it considers that observations are functions or single units rather than consecutive measurements (Illian et al., 2006).

3.3 Functional data analysis

A dataset in the FDA is a sample of the following form (Kokoszka and Reimherr, 2017):

$$x_n(t_{j,n}) \in \mathbb{R}, \quad t_{j,n} \in [T_1, T_2], \quad n = 1, \ldots, N, \quad j = 1, \ldots, J_n \tag{3.1}$$

where we have n observed curves over the same interval $[T_1, T_2]$. The basic idea is that the objects of study are the smooth curves

$$\{x_n(t) : t \in [T_1, T_2], \quad n = 1, \ldots, N\} \tag{3.2}$$

defined for all values of t but observed only at selected points $t_{j,n}$.

Then, the first step in the analysis involves rebuilding, through the sample equation 3.1, the functions by using smoothing techniques, which includes determining a set of *functional blocks* or *basis functions* ϕ_m, $m = 1, \dots, M$ and a set of coefficients c_m, $m = 1, \dots, M$ to define each function as a linear combination of these basis functions; thus, $x_n(t) = \sum_{m=1}^{M} c_{nm}\phi_m(t)$, $n = 1, \dots, N$. Although several types of bases exist, it is common to use Fourier basis systems for periodic data or spline basis (b-splines) for aperiodic data (Ramsay et al., 2009).

Conventionally, by using least squares or localised least squares fits, it is possible to estimate the coefficients c_m, $m = 1, \dots, M$. However, such methods are not efficient when observations exhibit a significant level of noise, causing their functional representation to exhibit multiple local fluctuations. Therefore, a *penalised smoothing* approach is preferred to minimise the effect of the random variability. This approach uses a large number of basis functions and penalises the sum of the squares through a *smoothing parameter* λ to enforce a tradeoff between overfitting and oversmoothing of the data to the smooth functions (Kokoszka and Reimherr, 2017).

Extensions of the classical summary statistics for functional data are useful to describe the behaviour of the smoothed functions. Let x_n, $n = 1, \dots, N$, be a set of functions fit to data. The *mean* and *variance functions* are (Ramsay et al., 2009):

$$\bar{x}(t) = \frac{1}{N}\sum_{n=1}^{N} x_n(t), \quad \texttt{var}(t) = \frac{1}{N-1}\sum_{n=1}^{N} \left[x_n(t) - \bar{x}(t)\right]^2 \tag{3.3}$$

In this sense, e.g. the mean function represents the average of the functions point-wise across the replications. Also, as in multivariate data analysis, it is possible to extend the concept of measurements of dependence between curves for different argument values through the *covariance function*, which is defined as (Ramsay and Silverman, 2005):

$$\hat{\sigma}(t,s) = \texttt{cov}(t,s) = \frac{1}{N-1}\sum_{n=1}^{N} \left(x_n(t) - \bar{x}(t)\right)\left(x_n(s) - \bar{x}(s)\right) \tag{3.4}$$

As usual, most of the statistical methods have an adapted version under the scope of FDA. For example, principal component analysis (PCA), discriminant analysis, and the regression techniques, among others (Martínez-Camblor and Corral, 2011) which mainly assume a sample of independent functions (Chen and Müller, 2012). Similarly, when curves are observed through the time, space, or the space-time, there are variants for correlated data, such as repeated measures (Park and Staicu, 2015), time series analysis (Hyndman and Ullah, 2007; Hyndman and Booth, 2008; Hyndman and Shang, 2018), and spatial statistics modelling (Delicado et al., 2010; Mateu and Romano, 2016).

FPCA is a valuable tool to explore and identify features in the curves and the number of types of them. As with the PCA used in classical multivariate methods, FPCA defines a new set of scalar variables $f_{j,n}$, $n = 1, \dots, N$, $j = 1, \dots, J_n$, called *scores*, as linear combinations of the smooth functions. Thus (Ramsay and Silverman, 2005),

$$f_{j,n} = \int \xi_n(t) u_j(t) dt \tag{3.5}$$

where $\xi_n(t)$ is a weight function that maximises $N^{-1}\sum\limits_{n=1}^{N} f_{j,n}^2$ subject to the constraint $\|\xi_n\|^2 = \int \xi_n(t)^2 dt = 1$. This process defines an eigenequation:

$$\int \hat{\sigma}(s,t)\xi_n(t) = \lambda\xi_n(t) \tag{3.6}$$

The solution of equation 3.6 gives the eigenvalues, λ, and the scores. Following the PCA, the scores associated with the first eigenvalue retain the maximum variability of the smooth curves, and so on with the next ones. Then, for subsequent analysis, it is customary to study the first d principal components with $d \ll N$.

3.4 Epidemic data

Epidemic data are conceived as realisations of spatio-temporal processes with autoregressive behaviour which do not come from planned experiments. Its observations, number of events, are not independent, and phenomena are only partially observed. Statistical analysis can be addressed for modelling (model-based approach) or monitoring (test-based approach) epidemic processes, generally in the context of infectious diseases surveillance (Robertson et al., 2010). It is common that data are spatially and temporally aggregated, and there is no information available about the number of susceptibles per spatial unit (Meyer et al., 2017; Paul et al., 2008). Also, time series on counts of infectious diseases exhibit regular patterns over time, i.e. long-term trends, seasonality, and occasional outbreaks (Salmon et al., 2016; Meyer et al., 2017).

For modelling epidemic data several approaches have been developed, among others, generalised linear mixed models (GLMM), Bayesian models, and models of specific space-time processes (Robertson et al., 2010). In this latter direction, Held et al. (2005) and Paul et al. (2008) developed a model for multivariate infectious disease surveillance counts based on a branching process with immigration. This model decomposes the number of events into two parts: (1) a *regular* or *endemic component* that shows the baseline rate with a regular temporal pattern and (2) an *anomalous* or *epidemic component* that reflects occasional outbreaks. It also allows the inclusion of overdispersion and seasonal effects and provides tools to identify sudden happening that is useful in surveillance systems.

Let $y_{i,t}$ be the number of events in the i-th area at the period t, $i = 1,\ldots,m$, $j = 1,\ldots,T$. Those counts are assumed negative binomial distributed (accounting for potential overdispersion), $y_{i,t}|y_{i,t-1} \sim \text{NegBin}(\mu_{i,t},\psi)$ with conditional mean:

$$\mu_{i,t} = \underbrace{\lambda_i y_{i,t-1} + \phi_i \sum_{j\neq i} w_{ij} y_{j,t-l}}_{\text{Epidemic component}} + \underbrace{\alpha_i + \sum_{s=1}^{S_i}(\gamma_{i,s}\sin(\omega_s t) + \delta_{i,s}\cos(\omega_s t))}_{\text{Endemic component}} \tag{3.7}$$

where λ_i represents the autoregressive parameter for the i–th area, ϕ_i quantifies the influence of the counts between connected regions, and w_{ij} are weights defined as a power law of the adjacency order o_{ji} between zones $w_{ij} = o_{ji}^{-d}$ for $i \neq j$ and $w_{jj} = 0$ to consider that humans travel through the areas (Meyer and Held, 2014). Additionally, S_i are the number of harmonics to include and ω_s are Fourier frequencies, e.g. $\omega_s = 2\pi s/24$ for hourly data. The parameter α_i allows different incidence level in the regions.

3.5 Spatio-temporal graphs

In mathematical terms, a *network* or *graph* is a structure utilised to represent the relationships between a pair of objects. Thus, a graph is an ordered pair $G = (V, E)$ where V is a non-empty set of *vertices*, *nodes*, or *points* and E is a set of *links* (Otte and Rousseau, 2002). A link is an ordered pair that represents the association between two nodes. Moreover, the relations between nodes can be *directional*, i.e., the connection from a node to another does not mean that the contrary connection exists, in which case the links are called *arcs*. The relationships can be *non-directional*, i.e., when a node is related to another, it also implies the existence of the reciprocal relationship, where links are called *edges*. Due to this, networks can be classified into two groups, based on the type of relationships established between nodes, are called *direct graph* in case non-directional and *digraph* in directional case (Snijders, 2011; Otte and Rousseau, 2002).

The network structure is described from a set of indicators, including *density* and *centrality*, although they are not unique. The first is also called *connectedness* and describes the general level of interconnection between nodes in a graph. Whilst, the second is generally referred to a particular actor, and measures different aspects, such as the *degree* (number of links that have a node with others), *closeness* (the sum or the average of the shortest distances from a node to all others) and *betweenness* (frequency or number of times a node acts as a bridge along the shortest path between two nodes) (Robins, 2013; Scott, 2012; Otte and Rousseau, 2002).

Statistical models in networks analysis are focused on the study of links, which are generally considered binary random variables, where 1 represents that there is a tie while 0 means, there is not. Models are developed for explaining dependencies between variables, i.e., between links. Although there are many possible types of dependencies, principal ones are the reciprocation of directed ties, homophily, transitivity of ties, degree differentials and hierarchies in oriented networks (Robins, 2013; Snijders, 2011).

When the nodes of a graph represent locations or areas, and each link symbolises the interaction between two nodes (locations), the formed graph is called a spatial network (Guo, 2009). Thus, the inclusion of geographic features is studied to understand how the graph structures are presented in space. For example, in the context of location-based social networks analysis where geographical properties are embedded through location services (Gao and Liu, 2015), people usually connect to others with comparable socioeconomic characteristics such as income, education, and language, among others. Socially it implies that similar people tend to live nearby, i.e., it is likely that spatial dependence schemes exist in social networks (Gao and Liu, 2015).

According to Brugere et al. (2014) a spatio-temporal network can be considered a network representation of relations among nodes which are oriented in geographical locations over time. The models for studying spatio-temporal networks should meet two requirements: (1) accommodate changes in the relations and spatial positioning over time and (2) facilitate efficient computation of results to ensure scalability.

4 A Statistical Approach for Studying the Spatio-temporal Distribution of Geolocated Tweets in Urban Environments

4.1 Context

An in-depth descriptive approach to the dynamics of the urban population is fundamental as a first step towards promoting effective planning and designing processes in cities. Understanding the behavioural aspects of human activities can contribute to their effective management and control. We present a framework, based on statistical methods, for studying the spatio-temporal distribution of geolocated tweets as a proxy for where and when people carry out their activities. We have evaluated our proposal by analysing the distribution of collected geolocated tweets over a two-week period in the summer of 2017 in Lisbon, London, and Manhattan. Our proposal considers a negative binomial regression analysis for the time series of counts of tweets as a first step. We further estimate a functional principal component analysis of second-order summary statistics of the hourly spatial point patterns formed by the locations of the tweets. Finally, we find groups of hours with a similar spatial arrangement of places where humans develop their activities through hierarchical clustering over the principal scores. Social media events are found to show strong temporal trends such as seasonal variation due to the hour of the day and the day of the week in addition to autoregressive schemas. We have also identified spatio-temporal patterns of clustering, that is, groups of hours of the day that present a similar spatial distribution of human activities.

Thus, in this work, we aim to offer, to practitioners and urban researchers, a robust and straightforward methodological strategy for processing significant volumes of human-generated social media data by using efficiently performed and replicable methods that can include new data in the analysis as soon as additional information is available. This effort provides meaningful insights regarding city environments and a picture of the urban population dynamics through knowing the spatio-temporal changes where humans develop their activities (Thakur et al., 2018).

To this end, we suggest a spatio-temporal statistical approach to analyse the collective dynamics of urban environments through the analysis of locations and timestamps of geolocated tweets generated by people in cities. This approach involves the estimation of regression models to characterise the temporal trends of the usage of social media and the use of classification algorithms to identify spatio-temporal patterns of places where humans develop their activities. Our method mainly uses the tools of regression for count data, spatial point patterns, functional principal component analysis (FPCA), and hierarchical clustering. This alternative considers that social media usage is a proxy for when and where humans develop their activities that can impact and shape policies and action plans in cities. Thus, we aim to study the spatio-temporal components of the dynamics of human activities by investigating the distribution of locations and timestamps in geolocated tweets.

Hence, we wish to prove that statistical modelling—and mainly spatio-temporal statistics—is an alternative approach to study urban dynamics. It provides advantages such as (1) the possibilities to analyse significant volumes of human-generated data in cities, (2) a way to gain insight into human behaviour almost in real time, and (3) tools to include implicitly and explicitly spatio-temporal correlation schemas in models and predictions. Besides, statistical modelling provides, by estimating the parameters of the models, a way to explain the processes that generate the data (Diggle, 2013). In such a sense, our approach can be useful in monitoring, comparing, and simulating urban environments more reliably. In this context, techniques such as regression models for count data allow for the inclusion of specific temporal structures such as autoregressive and seasonality effects (Liboschik et al., 2017). On the other hand, the statistical analysis of spatial point patterns identifies schemas of spatial distribution through a set of summary measures defined for different spatial scales (Baddeley et al., 2015; Illian et al., 2008). Furthemore, FPCA brings the possibility of reducing dimensionality, highlighting the relevant underlying characteristics in spatial summary measures (Lee et al., 2015).

To evaluate our proposal, we collect geolocated tweets, accessing the Twitter application programming interface (API) on streaming, for a two-week period in the summer of 2017 in three urban scenarios, namely, Lisbon, London, and Manhattan. We first address the analysis of temporal trends in the usability of social networks at the city level with explanatory models for count data, such as Poisson and negative binomial regression. Those models allow for identifying factors that explain the changes in the number of geolocated tweets collected per hour as a function of the number of tweets in previous hours (autoregressive parameters) in addition to the hour of the day and the day of the week data (seasonal effects). We then study the hourly spatial distribution of the places where people perform social media activities. To do that, we label each location within the hour when the tweet was created to form a multitype spatial point pattern. We estimate second-order summary statistics for each type, such as Ripley's K and pair-correlation functions. We then convert these summary statistics into functional curves by smoothing with the B-spline basis. We apply FPCA over curves and obtain the functional scores. We finally cluster those scores to obtain hours of the day with a similar spatial arrangement of places with events of Twitter activity.

Our approach demonstrates that spatio-temporal statistical analysis provides valuable tools to analyse a significant amount of geolocated human-generated data and provides insights into how human activities occur in the cities. The obtained results in the studied urban environments highlight the presence of several types of patterns through time across space in the usage of social networks by humans. Then, considering those patterns as a reflection of population dynamics in the cities, this line of investigation can provide instruments to define public policies regarding the provision of services and infrastructures and the planning, management, and mitigation of risks. For example, identifying of places commonly visited by people and hours of the day when that happens can suggest changes in the frequency of service of public transport systems and define strategies for disaster management, among others (García-Palomares et al., 2018; Resch et al., 2017; de Albuquerque et al., 2015).

4.2 Data

4.2.1 Collection

Our approach depends on collecting human-made social media activity over a period in an urban environment. Figure 4.1 summarises the process of data collection, which starts with the city and the places where people interact with their ubiquitous devices (smart devices) and share content on social media. Social network services store this content and the associated metadata for several purposes. In some cases, those services provide access to samples of their databases by connecting their APIs. In particular, Twitter offers the possibility to obtain (almost in real time) user-generated data by accessing its streaming API. Several software libraries, such as `twitter4j` of **Java**, `tweety` of MATLAB, `streamR` of **R**, and `tweepy` of Python, among others, allow for researchers to perform this task. We used **R** (R Core Team, 2018), the language and environment for statistical computing, and its package `tweet2r` (Aragó et al., 2018) to download geolocated tweets. `tweet2r` requires the definition of two parameters for the query: (1) a bounding box to establish the spatial scope and (2) a temporal window to set the period when R connects to the API. The downloading process builds files in GeoJSON format, and each file stores up to 3000 tweets. Since streaming collects approximately 1% of the overall activity (Morstatter et al., 2013; Hawelka et al., 2014; Steiger et al., 2016; Steinert-Threkeld, 2018), the gathered amount of data depends on the volume of usage of the social network in the city.

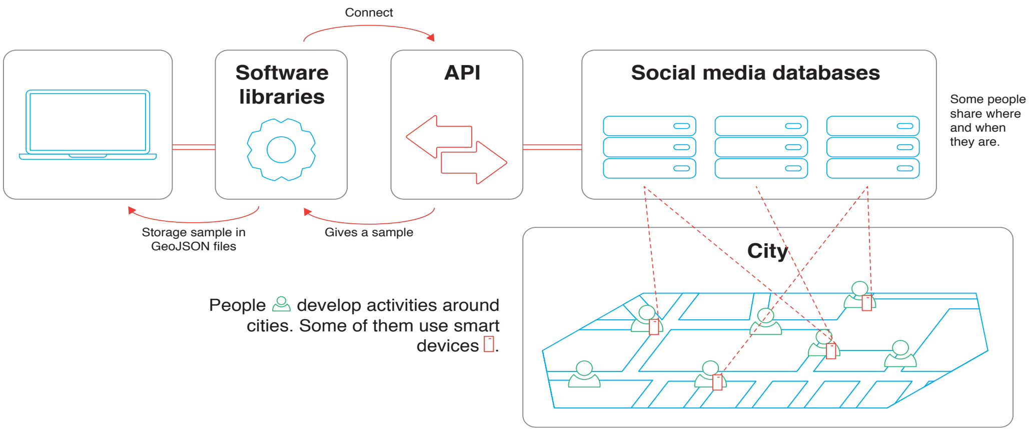

Connect
Software libraries
API
Social media databases
Some people share where and when they are.
Storage sample in GeoJSON files
Gives a sample
City
People develop activities around cities. Some of them use smart devices.

Figure 4.1: Schema of gathering geolocated social media data.

4.2.2 Pre-processing

Human-generated tweets

Once the data are collected, it is necessary to carry out a procedure of pre-processing to identify the information generated exclusively by humans and leave the databases ready for the subsequent analyses. Figure 4.2 presents a schema with the main steps to implement. Initially, GeoJSON files are merged and converted into a table that has by rows each gathered tweet and by columns its metadata. Due to the significant amount of recorded tweets, the gathered data has noises which not necessarily represent people's activities, and it requires filtering to access only those generated by humans before performing any analysis and also, to avoid biases in the results (Yin et al., 2016). The cleaning and removal of the noises is a semiautomatic iterative task that evaluates several sources of perturbation. (Tsou et al., 2017) mention system errors, commercial bot and cyborg tweets, along with user tweeting frequency. On the other hand, (Hawelka et al., 2014; Frias-Martinez et al., 2012) point out the tweeting frequency in the same location, as another aspect to review.

System errors are related to the API since it can provide tweets that do not have geolocation, as well as, information outside of the bounding box. Then, our method removes those rows with missing values in the attributes called *'lat'* and *'lon'* and rules out events registered outside of the boundaries of the box. For detecting the content associated with advertising, (Tsou et al., 2017) suggest reviewing the field called '*source*' in the metadata of the tweets which allows identifying a significant amount of accounts that are continuously sharing commercial information. Then, after tabulating all sources and counting their activity, it is possible to remove tweets that belong to cyborgs by manual inspection. We finally analyse the user and location tweeting frequency, by enumerating unique users and coordinates and then, counting the number of tweets in each case that permit identifying and eliminating those that are related to users and places with a high and unusual frequency.

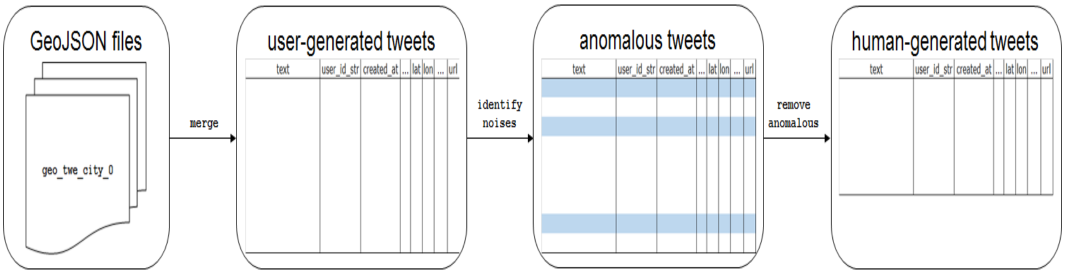


Figure 4.2: Process for pre-processing samples of geolocated tweets.

4.2.3 Datasets construction for statistical analysis

For performing the statistical analysis, our approach builds a new dataset that keeps only three fields, the coordinates (lat, lon) and the timestamps $(created\_at)$. We then add two new columns related to the temporal mark in the following way:

1. We obtain the hour of the day when people created those tweets, labelling each row with corresponding numbers $0, 1, \ldots, 23$.

2. We set, inside of the temporal window of data gathering, a study period, i.e., a start point t_0 and an endpoint t_{T+1}. It is necessary to ensure that the start point is at least 30 hours after the lower boundary of the collecting window to allow for obtaining past information about the process. In addition, we assume that t_i denotes the timestamp of the i-th tweet, $i = 1, 2, \ldots, N$ where N is the total number of collected tweets. Then, by subtracting t_0 from t_i, we obtain the number of elapsed hours from start point until a user shared the i-th tweet. That process allows for defining another timestamp, represented by t_N, through applying the floor function: $t_{N_i} = \lfloor t_i - t_0 \rfloor$. For instance, if $t_0 =$'`2017-07-30 00:00:00`', $t_{T+1} =$'`2017-08-13 00:00:00`' and if the timestamp for a particular tweet is $t_i =$'`2017-08-05 15:18:32`', then the t_N values associated with that study period are between 0 and 335 hours; the elapsed time for that tweet is 159.31 hours, and $t_{N_i} = 159$.

Temporal dataset

Our temporal data analysis approach requires to create a table based on the field called t_N. Let n_h be the count of the number of obtained tweets at hour h, $h = 0, 1, \ldots, T$. So, this procedure gives a discrete-time time series of counts $\{n_0, n_1, \ldots, n_T\}$. We complement the dataset building two sets of dummy variables, as follows: (1) six Boolean variables for the days-of-the-week leaving out the corresponding variable to the Monday and (2) 23 indicator variables for the hours-of-the-day assuming as the reference category the `00:00` hour. Finally, the table includes variables related to the count of tweets in previous hours for catching autoregressive and seasonal autoregressive schemas, the five last hours ($n_{-1}, n_{-2}, \ldots, n_{-5}$) and the same hours the day before ($n_{-24}, n_{-25}, \ldots, n_{-29}$), respectively. Following our previous example, where $t_0 =$'`2017-07-30 00:00:00`' and $t_{T+1} =$'`2017-08-13 00:00:00`', the Table 4.1 shows an schema of a possible temporal dataset.

Table 4.1: Schema of a temporal dataset.

date	t_N	n	*autoregressive*			*seasonal autoregressive*			*day-of-the-week*			*hour-of-the-day*		
			n_{-1}	$\ldots$	n_{-5}	n_{-24}	$\ldots$	n_{-29}	tuesday	$\ldots$	sunday	00:00	$\ldots$	23:00
`2017-07-30 00:00`	0	n_0	n_{-1}	$\ldots$	n_{-5}	n_{-24}	$\ldots$	n_{-29}	0	$\ldots$	1	0	$\ldots$	0
`2017-07-30 01:00`	1	n_1	n_0	$\ldots$	n_{-4}	n_{-23}	$\ldots$	n_{-28}	0	$\ldots$	1	1	$\ldots$	0
`2017-07-30 02:00`	2	n_2	n_1	$\ldots$	n_{-3}	n_{-22}	$\ldots$	n_{-27}	0	$\ldots$	1	0	$\ldots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
`2017-mm-dd hh:00`	h	n_h	n_{h-1}	$\ldots$	n_{h-5}	n_{h-24}	$\ldots$	n_{h-29}	0	$\ldots$	0	0	$\ldots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
`2017-08-12 23:00`	335	n_{335}	n_{334}	$\ldots$	n_{330}	n_{311}	$\ldots$	n_{306}	0	$\ldots$	0	0	$\ldots$	1

Spatio-temporal dataset

To perform the spatio-temporal analysis, we define another dataset based on the locations of the tweets and the hour of the day previously calculated by selecting only the rows that cover the study period. That is, we aggregate and label the data in hourly units of time. We then transform the spatial coordinates to a local coordinate reference system (CRS) through the **R** package `sp` (Pebesma and Bivand, 2005). Table 4.2 shows a schema of a possible spatio-temporal dataset, where (x_{j_h}, y_{j_h}, h) means the location of the j-th tweet shared at the hour of the day h.

Table 4.2: Schema of a spatio-temporal dataset.

east	north	hour
x_{1_0}	y_{1_0}	0
x_{2_0}	y_{2_0}	0
$\vdots$	$\vdots$	$\vdots$
x_{n_0}	y_{n_0}	0
x_{1_1}	y_{1_1}	1
x_{2_1}	y_{2_1}	1
$\vdots$	$\vdots$	$\vdots$
x_{n_1}	y_{n_1}	1
$\vdots$	$\vdots$	$\vdots$
$x_{1_{23}}$	$y_{1_{23}}$	23
$x_{2_{23}}$	$y_{2_{23}}$	23
$\vdots$	$\vdots$	$\vdots$
$x_{n_{23}}$	$y_{n_{23}}$	23

4.2.4 Dataset biases

The representativeness of the harvested human-generated data through the connection to the social media APIs has been a matter of discussion in previous research. There is a consensus regarding the high variation of the spatio-temporal distribution of the tweets (Steiger et al., 2015). Yet, it is not possible to argue that LBSN data are a representative of actual activity in the cities (Celikten et al., 2017) and it requires an assessment that is outside of the scope of this paper. Then, the findings of our approach only represent the contained activity within our Twitter datasets.

4.3 Methods

Our data analysis approach focuses on three main aspects. First, we implement several statistical methods to analyse the spatio-temporal distribution of human-generated social media data, followed by the processing of a significant amount of information almost in real time. Finally, we use easily implemented and reproducible techniques that allow for the inclusion of new data to the models as soon as further information is collected.

Additionally, we include in the analysis spatio-temporal structures that reflect the characteristics of human activity adequately. To this end, we decompose the statistical analysis into two parts (*see* Figure 4.3): the study of the temporal distribution of the hourly number of geolocated tweets in a city and the description of the spatial distribution by hours of the places where people generated the collected tweets.

Many factors can explain the temporal changes in the amount of human-generated data in cities, and some of them can be more relevant to the understanding of human behaviours. For example, in the scope of urban planning and decision-making processes, we can rapidly obtain insights regarding urban dynamics through the identification and impact quantification of issues related to the day of the week, hour of the day, and other temporal trends, such as autoregressive and seasonal effects. In this sense, the regression techniques are flexible statistical methods that can provide valuable tools to study temporal variations in the frequency of social media use.

Although human activity exhibits a high degree of spatio-temporal regularity, the exact place and time of where and when people carry out their activities can neither be fixed nor established by some sampling mechanism. In that sense, the statistical analysis of spatial point patterns plays an important role to study the distribution of the locations where people generate social media data since such an analysis provides elements to describe if those locations present some particular spatial arrangement. Furthermore, determining the temporal variations of that distribution gives a vision of the dynamics of the activities in the urban spaces, e.g., how people move from residential areas in the mornings to working places throughout the day or the frequency of visits to places of interest in the cities. In addition, multivariate techniques allow for us to identify groups of hours that show a similar spatial distribution, ultimately displaying in a synthetic way pictures of urban human activity variations through the day.

This section first presents the main elements of the statistical methods that make up our methodological proposal. It then describes the essentials of regression models for the count data and establishes our procedure to estimate, select, and validate those type of models. We then explain the general framework of the statistical analysis of spatial point patterns. Finally, we address the functional data analysis (FDA) and its application in the context of the multitype spatial point patterns. Let y_t, $t = 0, \ldots, T$ be the number of geolocated tweets at the hour t. We will assume those counts follow a Poisson or a negative binomial distribution with conditional mean μ_t given by:

$$\begin{aligned}\log(\mu_t) = \eta_t = \beta_0 + &\underbrace{\beta_1 I_{tue(t)} + \cdots + \beta_6 I_{sun(t)}}_{\text{day of the week}} + \underbrace{\beta_7 I_{01:00(t)} + \cdots + \beta_{29} I_{23:00(t)}}_{\text{hour of the day}} \\ &+ \underbrace{\beta_{30} n_{-1(t)} + \cdots + \beta_{34} n_{-5(t)}}_{\text{autoregressive}} + \underbrace{\beta_{35} n_{-24(t)} + \cdots + \beta_{40} n_{-29(t)}}_{\text{seasonal autoregressive}}\end{aligned} \tag{4.1}$$

where β_j, $j = 0, \ldots, 40$ represents the parameters of the model, I, the corresponding dummy variables for the day of the week and the hour of the day, and n_{-s}, the counts of the number of the tweets in previous hours. Equation 4.1 describes the *full model*. We use that specification to estimate the parameters of two models, one for each type of response variable, following the procedure suggested by (Katsouyanni et al., 1996). We carry out a stepwise process to select explanatory variables based on the Bayesian information criterion (BIC) (Venables and Ripley, 2002). The obtained models are compared to choose the best model regarding the probability distribution of the response variable by using a likelihood ratio contrast (Cameron and Trivedi, 1986). We finally identify the preferred model and test for normality of the residuals using the Shapiro-Wilk test and for residual autocorrelation using empirical autocorrelation function plots.

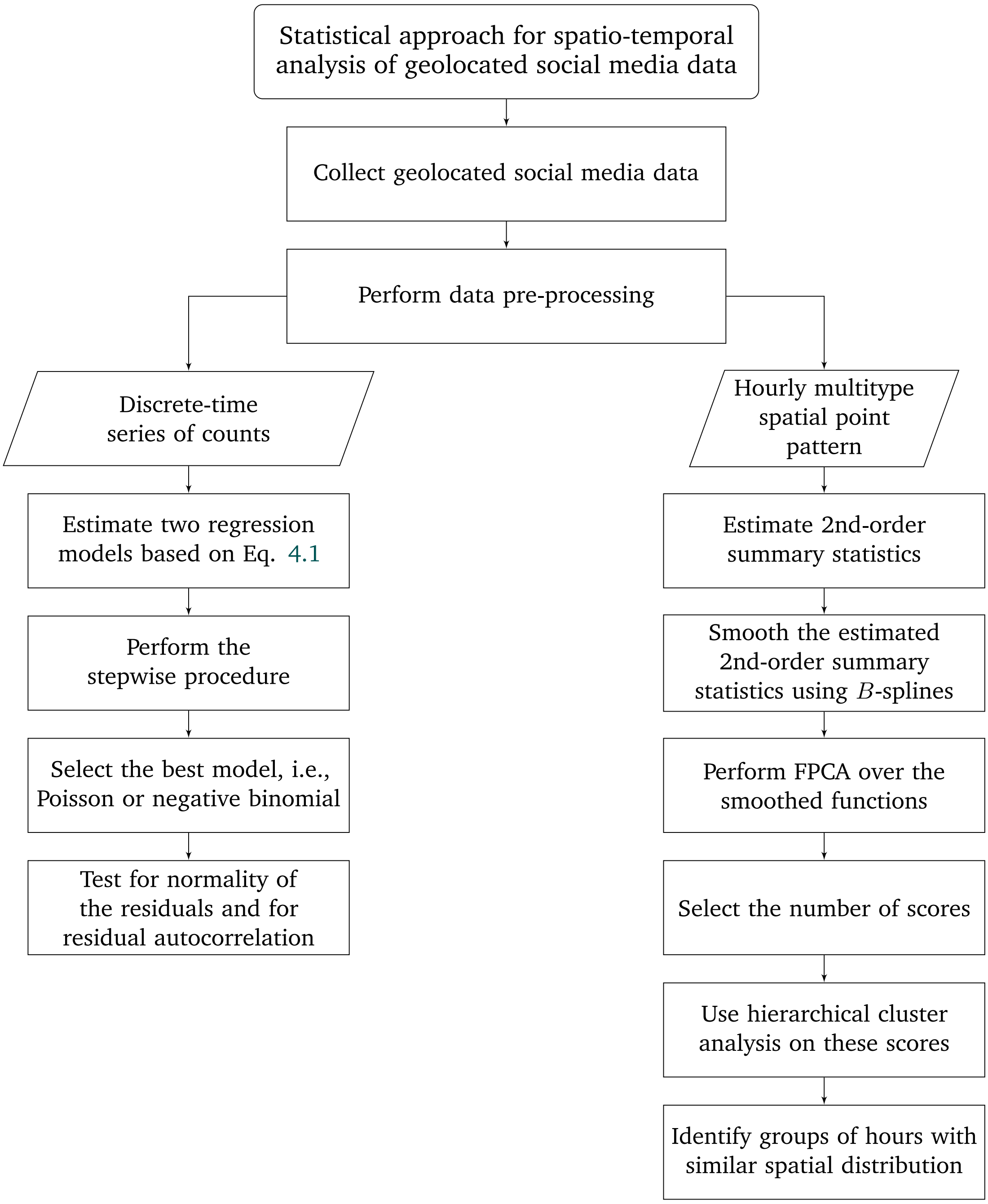

Statistical approach for spatio-temporal analysis of geolocated social media data
Collect geolocated social media data
Perform data pre-processing
Discrete-time series of counts
Estimate two regression models based on Eq. 4.1
Perform the stepwise procedure
Select the best model, i.e., Poisson or negative binomial
Test for normality of the residuals and for residual autocorrelation
Hourly multitype spatial point pattern
Estimate 2nd-order summary statistics
Smooth the estimated 2nd-order summary statistics using B-splines
Perform FPCA over the smoothed functions
Select the number of scores
Use hierarchical cluster analysis on these scores
Identify groups of hours with similar spatial distribution

Figure 4.3: Methodological approach.

The format of the data (*see* Table 4.2) is $\{\mathbf{s}_{j_h}, h\} : j = 1, \dots, n_h$, where each $\mathbf{s}_{j_h} \in W \subset \mathbb{R}^2$ denotes the location, and h, $h = 0, \dots, 23$, the corresponding hour of theday of a tweet shared in the city W. We assume that these data constitute a full register of all events that happen within W at the hour h. We will consider this dataset as an hourly *multitype spatial point pattern*.

Thus, we first estimate the Ripley's K-function for each hourly spatial point pattern using the following estimator:

$$\hat{K}_h(r) = \frac{|W|}{n_h(n_h - 1)} \sum_{i=1}^{n_h} \sum_{\substack{j=1 \\ j \neq i}}^{n_h} \mathbf{1}\{\|\mathbf{s}_{i_h} - \mathbf{s}_{j_h}\| \leq r\} e_{i_h j_h}(r) \tag{4.2}$$

where $|W|$ is the area of the city, n_h is the number of tweets at the hour h, $\|\mathbf{s}_{i_h} - \mathbf{s}_{j_h}\|$ is the distance between the geolocation of two collected tweets at the hour h, $\mathbf{1}\{\cdot\}$ takes the value 1 when the distance is less than or equal to r or 0 otherwise, and $e_{i_h j_h}(r)$ is an *edge correction weight* defined by the geometry of the window and the number of the events in the point pattern (Diggle, 2013).

Based on the estimator 4.2, we calculate the Besag's L-function $\hat{L}_h(r)$, $h = 0, \ldots, 23$ for distances r_q, $q = 1, \ldots, m$, where $r_i < r_j$ $\forall i \neq j$. To decrease the bias in the estimation of the function, we consider r_q values up to $1/4$ of the smallest side length of the rectangle that circumscribes the window W (Baddeley et al., 2015). From those estimations, we then obtain a functional representation of the 24 curves through smoothing techniques with cubic $b-$splines by imposing a roughness penalty based on a harmonic acceleration operator (Kokoszka and Reimherr, 2017). We establish the number of the functional blocks by using the rule $F = m + 2$ (Ramsay et al., 2009). We posteriorly perform an FPCA over the smoothed functions and select as many scores as is necessary to obtain at least 70% of the retained variability. Finally, we perform hierarchical clustering on the selected scores with Ward's procedure (Husson et al., 2017) which allows for the detection of groups of similar second-order summary statistics, i.e., groups of hours whit similar spatial distribution of tweet locations.

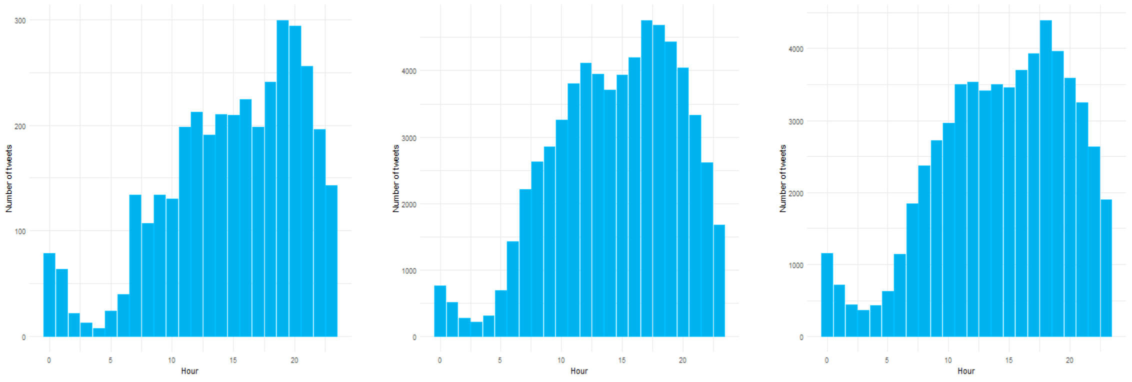
4.4 Results

To evaluate our data analysis approach, we collected geolocated tweets in a two-week period from July 28, 2017, at 12:22:00 UTC/GMT+1 hour to August 14, 2017, 12:21:59 UTC/GMT+1 for the metropolitan areas of Lisbon, London, and New York City. Table 4.3 shows the geographical limits of the corresponding bounding boxes that establish the parameters of the query to connect the Twitter's API in addition to the total number of downloaded tweets and the number of tweets after preprocessing the data. In the case of New York City and the metropolitan area of Lisbon, we restricted the study to the information coming from Manhattan Island and the municipality of Lisbon, to avoid the impact of bodies of water. Thus, we discarded tweets outside of the administrative boundaries of those cities. We collected $4,373$, $79,519$, and $79,649$ tweets for Lisbon, London, and Manhattan, respectively. For the subsequent analysis, we set t_0 =`'2017-07-30 00:00:00'` and t_{T+1} =`'2017-08-13 00:00:00'`. This step provided a study period of 336 hours, between 0 and 335. We then processed $3,626$, $64,404$, and $59,472$ tweets in each urban scenario. We finally transformed the coordinates to the local CRS EPSG:3763 for Lisbon, EPSG:27700 for London, and EPSG:2263 for Manhattan. Figures 4.4 and 4.5 show the bar charts of the temporal distribution of the collected tweets during the study period. We

Table 4.3: Parameters of the query in the Twitter's API.

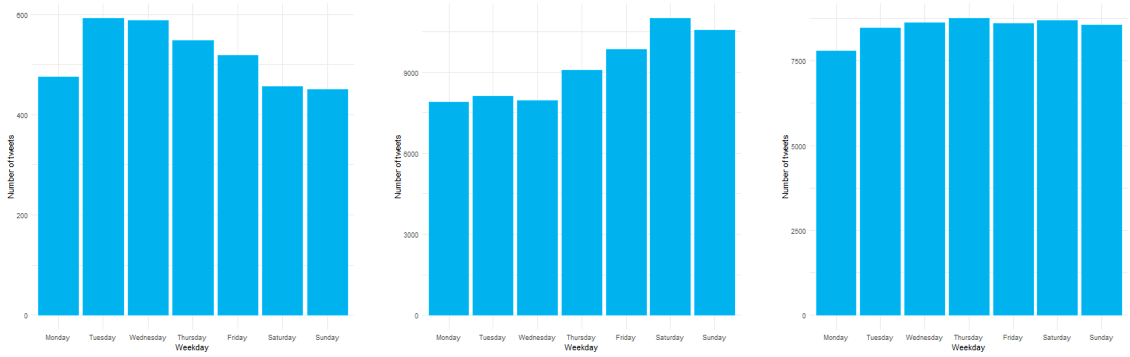
Metropolitan area		Lisbon	London	New York City
Bounding box	(Left, Bottom)	(−9.503, −38.35)	(−0.516, −51.30)	(−73.995, −40.523)
	(Right, Top)	(−8.4925, −39)	(0.36, −51.69)	(−73.695, −40.923)
Number of collected tweets	Total	213,253	1,084,059	1,370,963
	Clean	11,817	87,448	119,802

aggregated the tweets for the two-weeks period by hours of the day and days of the week. We found considerable differences between the amount of gathered information in each urban settlement, but their distribution throughout the day presents similar patterns of behavior. We discovered a profound decreasing on the usage of twitter after midnight and till early in the morning, followed by an increase that gets a peak in the evening. Those maximums did not occur at the same hour, being among 19:00 and 21:00 in Lisbon, from 17:00 to 19:00 in London, and at 18:00 in Manhattan. On the other hand, regarding the day-of-the-week, the three cities showed marked variations, while Lisbon had more social media activity from Tuesday to Thursday, London recorded more of tweets in the weekends than in the weekdays, and there was not a considerable difference in the volume of human-generated data into days of the week in Manhattan. Figure 4.6 displays the three count time series for the 336 hours of analysis where it is evident a daily seasonal effect in the frequency of interaction of people with their social networks. Lisbon's time series exhibits unusual activity, a high number of tweets in the evenings on August 1st, August 3rd, and August 9th, 2017. For each city, we adjusted two count regression models based on the equation 4.1, by considering Poisson and negative binomial responses. Table 4.4 presents the statistics for the goodness of fit to the selected best model in each urban settlement. The likelihood ratio test shows that in all cases the models have a better fit using a negative binomial distribution for the random component in the corresponding GLM. After performing the stepwise variable selection procedure, we concluded that the preferred models are suitable to explain the number of geolocated tweets per hour as a function of the examined explanatory variables since deviance statistics are statistically significant. Table 4.5 presents the summary of the estimation for the selected negative binomial regression models. The results show that the parameters related to the day-of-the-week are positively correlated with the Twitter activity



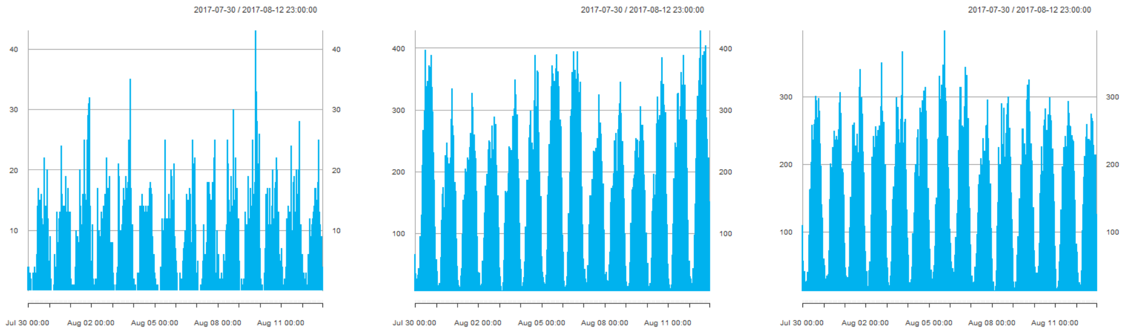
(a) Lisbon (b) London (c) Manhattan

Figure 4.4: Hourly distribution of geolocated tweets.



(a) Lisbon (b) London (c) Manhattan

Figure 4.5: Weekly distribution of geolocated tweets.



(a) Lisbon (b) London (c) Manhattan

Figure 4.6: Time series of hourly geolocated tweets in three urban environments.

Table 4.4: Statistics of goodness of fit for estimated count regression models.

Test	Lisbon		London		Manhattan	
Likelihood ratio (LR)	26.25	***	165.02	***	281.97	***
Deviance (D)	363.77	*	397.32	***	388.88	**

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

in Lisbon, from Tuesday to Thursday, and in London, from Thursday till Sunday, and have no impact on Manhattan. The models indicate that there is a notable correlation between the hour-of-the-day and the amount of social media data shared by people. The pattern is almost the same in the three urban environments, being a negative association from midnight until early in the morning and a positive relationship that increases with the course of the day, peaking at 19:00 in Lisbon, 17:00 in London, and 18:00 in Manhattan. Additionally, some of the parameters related to the autoregressive effects were significant. In Lisbon, the number of tweets created 5 hours before is negatively correlated with the activity for the current hour. In the case of London, an increase of the social media activity in the one hour and three hours before is likely to produce an increase in the number of tweets in the present time. In the same way as the selected model for Lisbon, the estimated model for London reported a negative relationship with the number of tweets five hours before and the activity at the current moment. For Manhattan, only the amount of the tweets in the two previous hours exhibits a positive correlation with the amount of interaction with Twitter in the current time. The stepwise procedure removed all the variables included for the seasonal autoregressive trends.

Table 4.5: Estimated regression coefficients and 95% confidence intervals in the fitted negative binomial regression models for the number of geolocated tweets per hour.

(a) Lisbon

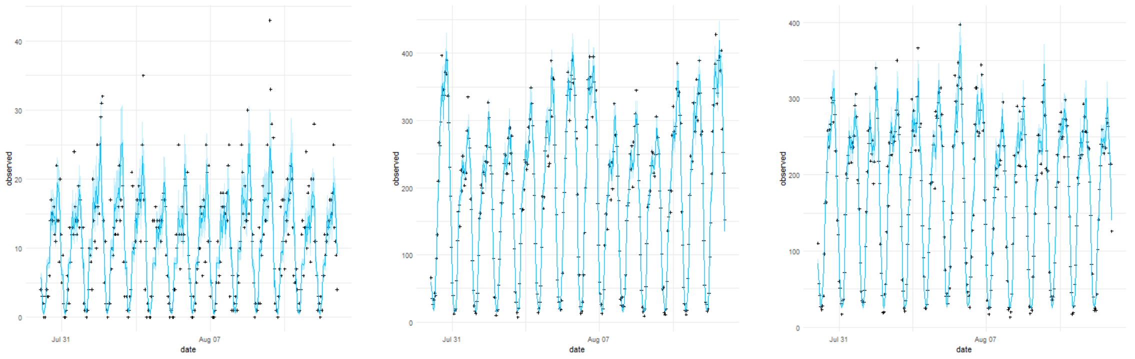
Parameter	Estimate	95% CI
Intercept	2.062	(1.952,2.172)
Tuesday	0.237	(0.112,0.362)
Wednesday	0.239	(0.113,0.364)
Thursday	0.197	(0.069,0.325)
02:00	-1.368	(-1.85,-0.934)
03:00	-1.986	(-2.604,-1.458)
04:00	-2.536	(-3.333,-1.893)
05:00	-1.523	(-1.977,-1.115)
06:00	-1.033	(-1.393,-0.698)
11:00	0.531	(0.321,0.741)
12:00	0.736	(0.53,0.942)
13:00	0.585	(0.374,0.795)
14:00	0.723	(0.515,0.929)
15:00	0.712	(0.505,0.919)
16:00	0.865	(0.653,1.076)
17:00	0.751	(0.533,0.97)
18:00	0.932	(0.725,1.14)
19:00	1.161	(0.959,1.365)
20:00	1.144	(0.941,1.348)
21:00	1.036	(0.825,1.249)
22:00	0.731	(0.513,0.948)
23:00	0.471	(0.229,0.711)
n_{-5}	-0.018	(-0.026,-0.01)

(b) London

Parameter	Estimate	95% CI
Intercept	3.803	(3.721,3.884)
Thursday	0.07	(0.027,0.112)
Friday	0.125	(0.08,0.17)
Saturday	0.174	(0.122,0.226)
Sunday	0.156	(0.106,0.206)
01:00	-0.291	(-0.414,-0.169)
02:00	-0.859	(-1.004,-0.717)
03:00	-1.055	(-1.214,-0.9)
04:00	-0.741	(-0.882,-0.603)
06:00	0.685	(0.578,0.791)
07:00	1.019	(0.909,1.129)
08:00	1.077	(0.958,1.196)
09:00	1.078	(0.954,1.203)
10:00	1.163	(1.037,1.289)
11:00	1.289	(1.165,1.412)
12:00	1.33	(1.204,1.456)
13:00	1.251	(1.12,1.382)
14:00	1.201	(1.073,1.33)
15:00	1.292	(1.169,1.414)
16:00	1.37	(1.249,1.491)
17:00	1.496	(1.371,1.621)
18:00	1.401	(1.263,1.54)
19:00	1.327	(1.188,1.466)
20:00	1.248	(1.111,1.384)
21:00	1.119	(0.991,1.247)
22:00	0.998	(0.883,1.114)
23:00	0.646	(0.537,0.755)
n_{-1}	0.002	(0.001,0.002)
n_{-3}	0.001	(0.0002,0.001)
n_{-5}	-0.001	(-0.001,-0.0003)

(c) Manhattan

Parameter	Estimate	95% CI
Intercept	3.965	(3.823,4.108)
01:00	-0.321	(-0.455,-0.187)
02:00	-0.698	(-0.854,-0.542)
03:00	-0.813	(-0.984,-0.644)
04:00	-0.618	(-0.789,-0.447)
05:00	-0.252	(-0.416,-0.089)
06:00	0.318	(0.166,0.47)
07:00	0.697	(0.555,0.839)
08:00	0.824	(0.693,0.956)
09:00	0.837	(0.714,0.959)
10:00	0.848	(0.728,0.968)
11:00	0.955	(0.835,1.076)
12:00	0.869	(0.739,0.999)
13:00	0.791	(0.661,0.922)
14:00	0.841	(0.714,0.968)
15:00	0.82	(0.691,0.95)
16:00	0.884	(0.756,1.013)
17:00	0.919	(0.787,1.052)
18:00	0.976	(0.837,1.114)
19:00	0.795	(0.645,0.945)
20:00	0.731	(0.586,0.877)
21:00	0.711	(0.577,0.846)
22:00	0.572	(0.445,0.699)
23:00	0.354	(0.233,0.474)
n_{-1}	0.002	(0.001,0.003)
n_{-2}	0.001	(0.000,0.002)

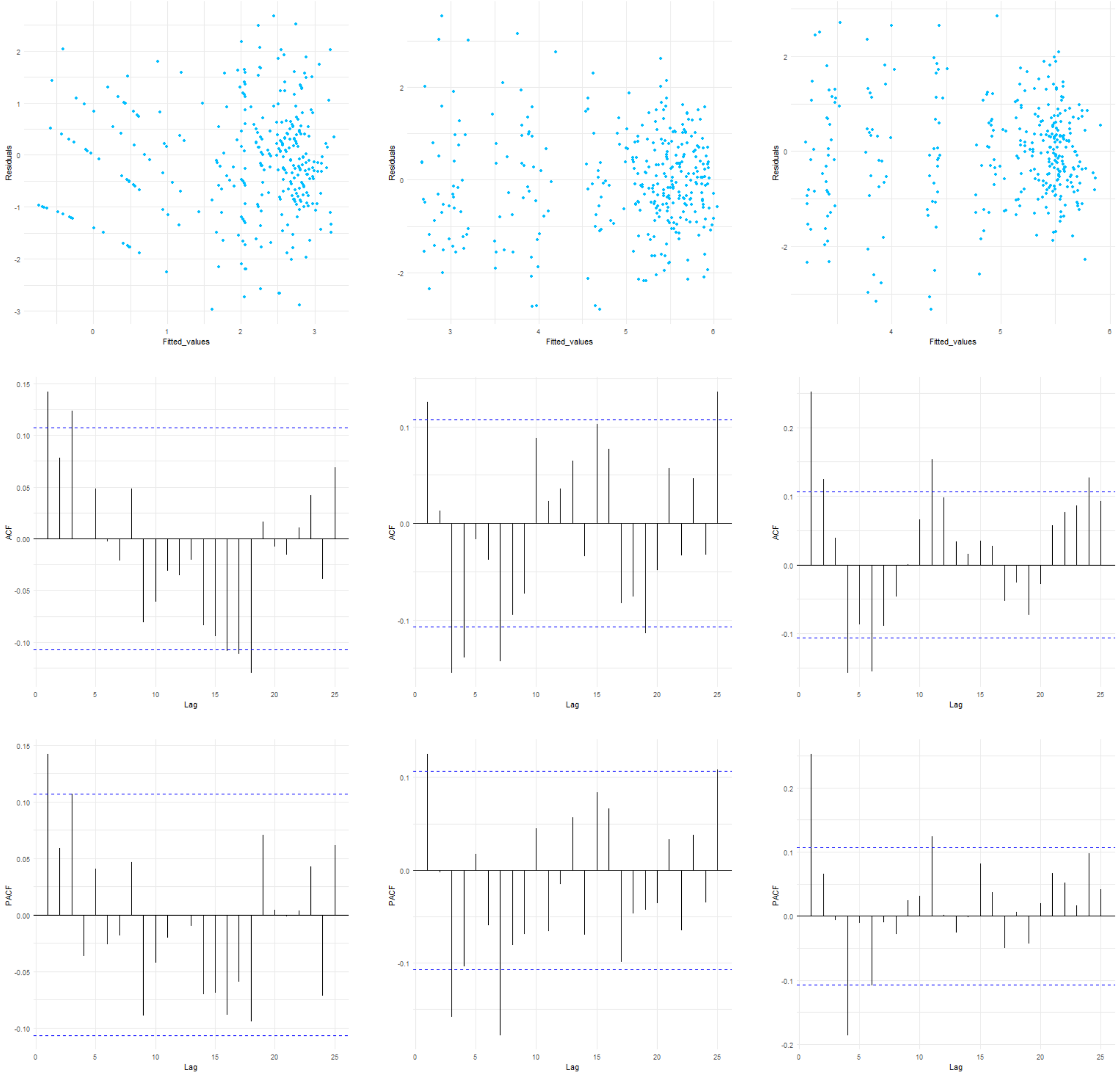


(a) Lisbon (b) London (c) Manhattan

Figure 4.7: Observed temporal variation of geolocated tweets (black dots) together with the fitted variation from a negative binomial regression model (deep sky blue areas).

Figure 4.7 compares observed and fitted numbers of geolocated tweets over the observation period in the three cities. The predicted values through the models represent most of the trends that data exhibit adequately. To evaluate the validity of the models and to identify departures from the statistical assumptions, we conducted a residual analysis. The results of the Shapiro-Wilk's test, Lisbon: $W = 0.996$, $p-$value$= 0.65$; London: $W = 0.995$, $p-$value$= 0.34$;

Manhattan: $W = 0.996$, $p-$value$= 0.27$, present that there is no statistical evidence to reject the null hypothesis that states the residuals of the models follow a Gaussian probability distribution. Additionally, Figure 4.8 shows the residuals versus the fitted values and autocorrelation function and partial autocorrelation function plots. We note that no apparent patterns arise from the relation between residuals and adjusted values, as well as, that the residual autocorrelation is not significant. For each city, we built a multitype spatial point pattern, labelling each location with the hour when a user created the corresponding tweet. Figure 4.4 shows the distribution of the number of events for each mark, displaying the dynamics of the social media activity through the day. We established the length of the smaller side of the rectangles that circumscribe the city of Lisbon, the London's metropolitan area, and the island of Manhattan. Based on those lengths, we defined the maximum distances (r_m) to estimate the L-Besag's function. We worked with a sequence of values from 0 metres up to r_m, each 25 metres. Table 4.6 shows a summary of



(a) Lisbon (b) London (c) Manhattan

Figure 4.8: Residual plots for the selected regression models.

the obtained ranges. We then estimated the centred version $\hat{L}(r) - r$ of the function, for every hourly formed spatial point pattern, over the set of those distances. We obtained the functional representation of those estimations by using functional blocks of 117, 451, and 304 for Lisbon, London, and Manhattan, respectively and a roughness penalty with a smoothing parameter $\lambda = 0.00001$. We after calculated the FPCA over the smoothed curves and kept the first two principal scores since they cumulated more than 70% of the variability in the three scenarios. Additionally, to disclose more significant components of variation, we rotated the functional principal components with the VARIMAX rotation algorithm (Ramsay et al., 2009). We finally got the dendrogram by applying agglomerative hierarchical clustering through Ward's method on the matrix of dissimilarities computed with the Euclidian distance between the scores. Figures 4.9, 4.10, and 4.11 present the results of the spatio-temporal data analysis approach in the three studied urban scenarios. In the case of Lisbon, the smoothed functions reveal schemes of spatial clustering for almost all of considered distances and hours of the day, except for the case of the events registered at 04:00 whose curve decreases rapidly and reach negative values after 1.75 km. Also, those functional representations belonging to hours from midnight to early in the morning (light deep sky blue curves) are more irregular than those associated with later hours. The first two principal components retain 86.06% and 7.84% of the variability, respectively. As a functional principal component symbolises variation over the average curve, the interpretation depends on this capability. Thus, since the first component takes negative values for distances up to 500 metres, approximately the variation of the mean of the hourly second-order summary statistics, the relationship is strongest for distances longer than this value, and the second component captures primarily variations in the hourly summaries up to 1.5 km. Panel (c) of Figure 4.9 reveals that the spatial distribution, of the shared events at 04:00, is quite dissimilar in comparison with the behaviour of the distributions for the other hours of the day. There are approximately three groups of hours for human activities, thus: (1) between 00:00 and 01:00, (2) from 02:00 to 07:00, and (3) at the rest of the hours.

Table 4.6: Distance parameters for estimating the second-order summary statistics for the hourly multitype spatial point patterns of tweets in three urban settlements.

City	Length of the shorter side	1/4 of the length	r_m	m
Lisbon	$11,530.11$	$2,882.53$	2875	115
London	$44,819.03$	$11,204.76$	$11,200$	449
Manhattan	$30,153.90$	$7,533.96$	$7,525$	302

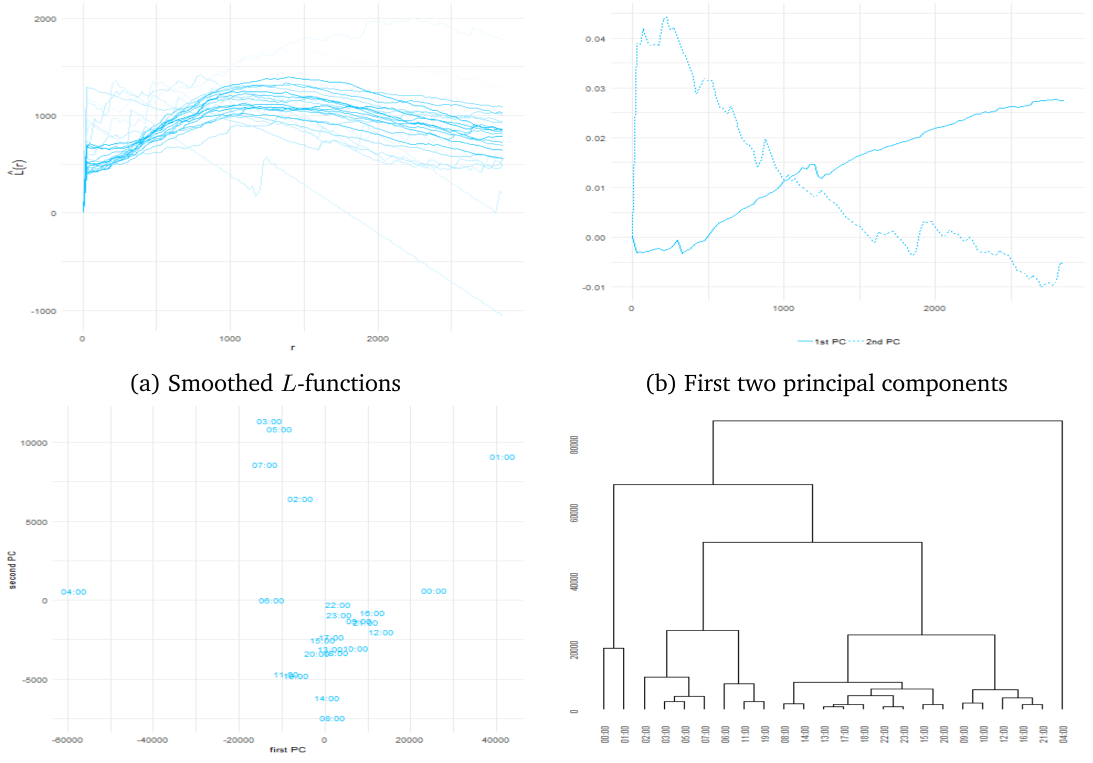
The smoothed centred L-Besag's functions for London show less irregularity than for Lisbon. The curves also exhibit a pattern of spatial clustering for all distances and hours of the day. The functional representation for the hourly second-order summary statistics reveals marked differences between the curves associated with tweets generated in dawn hours to the curves from tweets shared in other periods of the day. The first two principal components explain 97.5% and 1.73% of the variability of the summary statistics, respectively. The first harmonic portrays a continuous increase of the variation of the mean function with the distance, mainly from 3 km. The distribution of the hours through of the scatterplot of the first two functional principal components tells that there are four groups approximately, three of them for early hours in the morning and the other for the rest of the day.

Similarly to the other two cities, the obtained results for Manhattan indicate that the spatial arrangement of the collected geolocated tweets exhibits schemas of spatial clustering for every hour and all distances. Also, the smoothed curves that belong to hours after midnight until `05:00` have more irregularity than at other hours of the day. The first two harmonics keep 70.13% and 26.17% of the variability of the smoothed L-Besag's functions. The first component reveals that the variation in the mean function of the second-order summary statistics is increasing for all distances being higher from 2.8 km meanwhile the second component portrays increments of the variations up to 4.2 km. The hours form almost four groups. One for activities done from `00:00` to `02:00`, another for `03:00` to `05:00`, and the other two for later hours.

We finally obtained the intensity function of each mark of the multitype spatial point pattern, by using bidimensional density kernel estimation. We employed the quartic kernel and selected the bandwidth by using Scott's method (Scott, 2015). We later standardised all the estimated values and brought them to the scale 0 - 1 by subtracting their minimum and then dividing in their range. Figures 4.12 to 4.14 display the estimations. In Lisbon, there is a persistent accumulation of the human-generated data in the margin south that borders with the Tagus river where are located the main places of interest of the city. On the other hand, London's metropolitan area concentrates social media activity in the surrounding boroughs of the city of London where are located touristic places, big companies, and commercial areas. The island of Manhattan aggregates most of the Twitter's activity in the direction south-west from the Central Park to the limits with the Upper Bay and the Hudson River that locates Times Square, SOHO, and the Financial District, among others.

4.5 Discussion

We first examined the temporal distribution of the number of geolocated tweets per hour by using regression models for count data under the scope of the GLMs. We evaluated and found that in the three studied cities, the models have a better goodness of fit when we used the negative binomial distribution for the random component. This result implies that the counts exhibit a high heterogeneity, which reflects the complexity of the analysed systems and agrees with as stated by (Batty, 2009). We additionally detected strong temporal trends related to the day of the week and the hour of the day that reinforce the idea that people who publish geolocated tweets tend to develop their activities approximately at the same times. (França et al., 2015; Frias-Martinez and Frias-Martinez, 2014; Steiger et al., 2016) studied social media data from London and Manhattan and identified areas with high social media activity and differences in behaviour between weekdays and weekends and in hours of the day. However, those temporal patterns change between cities. Our results in the case of Manhattan show that the estimated model did not establish a significant difference in the days of the week, which is contrary to the findings in previous research. These divergences can be due to those aforementioned studies used



(a) Smoothed L-functions

(b) First two principal components

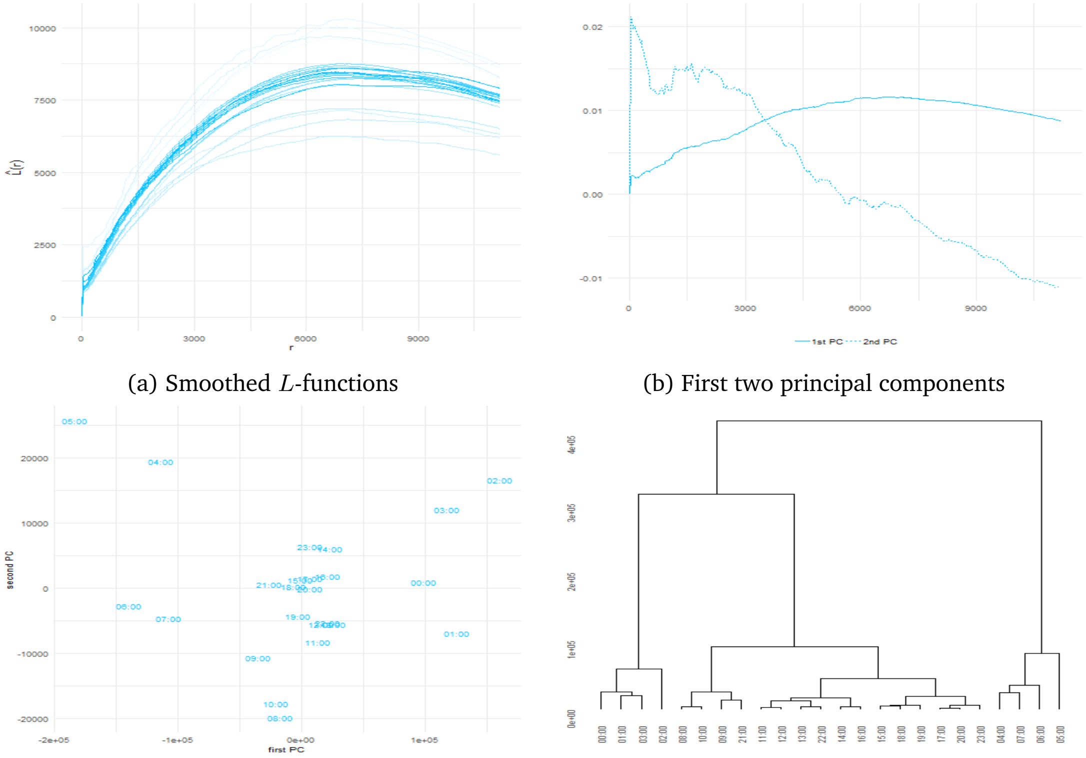
(c) A plot of the scores of all hours

(d) Dendrogram of the scores

Figure 4.9: Results from a FPCA on L-functions in Lisbon.

a more extended period of data collection than ours, which allowed the authors to have a broader image of the urban dynamics, not just a two-week period in a summer, and their conclusions are based on frequencies while we used a more sophisticated approach that included regression modelling and statistical hypothesis testing. The dissimilarity of the frequency of people's use of social media in the other two cities through the weekdays might be an effect of the unusual counts registered in the time series of Lisbon that increased the volume of human-generated data from Tuesday to Thursday. After a thorough review, we attribute the outlier occurring on August 9, 2017, at 19:00 to the prematch tweets of the Portuguese local soccer league between Benfica versus Braga. Our approach also involved the estimation of parameters associated with autoregressive trends. The findings highlight that those temporal effects are also significant to explain the number of tweets and can be meaningful as a measure to anticipate the pressures of increasing the amount of human activity.

We then investigated the spatio-temporal distribution of the geolocated tweets. We linked elements of statistical analysis of spatial point patterns, FDA, and hierarchical clustering. We discovered that locations, where people create and share social media data, exhibit a pattern of spatial clustering for every hour during the day and for all considered spatial scales. This result agrees with the fact the people tend to visit the same places at the same times (Gao and Liu, 2014; Song et al., 2010b; González et al., 2008). Furthermore, we detected that those schemes of clustering change through the day, being more similar from 08:00 to midnight and highly unlikely between midnight and early hours in the morning. We also found that the measures



(a) Smoothed L-functions

(b) First two principal components

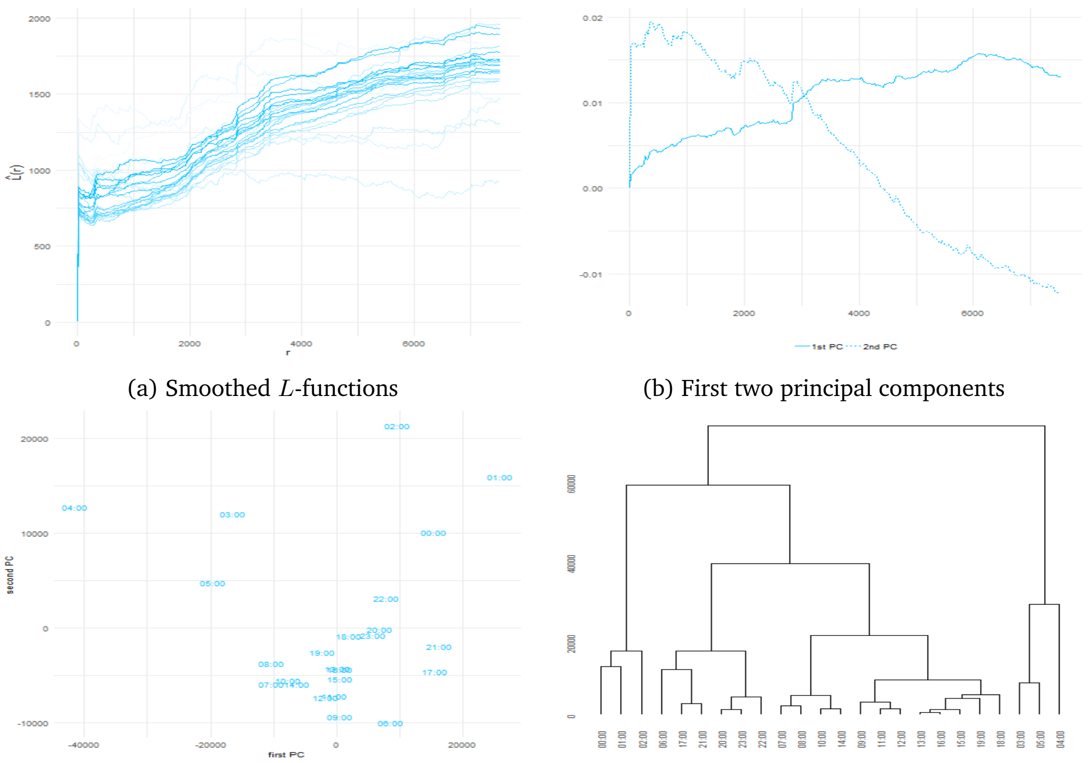
(c) A plot of the scores of all hours

(d) Dendrogram of the scores

Figure 4.10: Results from a FPCA on L-functions in London.

of spatial correlation through the time tend to be more homogeneous in short distances, at 500 metres, 3 km, and 2.8 km for Lisbon, London, and Manhattan, respectively. These values differ significantly with travel distance of 1.5 km reported in human mobility studies (González et al., 2008; Simini et al., 2012; Song et al., 2010b). The behaviour of the smoothed second-order summary statistics showed more uniform curves in London and more erratic curves in Lisbon, which might be an effect of the number of gathered tweets in each city in the two-week period. The analysis also revealed that the places where people share content in Twitter are located in the same areas at the same hours, which is a common feature in the social conduct of humans. The irregular shape of the curves for dawn hours retained most of the variability of the L-Besag's functions and covered other spatial effects that might occur in other periods of the day.

Considering our results, we suggest that an approach based on epidemic data can more effectively accommodate the presence of outliers and might even be capable of predicting them. Epidemic data are conceived as realisations of spatio-temporal processes with autoregressive behaviour which do not come from planned experiments. Its observations, number of events, are not independent, and phenomena are only partially observed (Meyer et al., 2017). There is a high similarity with the distribution of the number of geolocated tweets. Those methods also include autoregressive trends and spatio-temporal structures in the estimation of the parameters of the models that might describe human social conduct more accurately. Our analysis has shown the data coming from the hours commonly dedicated to rest might hide spatio-temporal patterns in the behaviour of the people in the cities at other times of the day. Therefore, we also suggest



(a) Smoothed L-functions

(b) First two principal components

(c) A plot of the scores of all hours

(d) Dendrogram of the scores

Figure 4.11: Results from a FPCA on L-functions in Manhattan.

that to avoid the randomness associated with activity during those hours, the analysis of the human activity in cities should be restricted to the hours of the day where humans are more active and developing their daily life.

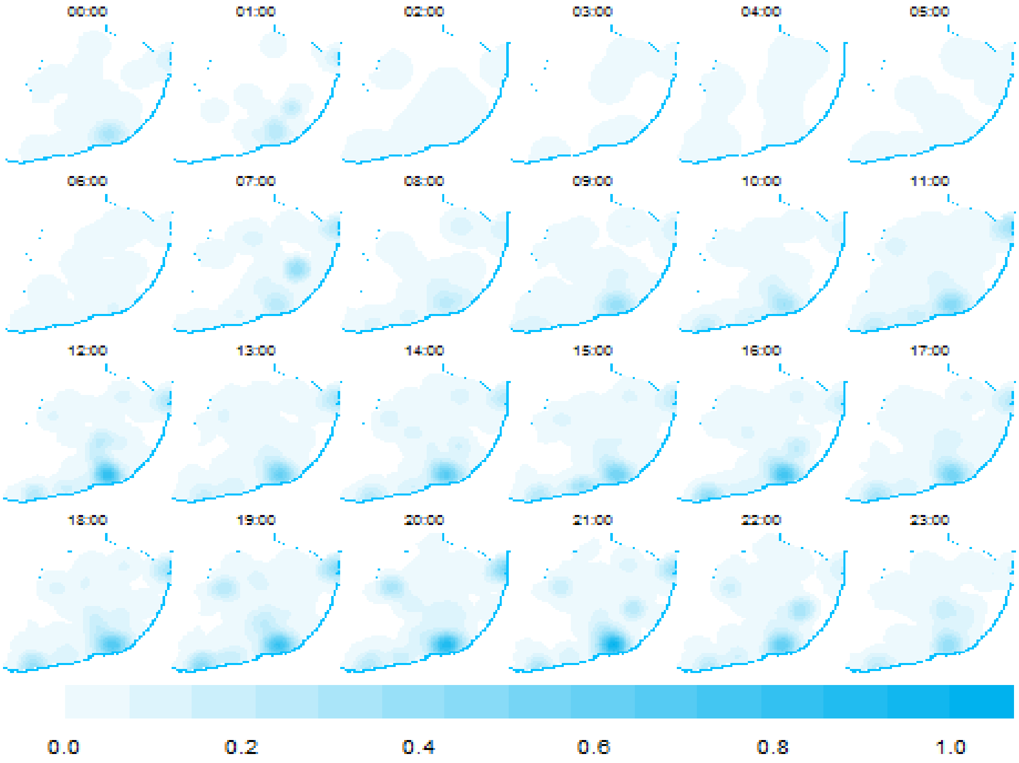


Figure 4.12: Estimated intensity function of the hourly multitype spatial point pattern in Lisbon.

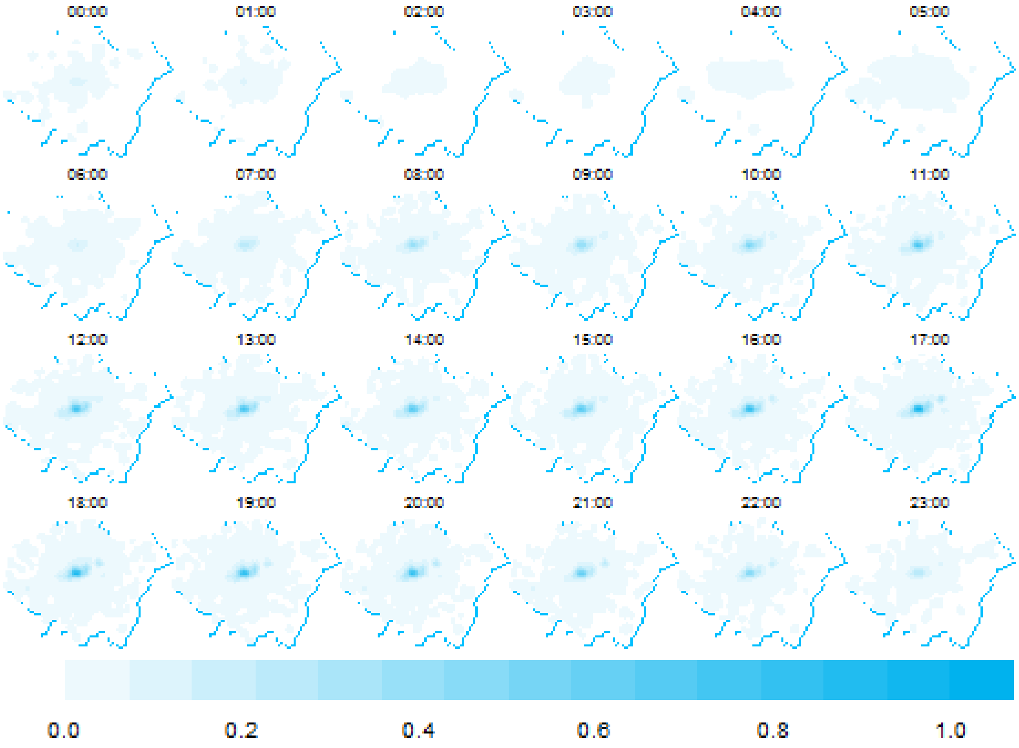


Figure 4.13: Estimated intensity function of the hourly multitype spatial point pattern in London.

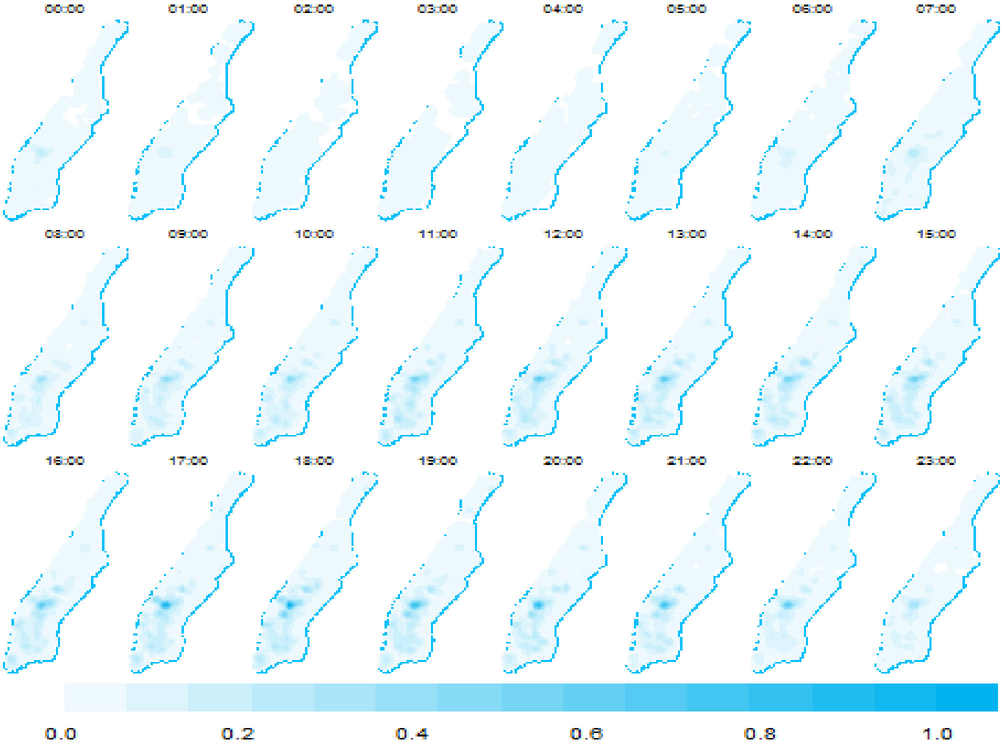


Figure 4.14: Estimated intensity function of the hourly multitype spatial point pattern in Manhattan.

Understanding Human Urban Activity Through the Statistical Analysis of Epidemic Data

5

Understanding urban dynamics is a crucial task for the analysis of cities. It provides core elements for informed urban planning and decision-making processes. Social media data has become a sensor of human activities and a valuable source of information to study where and when these occur. This research is proposing a framework based on the use of statistical modelling developed in the context of infectious disease surveillance for explaining the spatio-temporal distribution of social media data as a proxy of human activity in cities. To evaluate this, we gathered live stream tweets of three urban environments in a two weeks period during the summer of 2017 and estimated non-linear random effects multivariate models to explain the number of geolocated tweets per region per hour. This approach assumes that the count of tweets follows a negative binomial probability distribution and decomposes the conditional mean into two elements. First, an epidemic component consisting of temporal autoregressive effect and spatial neighbourhood defined by a power law concerning distance. Second, an endemic part that includes seasonality and day of the week effects. The model selected caught the temporal trends in social media activity accurately and showed the differences in the dynamics of the cities. It was also able to identify regions and times with unusual behaviour. Our proposal provides through the parameters for the identification of endemic and epidemic components of the social media data related to human activity an alternative reading to this phenomenon that can replicate or simulate complex systems and be useful in monitoring urban dynamics in real-time.

5.1 Introduction

The study of city dynamics has been receiving considerable attention in the geospatial investigation due to the diversity and the complexity of issues appearing in the urban systems (França et al., 2015; Batty, 2009; Celikten et al., 2017). Although, the lack of data for dealing with this matter was a substantial problem. Nowadays, several actors including geographers, social researchers, and data scientists consider social media data as a core source for studying cities (Silva et al., 2013). Despite of it, this information is sparse in the geographical space, incomplete in a time interval (Ferrari et al., 2011), and might not be representative (Toole et al., 2015). It is considered a better alternative for analysing human activity than survey sample techniques through the use of questionnaires since it catches people's perceptions and spatio-temporal changes more accurately in real-time (França et al., 2015; Frias-Martinez et al., 2012; Wakamiya et al., 2011). Such data has allowed developing specific techniques which involve network analysis, data mining, and statistics in a new branch of knowledge called urban informatics or urban analytics (Stimmel, 2015; Zheng et al., 2014).

This work develops an alternative, built on statistical methods of epidemic data, for studying data from LBSN as a proxy for human urban activity. It considers a model-based approach (Robertson et al., 2010) that has been used in the context of multivariate modelling of infectious disease surveillance counts (Paul et al., 2008; Held et al., 2005). Through this approach, it aims to understand how social media data behaves across space and time. This model decomposes the counts into two parts, the regular (endemic component) and the anomalous (epidemic component) activity and allows the inclusion of overdispersion and seasonal effects that are common in human behaviour. Additionally, it provides a tool for identifying outbreaks, characteristic of surveillance systems, which can be useful for urban planners and decision-makers.

However, the use of social networks and their data has not been fully explored in this matter. Specially, statistical modelling can provide elements (parameters of the models) to understand underlying processes generating the data and be useful in monitoring urban environments more reliably. In this context, epidemic phenomena exhibit similar characteristics to social media activity. Thus, statistical methods for this type of data can be potentially used for analysing data coming from LBSN. Even though to the best of our knowledge, there has not been developed a specific application of epidemic data for studying human urban activity, reports can be found about syndromic surveillance systems in the frame of preventing bioterrorism attacks. (Bradley et al., 2005; Buehler et al., 2003).

Prior studies that use location-based social network (LBSN) data as the primary data source have been narrowed to data mining techniques for examining urban dynamics and human activity and extracting urban patterns. In modelling urban dynamics and human activity, (Celikten et al., 2017) implemented a probabilistic topic modelling in a dataset of geotagged activity from Foursquare that was accessed from check-ins via Twitter. The authors included in their analysis the exact location of the users and the timestamps of the events. They reported unique features of the geographical areas and similar regions across different cities. (França et al., 2015) studied the dynamics of Manhattan using five months of geolocated tweets. The authors aggregated the data in the corresponding days of the week and in hourly units of time. In addition, the authors identified areas with high social media activity and differences in behaviour between weekdays and weekends and in hours of the day. In extracting urban patterns, (Ferrari et al., 2011) analysed 13 million Twitter posts in New York City using latent Dirichlet allocation (LDA) algorithms. Such an approach allowed for the authors to identify hotspots in the city life that persist over time and space in the urban scenario. However, to the best of our knowledge, no previous research has been perfomed in the direction of spatio-temporal statistics.

Epidemic data are conceived as realisations of spatio-temporal processes with autoregressive behaviour which do not come from planned experiments. Its observations (number of cases) are not independent, and phenomena are just partially observed. Statistical analysis can be addressed for monitoring or modelling epidemic processes, generally in the context of the infectious diseases (Meyer et al., 2017; Salmon et al., 2016). Moreover, time series on counts of infectious diseases exhibit regular patterns over time, i.e. long-term trends, seasonality, and occasional outbreaks (Paul et al., 2008; Held et al., 2005), aspects that also present human activity.

5.2 Data

Geolocated social media data coming from Twitter Streaming API was collected using tweet2r (Aragó et al., 2018) library from R that enables to gather a sample of all tweets created in a specific spatial boundary box in near real time. The process was carried out during a period in the summer of 2017 in a window that covers the metropolitan area of Lisbon, Portugal (*see* details in Table 5.1 and Figure 5.1). The process produced files in GeoJSON format, which were transformed into a table with the location (longitude and latitude) of each downloaded tweet. The analysis of the information ruled out the events registered outside of the boundary of the city. The coordinates were projected to the local coordinate reference system (CRS) of Lisbon (EPSG: 3763).

Table 5.1: Parameters of the query in Twitter Stream API.

Local time	
Start	End
2017-07-28 12:22:00	2017-08-14 12:21:59
Boundary box	
(Left - Bottom)	(Right - Top)
(-9.503; - 38.35)	(-8.4925; - 39)

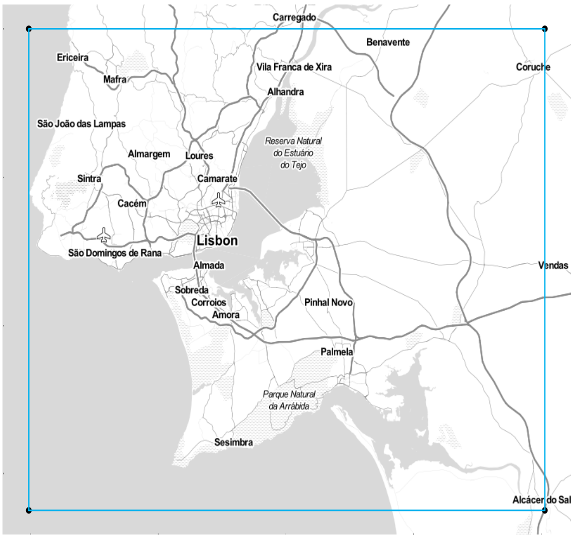


Figure 5.1: Bounding box around Lisbon Metropolitan Area

The dataset was transformed as follows. The runway of the airport and an extensive green area (called Monsanto Forest Park) were not considered as potential zones with high impact on urban human activity whereby they are not included in further analysis. The city was divided into rectangles whose size was determined using Scott's method (Scott, 2015) to estimate the bandwidth in kernel smoothing over the geolocations of the collected tweets (*see* Figure 5.2). Finally, a new spatio-temporal dataset was built where each rectangle is associated with an hourly time series of the count of the number of tweets gathered for that region.

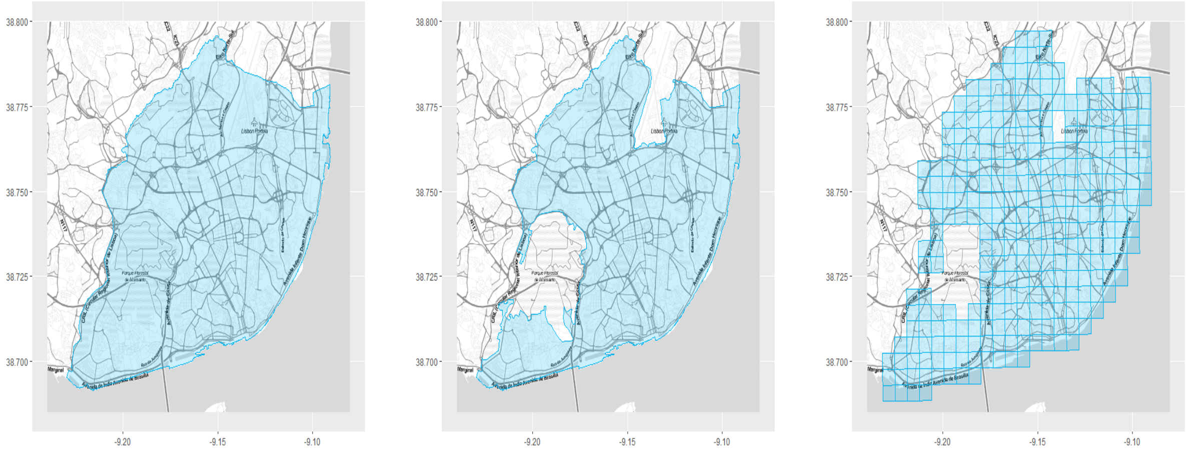


Figure 5.2: Spatial arrangement of the area

5.3 Statistical framework and methods

A multivariate model of infectious diseases surveillance data was considered that can be seen it as a branching process with immigration (Paul et al., 2008; Held et al., 2005). In the context of surveillance setting, data are spatially and temporally aggregated, and there is no information available about the number of susceptibles per region (Meyer et al., 2017; Paul et al., 2008). Let $y_{i,t}$ be the number gathered of geolocated tweets in the i-th rectangle at the hour t, $i = 1, \dots, m$, $j = 1, \dots, T$. Following Paul et al. (2008), those counts are assumed negative binomial distributed (accounting for potential overdispersion), $y_{i,t}|y_{i,t-1} \sim \text{NegBin}(\mu_{i,t}, \psi)$ with conditional mean:

$$\mu_{i,t} = \underbrace{\lambda_i y_{i,t-1} + \phi_i \sum_{j\neq i} w_{ij} y_{j,t-l}}_{\text{Epidemic component}} + \underbrace{\alpha_i + \sum_{s=1}^{S_i}(\gamma_{i,s} \sin(\omega_s t) + \delta_{i,s} \cos(\omega_s t))}_{\text{Endemic component}} \tag{5.1}$$

where λ_i represents the autoregressive parameter for the i-th rectangle, ϕ_i quantifies the influence of the counts between connected regions, and w_{ij} are weights defined as a power law of the adjacency order o_{ji} between zones $w_{ij} = o_{ji}^{-d}$ for $i \neq j$ and $w_{jj} = 0$ to consider that humans travel through metropolitan areas (Meyer and Held, 2014). Additionally, S_i are the number of harmonics to include and ω_s are Fourier frequencies, e.g. $\omega_s = 2\pi s/24$ for hourly data. The parameter α_i allows different incidence level in the regions. While *epidemic component* reflects occasional outbreaks (for instance, anomalous situations caused by massive events), *endemic* *component* shows the baseline rate with a regular temporal pattern. A particular case appears for the overall time series, which is called *univariate model* where the general form presented in the equation (3.7) can be written in the following way:

$$\mu_t = \underbrace{\lambda y_{t-1}}_{\text{Epidemic component}} \underbrace{+ \alpha + \delta_{d(t)} + \eta_{h(t)} + \sum_{s=1}^{S} (\gamma_s \sin(\omega_s t) + \delta_s \cos(\omega_s t))}_{\text{Endemic component}} \tag{5.2}$$

where $d(t)$ identifies the day-of-the week and $h(t)$ the hour-of-the-day to reflect different incidence levels.

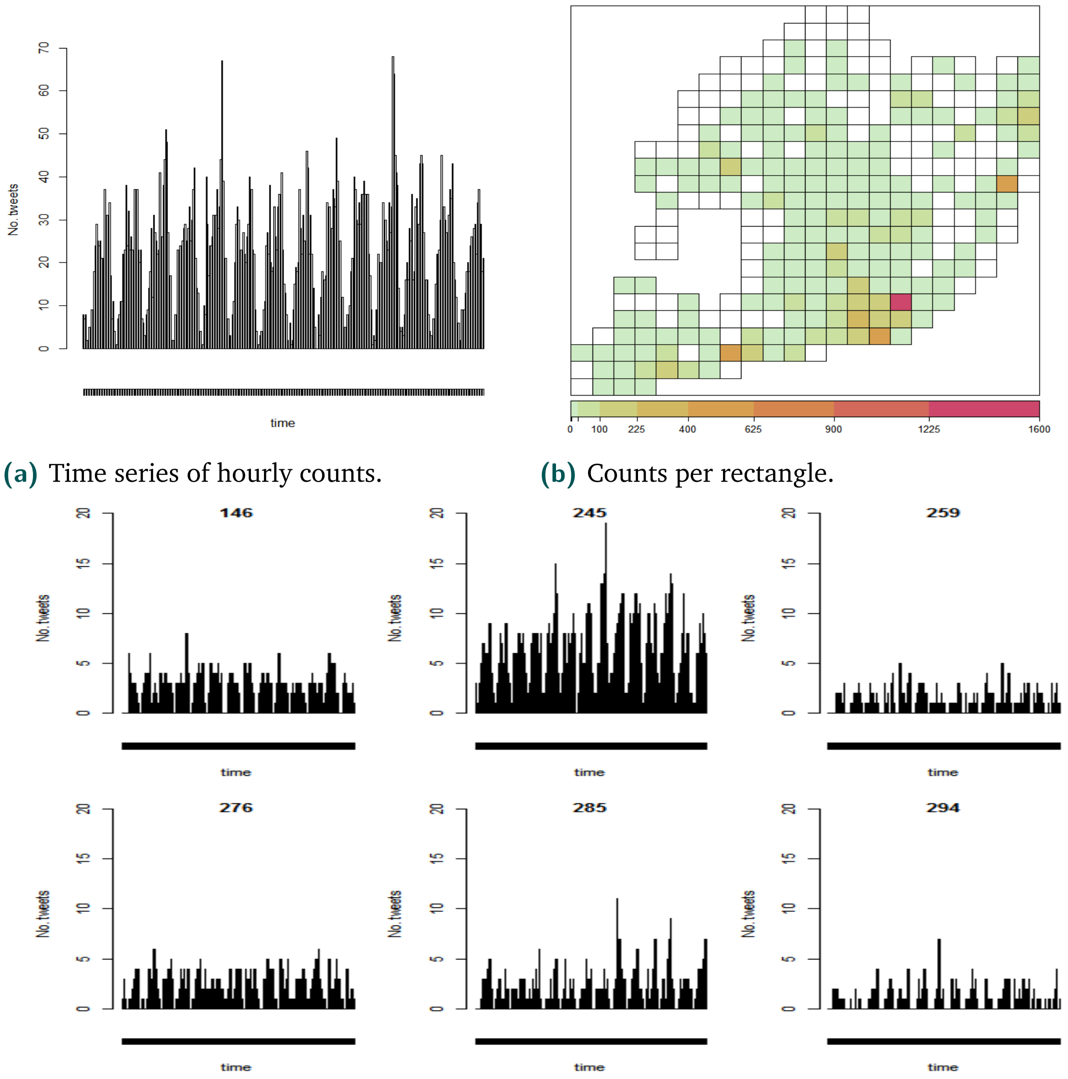
5.4 Results

During the whole period, 14500 geolocated tweets where collected inside of bounding box of which 8514 were created in the city of Lisbon. The statistical analysis was conducted considering tweets generated between 2017-07-30 00:00:00 and 2017-08-12 23:59:59, establishing a temporal horizon of 336 hours. Scott's method provided bandwidths of 524m for x-axis and 563m for y-axis respectively. Based on that, the city was divided into 302 rectangles of such size.

Figure 5.3 shows the visualisation of the data. The overall time-series plot in Figure 5.3a revealed a marked seasonality, peaking between 19:00 and 23:00 and reaching its minimum after midnight till early morning. There was a significant increase of geolocated tweets in the evenings of the sixth and eleventh days. The spatial plot in Figure 5.3b indicates the spatial clustering of the events over the river margin, especially in the south-east direction where the city centre is located. Finally, Figure 5.3c presents the individual time series for the rectangles with more than 200 tweets, which also exhibit seasonal behaviour and heterogeneity in the number of tweets.

The model was fitted using a stepwise procedure based according to Bayesian Information Criterion (BIC). Finally, it included linear terms for the autoregressive (epidemic) component and overall-trend, sine-cosine pairs corresponding to 24-hour, 12-hour, and 6-hour cycle lengths, dummy variables for Tuesday, Wednesday, and Thursday, and dummy variables for hours between 6:00 and 15:00 (*see* Table 5.2). Day-of-the week was positively correlated with twitter activity whereas the correlation with hour-of-the-day was negative. On the other hand, the autoregressive part suggested a high-pressure for increasing the number of counts in hours preceded hours for a low social media activity. Figure 5.4 compares observed and fitted numbers of geolocated tweets over the observation period.

The multivariate (spatio-temporal) infectious disease surveillance model was estimated using the same procedure as in the univariate case (*see* Table 5.3). Although it was considered terms for the day-of-the-week and the hour-of-the-day in the endemic component, those were not statistically significant whereby they were deleted into the selected model. Then, the long-term trend was established by $S = 3$ harmonics with Fourier frequencies related to 24-hour, 12-hour, and 6-hour. In this case, unlike univariate way, the epidemic component consisted of the temporal autoregressive and the spatial neighbourhood parameters which have a statistical contribution



(a) Time series of hourly counts.

(b) Counts per rectangle.

(c) Count time series of the regions with more than 200 tweets.

Figure 5.3: Geolocated tweets in Lisbon

in the explanation of the phenomenon. The dominant eigenvalue was 0.79 that represents the epidemic proportion of disease incidence, i.e., a considerable part of the fitted mean of the counts comes from the inside-rectangle autoregressive component with a small contribution of activity of adjacent areas and a slightly low endemic incidence. The estimation of the decay parameter of the adjacency order d is approximately 2 meaning that spatial interaction is presented around 1km (size of two cells). Figure 5.5 compares observed and fitted counts over the observation period in the areas with more than 200 of geolocated tweets, it can be observed that the epidemic component give the highest contribution in comparison with the other elements of the model.

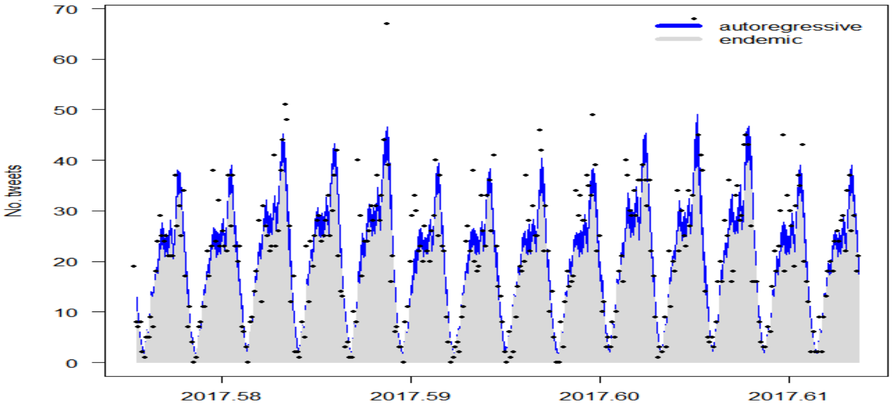


Figure 5.4: Observed temporal variation of geolocated tweets (black dots) together with the fitted variation from a univariate infectious disease surveillance model (grey-blue areas).

Table 5.2: Estimated regression coefficients, 95% confidence intervals, and p-values in the fitted univariate infectious disease surveillance model for the number of geolocated tweets per hour in Lisbon, Portugal.

Parameter	Estimate	95% CI	p-value
$\hat{\lambda}$	-1.87	(-2.51; -1.23)	1.05e-08
$\hat{\alpha}$	6.21	(5.14; 7.29)	0
Tuesday	0.17	(0.06; 0.27)	1.74e-03
Wednesday	0.15	(0.04; 0.25)	5.84e-03
Thursday	0.18	(0.07; 0.28)	7.80e-04
6:00	-3.21	(0.8; 1.67)	8.06e-11
7:00	-6.19	(-4.18; -2.24)	1.56e-09
8:00	-10.6	(-8.19; -4.18)	2.14e-10
9:00	-13.72	(-13.88; -7.33)	2.55e-10
10:00	-15.37	(-17.97; -9.47)	1.37e-10
11:00	-14.32	(-20.06; -10.67)	1.86e-10
12:00	-11.5	(-18.72; -9.92)	1.89e-10
13:00	-7.89	(-15.04; -7.96)	8.65e-11
14:00	-4.15	(-10.27; -5.5)	1.48e-10
15:00	-1.57	(-5.42; -2.88)	1.60e-09
sin $S = 1$	1.99	(-2.07; -1.06)	6.82e-06
cos $S = 1$	-6.17	(1.12; 2.86)	5.11e-12
sin $S = 2$	-3.73	(-7.93; -4.42)	3.59e-13
cos $S = 2$	2.26	(-4.74; -2.73)	1.05e-10
sin $S = 3$	1.23	(1.57; 2.94)	2.07e-08
$\hat{\psi}$	0.03	(0.02; 0.04)	5.37e-06

5.5 Discussion

It was proposed a novelty approach for analysing social media data coming from Twitter based on statistical modelling for epidemic data. This alternative was able to describe the phenomenon adequately and give meaningful insights about how humans behave across time

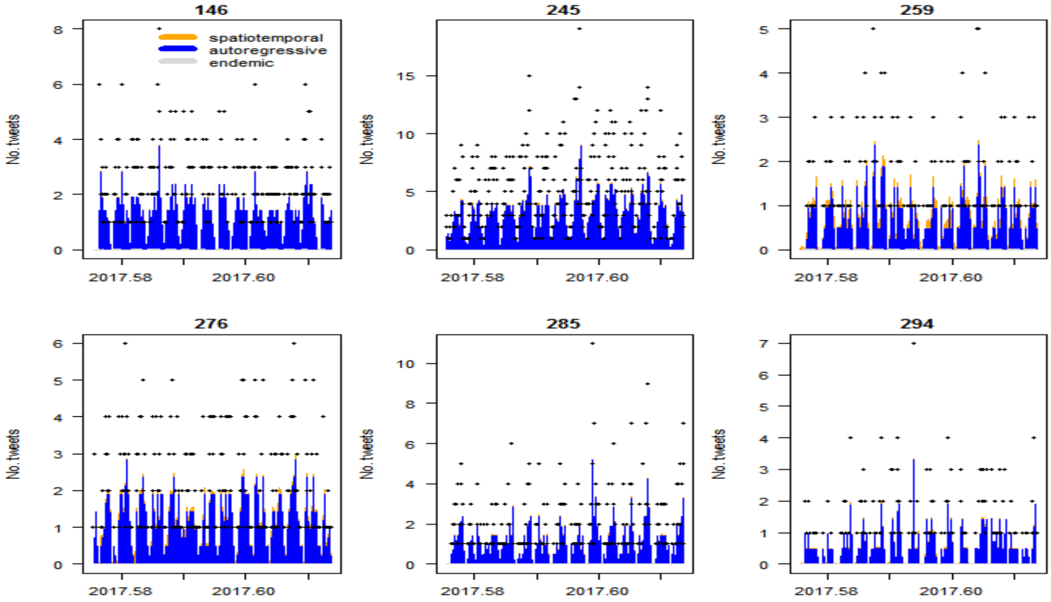


Figure 5.5: Observed temporal variation of geolocated tweets (black dots) together with the fitted variation from a multivariate infectious disease surveillance model (grey-blue-orange areas) in the rectangles with more than 200 geolocated tweets.

Table 5.3: Estimated regression coefficients, 95% confidence intervals, and p-values in the fitted multivariate infectious disease surveillance model for the number of geolocated tweets per hour in Lisbon, Portugal.

Parameter	Estimate	95% CI	p-value
$\hat{\lambda}$	-0.76	(-0.83; -0.68)	0
$\hat{\phi}$	-1.15	(-1.22; -1.07)	0
$\hat{\alpha}$	-4.73	(-4.9; -4.55)	0
sin $S = 1$	-0.11	(-0.25; 0.03)	0.13
cos $S = 1$	-0.85	(-1.07; -0.63)	2.82e-14
sin $S = 2$	-0.27	(-0.41; -0.12)	2.65e-04
cos $S = 2$	-0.5	(-0.68; -0.32)	5.19e-08
sin $S = 3$	-0.21	(-0.35; -0.08)	1.88e-03
cos $S = 3$	-0.13	(-0.28; 0.01)	0.07
d	1.98	(1.83; 2.12)	0
$\hat{\psi}$	3.32	(3.04; 3.61)	0

and space. The fitted models included: (1) an essential seasonal element of a 24-hour cycle, that has been highlighted in previous research about human activity, (2) an autoregressive component to indicate that hours with low content-generated precede hours with more usage of the social network, and (3) a spatial element to reflect how the amount of activity is affected by the behaviour in neighbouring areas with a radius of interaction is approximately 1km which also agrees with the referred as the distance travelled by people.

6 Urban human mobility patterns in origin-destination systems using functional data analysis and hierarchical clustering

6.1 Overview

We show a methodological approach to discover and describe human mobility patterns in origin-destination urban public transport systems. These systems register the entries and exits of the users to the system stations through smart cards. Conventionally, this information is transformed into a matrix called spatial interaction or origin-destination matrix where each of its positions represents the number of trips that start in one place and ends in another. Data analysis is oriented to understand how the systems work through establishing the demand of the origins and the attractiveness of the destinations.

Historically, origin-destination matrices were estimated by using survey sampling techniques, which were slow and expensive and restricted their constant updating. Nowadays, the use of sensors in the access points of public transport systems makes possible to collect large amounts of data that describe the mobility of users through the network of stations in the system. This availability of information requires for developing data analysis techniques that allow discovering, understanding, describing, and monitoring human mobility patterns, as well as, providing tools to establish adequate management plans for transport services and provide better travel time for passengers.

6.2 Method

6.2.1 Data

To test our method, we had access to the information of the registered transactions in the public transport system of Lisbon, Portugal in May 2015. We restricted the analysis to trips made in the metro network of the city. At that time, the system had 49 stations, which are distributed in 4 lines identified with colours, yellow, blue, green, and red (*see* Figure 6.1). The system operates every day between 6:30 am and 1:00 am. However, the frequency of service and the number of wagons of each train decrease on weekends and holidays. Users enter and leave the stations using a public transport card.

We carried out a sequential procedure for cleaning and preprocessing the information to dispose of a database in the required format to perform statistical analysis and modelling. We implemented such process using Microsoft SQL Server, and it consisted of the following steps:

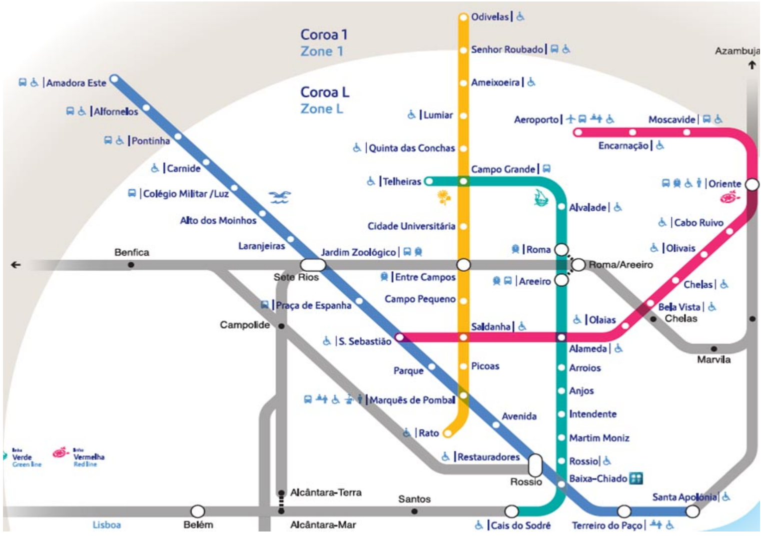


Figure 6.1: Map of Lisbon's subway.

- Only transactions made in the Lisbon subway were selected.
- All inbound to the subway that had an associated outbound were selected.
- It was verified that the exit time of the system happened after the entry time.
- Those transactions registered outside of the service hours of the metro and those carried out on weekends and holidays were eliminated.
- The total number of made outflows from each station of the system in fifteen minutes slots was counted.

6.2.2 Statistical analysis

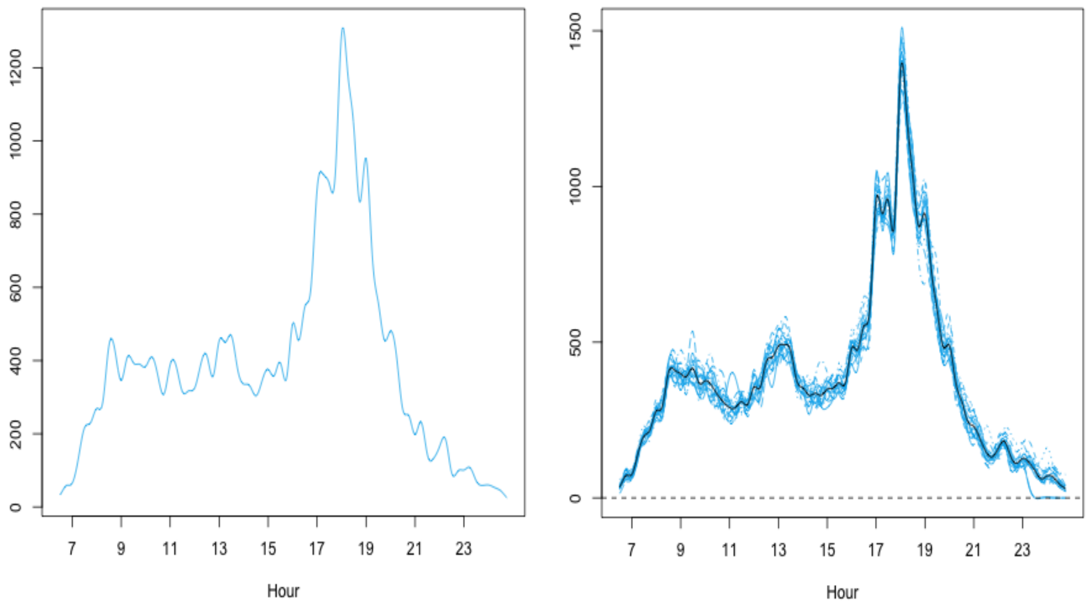
Data analysis consists of four main aspects. The first two are to establish the mobility pattern of each station based on the daily information of the number of trips that start there, while the last two are to define the overall behaviour of the system, by grouping stations that have similar temporal behaviours in the demand of the service. Thus, for each station, a functional representation of the daily number of trips is initially constructed, and then, such daily curves are summarised through the use of the functional mean, which we interpreted as the mobility pattern of said station. This process generates a characteristic curve for each station. These summary curves are reduced through the functional principal components analysis and the generated scores generated are classified by using hierarchical clustering methods.

6.3 Results

Lisbon's public transport system registered 47,101,706 transactions of which 20,968,691 were in the metro. The information had 10,542,403 entries and 10,426,288 exits. We verified the consistency of the information and found that 10,312,211 origin-destination flows. We ruled out 204,503 flows that were registered outside of the operational time. We also removed the flows of the Labor Day (May 1st) and Saturdays and Sundays (May 2nd, 3rd, 9th, 10th, 16th, 17th, 23rd, 24th, 30th, and 31st). Additionally, the Lisbon subway did not report transactions on May 19th, and 26th. Thus, to each station got 18 daily time series with the number of trips that started there in slot-times of 15-minutes.

For each station and each day, we obtained the functional representation of the corresponding curve to the number of trips by using 76 functional blocks a roughness penalty with a smoothing parameter $\lambda = 0.000001$. We then calculated the functional mean of the 18 obtained curves. The functional mean represents the mobility pattern of the station. Thus, we got 49 curves or individual mobility patterns of all stations. To establish the aggregate mobility patter, we got the functional expression of those 49 means. We after calculated the FPCA over the smoothed curves and kept the first two principal scores since they cumulated more than 70% of the variability. Additionally, to disclose more significant components of variation, we rotated the functional principal components with the VARIMAX rotation algorithm (Ramsay et al., 2009). We finally got the dendrogram by applying agglomerative hierarchical clustering through Ward's method on the matrix of dissimilarities computed with the Euclidian distance between the scores.

To illustrate our approach, we show in the Figure 6.2 the obtained functional representation of the number of trips in a particular station of the system. Panel (a) belongs for one specific time series, whilst, panel (b) displays the behaviour of the curves for all period and in colour black la functional mean or mobility pattern in that place. Thus, we found that Marques de Pombal station behaves in a unimodal mobility pattern with high demand iof the services between 17:00 and 21:00 reaching its maximum demand at 18:00.



(a) During a particular day

(b) During all observed period

Figure 6.2: The functional representation of inflows counts at Marques de Pombal station.

Figure 6.3 contains the plot of the smoothed curves or mobility patterns for each of 49 stations of the system. It shows that in general there two peaks of the demand of the services. First one occurs from 07:00 to 10:00 in the morning and between 17:00 to 20:00 in the evening. The results of the FPCA are displayed in the Figure 6.4. Panel (a) plots the components and panel (b) the scores for each station. The first two principal components explain 61.9% and 36.3% of the variability of the number of flows, respectively. The first harmonic portrays a decrease in the variation of the mean function for the flows that happen early in the morning up to 10:00 and a constant increase for the rest of the operation. This result implies that the number flows are more homogeneous in the morning and more heterogeneous in the evening.

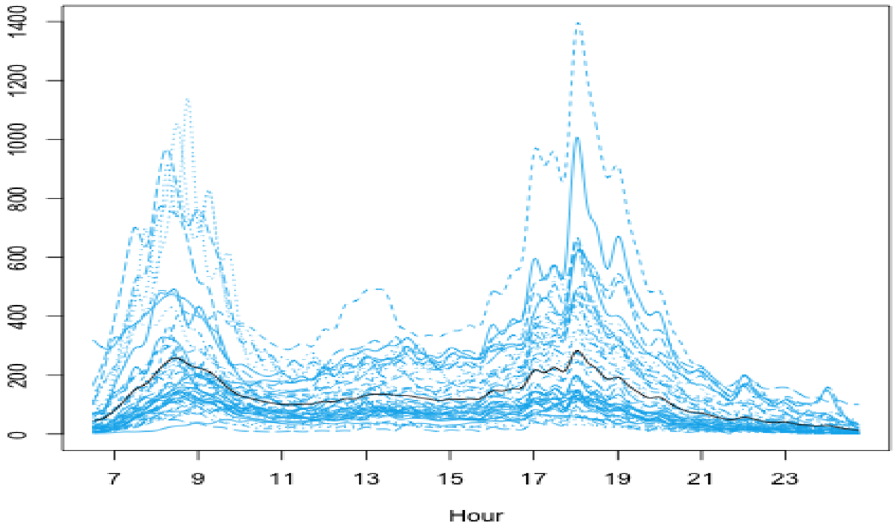
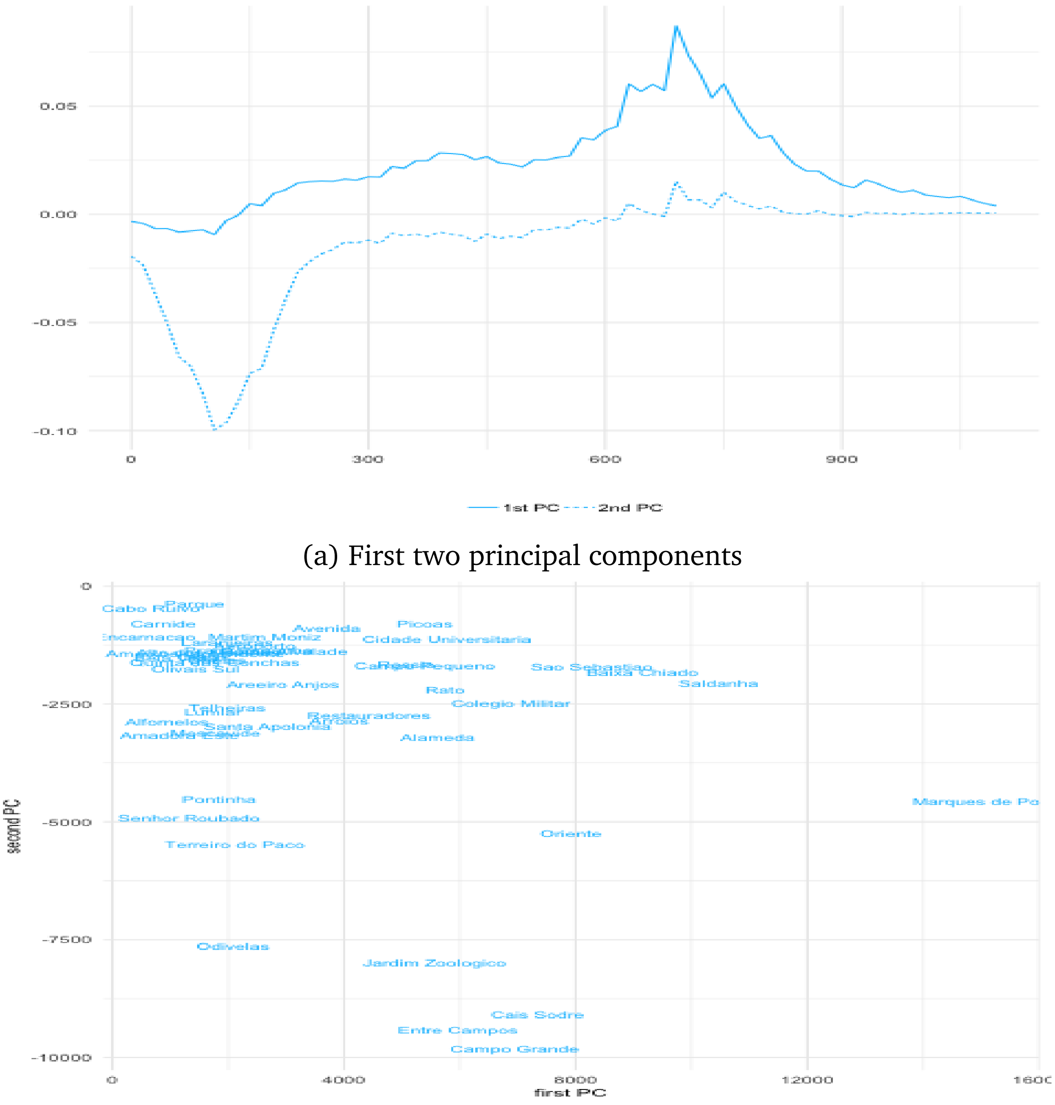


Figure 6.3: The functional representation of inflows counts at the Lisbon subway network.

Figure 6.5 and table 6.1 present the results of the hierarchical clustering of the functional scores. They shows that stations of the Lisbon subway can be classified into six groups according with the amount of demand of the service. Particularly, Figure 6.6 plots the mobility pattern of each group. We found mainly three summary behaviours: (1) Unimodal with high demand of system in the morning (clusters 2 and 5), (2) Unimodal with high demand of the system in the evening (Cluster 3, 4, and 6), and (3) Bimiodal with high demand in the morning and in the evenings.



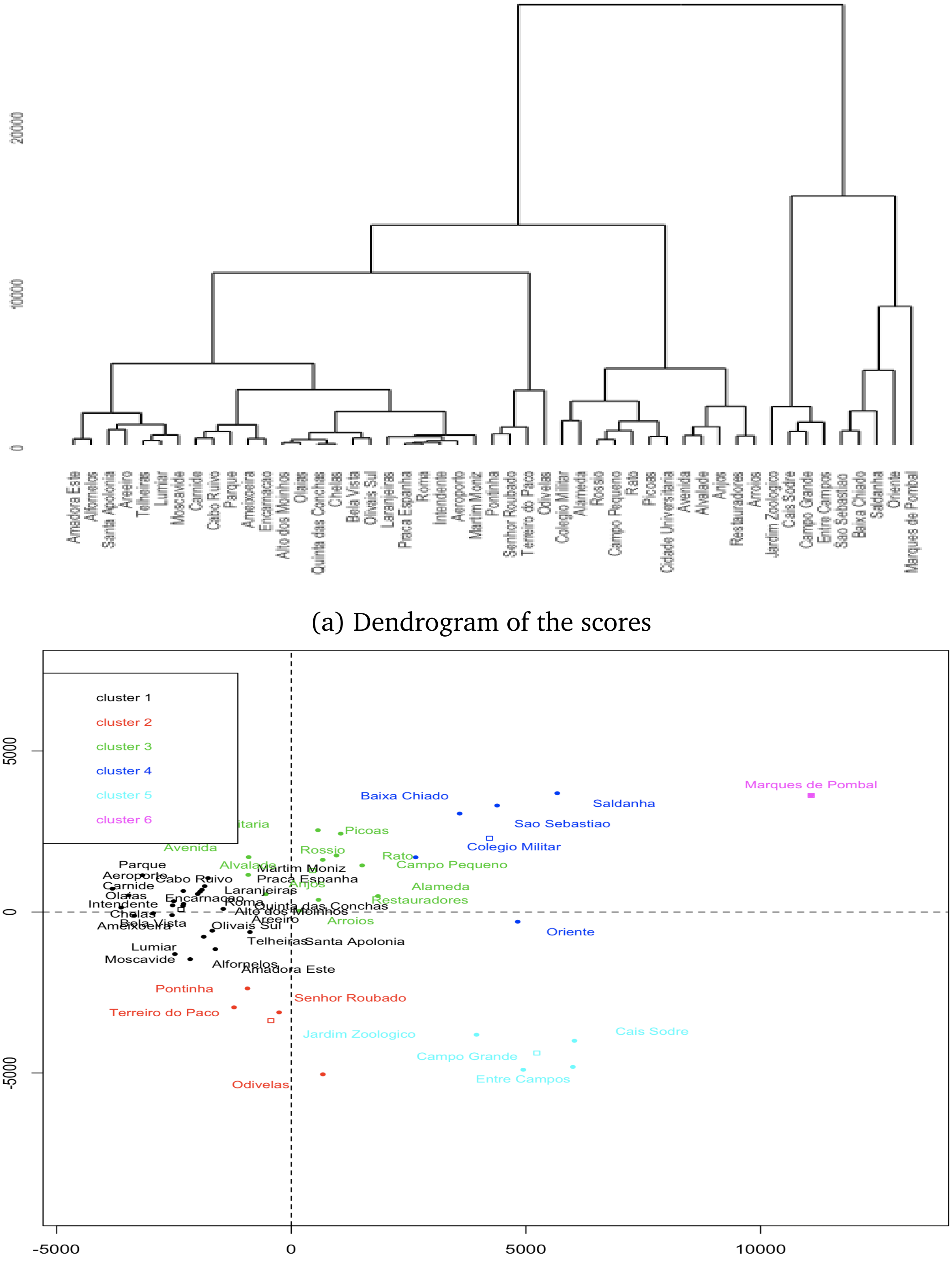
(a) First two principal components

(b) Plot of the scores for all stations

Figure 6.4: Results from a FPCA of inflows counts at the Lisbon subway network.

Table 6.1: Clusters of stations based on temporal patterns of flows

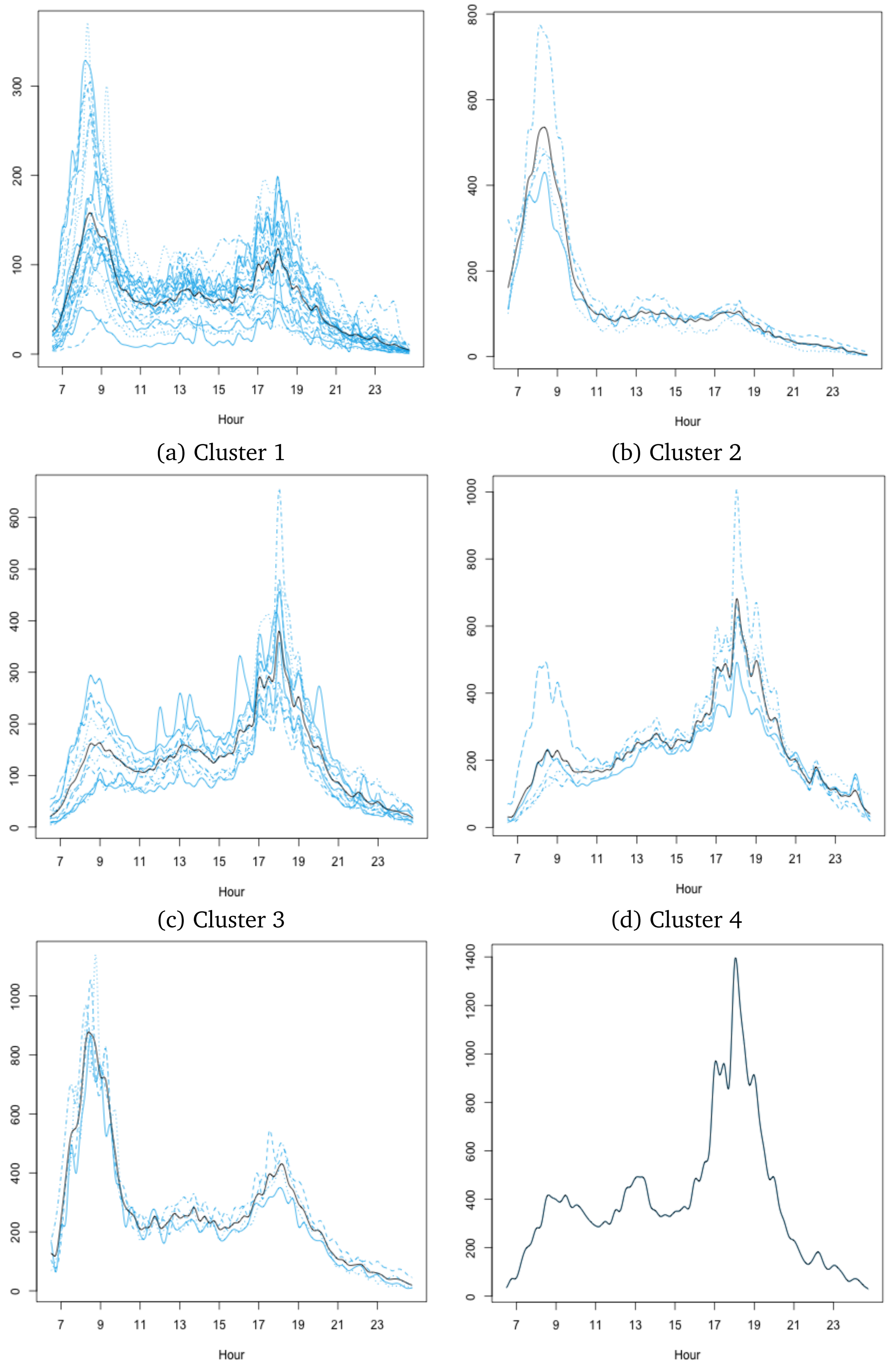
Cluster	Stations
1	Amadora Este, Alfornelos, Carnide, Alto dos Moinhos, Laranjeiras, Praca Espanha, Parque, Santa Apolonia, Martim Moniz, Intendente, Areeiro, Roma, Telheiras, Quinta das Conchas, Lumiar, Ameixoeira, Olaias, Bela Vista, Chelas, Olivais Sul, Cabo Ruivo, Moscavide, Encarnacao, Aeroporto
2	Pontinha, Terreiro do Paco, Senhor Roubado, Odivelas
3	Avenida, Restauradores, Rossio, Anjos, Arroios, Alameda, Alvalade, Rato, Picoas, Campo Pequeno, Cidade Universitaria
4	Colegio Militar, São Sebastião, Baixa Chiado, Saldanha, Oriente
5	Jardim Zoológico, Cais Sodré, Entre Campos, Campo Grande
6	Marques de Pombal



(a) Dendrogram of the scores

(b) Plot of clusters

Figure 6.5: Results from a hierarchical clustering of inflows counts at the Lisbon subway network.



(a) Cluster 1

(b) Cluster 2

(c) Cluster 3

(d) Cluster 4

(e) Cluster 5

(f) Cluster 6

Figure 6.6: Clusters of stations based on temporal patterns of flows

Conclusions 7

This dissertation reviewed relevant literature in the field of human dynamics and summarised several components in its study, as well as reflections on the statistical methods to find urban patterns. Discussions on Chapters 4 and 5 have been done on the potential opportunities and impact of the two main statistical approaches suggested for the use of social media data and epidemic-like data for studying spatio-temporal distributions related to the understanding of human urban activity. These have particularly reflected on how the proposed methods address some of the current gaps in the field of statistical modelling of human activity regarding monitoring, modelling, and predicting urban human dynamics patterns on a large-scale level. Furthermore, the statistical approaches here developed resource to the use of data fusion techniques to integrate information from diverse smart city´s sources as a way to improve the goodness of fit of models for pattern discovery.

7.1 Spatio-temporal distribution, social media data, and human activity

This dissertation proposed an alternative to studying the spatio-temporal distribution of geolocated tweets through the use of statistical methods for answering several questions. It first focused on developing a meaningful approach to analyse a considerable amount of human-generated data. In that direction, it was found that regression modelling, spatial point patterns, FDA, and hierarchical methods can process data coming from social networks and discover and describe regularities in the distribution of the human activities in the cites.

Additionally, this work aimed to characterise temporal and spatial structures in the data to capture the spatio-temporal behaviour of humans. The evaluated techniques showed that temporal trends and spatial autocorrelation are relevant to improve the goodness of fit in regression methods and adequately describe the spatial distribution of the places where people interact with social media, respectively. The analytical proposal was tested in three different cities to characterise and compare behaviours across those urban environments which allowed to gain information involving human conduct in cities with different structures and dynamics.

However, the presented statistical approach and its application had some limitations. One of the most relevant limitations referred to the analysis of content data which has provided meaningful insights into the field of Twitter Analytics. The inclusion of semantics components of geolocated tweets in our approach would give additional information about the cities.

Nonetheless, the previous point was not covered by the research here presented since its core objective was to understand the spatio-temporal distribution of where humans interact with their social networks as a proxy for human activity and avoid constraints related to the semantic analysis. The two main arguments for not considering semantics analysis as part of the approaches here developed were as follows. First, Twitter streaming API provides human-generated content that is not only text, which reduces the amount of data for analysing in procedures of sentiment analysis, and thus, representativeness of its results, for example, we found the more than 50% of the geotagged tweets came from third parties such as Instagram and Foursquare, among others. Second, the computational expense of text mining methods the problems in the processing of textual information due to UTF-8 characters, study more than one language, and hashtag parsing (Huang et al., 2018). And third, the privacy concerns associated with the identification of the users (Tasse and Hong, 2014; Gao and Liu, 2014; Frias-Martinez et al., 2012).

Additional limitations on this research study related to:

- The complex task of identifying and eliminating shared content by machines, bots, cyborgs and other sources who are not people which restricts the analysis. A
- The short study period and the season when was located can distort the found patterns.
- The representativeness of the harvested sample through the Twitter API which is only around one per cent of the overall activity.
- Finally, estimated models can vary greatly depending on population number, social structure, ethnicity, culture, traditions, and consumer preferences. Many control variables can enter modelling.

7.2 Spatio-temporal distribution, epidemic data, and human activity

Despite the above limitations, we have found that our approach was able to identify almost the same behaviours in a more straightforward way than alternatives developed in previous research. On the other hand, our proposal is looking for providing easily implemented and reproducible methods that can be automatised and thus, analyse a significant amount of geolocated data with the advantage of using more advanced techniques. Moreover, we included and statistically tested the effect of considering structures of spatio-temporal autocorrelation that might allow for predicting, monitoring and, simulating the activities accurately in the cities. For example, the inclusion of autoregressive parameters permits anticipates abnormal situations due to the pressure that immediate changes produce in the short-term forecasts.

Bibliography

Amini, A., Kung, K., Kang, C., Sobolevsky, S., and Ratti, C. (2014). The impact of social segregation on human mobility in developing and industrialized regions. *EPJ Data Science*, 3(1):1–20.

Anbaroglu, B., Heydecker, B., and Cheng, T. (2014). Spatio–temporal clustering for non–recurrent traffic congestion detection on urban road networks. *Transportation Research Part C: Emerging Technologies*, 48:47–65.

Aragó, P., Juan, P., and Staab, J. (2018). *tweet2r: Twitter Collector for R and Export to 'SQLite', 'postGIS' and 'GIS' Format*. R package version 1.1.

Arbia, G. and Petrarca, F. (2013). Effects of scale in spatial interaction models. *Journal of Geographical Systems*, 15(3):249–264.

Baddeley, A., Rubak, E., and Turner, R. (2015). *Spatial point patterns: methodology and applications with R*. CRC Press.

Bagrow, J. and Lin, Y.-R. (2012). Mesoscopic structure and social aspects of human mobility. *PloS one*, 7(5):e37676.

Bailey, T. and Gatrell, A. (1995). *Interactive spatial data analysis*. Longman Scientific & Technical Essex.

Bakci, T., Almirall, E., and Wareham, J. (2012). A smart city initiative: The case of barcelona. *Journal of the Knowledge Economy*, 4(2):135–148.

Bakerman, J., Pazdernik, K., Wilson, A., Fairchild, G., and Bahran, R. (2018). Twitter geolocation. *ACM Transactions on Knowledge Discovery from Data*, 20(3):1–17.

Batrinca, B. and Treleaven, P. C. (2014). Social media analytics: a survey of techniques, tools and platforms. *AI & SOCIETY*, 30(1):89–116.

Batty, M. (2009). Cities as complex systems: Scaling, interaction, networks, dynamics and urban morphologies. In Meyers, R. A., editor, *Encyclopedia of complexity and systems science*, pages 1041–1071, New York, NY. Springer.

Beckmann, M. J. (1967). On the theory of traffic flow in networks. *Traffic Quarterly*, 21(1).

Bettencourt, L. M. (2013). The origins of scaling in cities. *Science*, 340(6139):1438–1441.

Bradley, C. A., Rolka, H., Walker, D., and Loonsk, J. (2005). Biosense: implementation of a national early event detection and situational awareness system. *MMWR Morb Mortal Wkly Rep*, 54(Suppl):11–19.

Brockmann, D. (2012). Complex systems: Spotlight on mobility. *Nature*, 484(7392):40–41.

Brockmann, D., Hufnagel, L., and Geisel, T. (2006). The scaling laws of human travel. *Nature*, 439(7075):462–465.

Brugere, I., Gunturi, V., and Shekhar, S. (2014). Modeling and analysis of spatiotemporal social networks. In Alhajj, R. and Rokne, J., editors, *Encyclopedia of Social Network Analysis and Mining*, pages 950–960. Springer.

Buehler, J. W., Berkelman, R. L., Hartley, D. M., and Peters, C. J. (2003). Syndromic surveillance and bioterrorism-related epidemics. *Emerging infectious diseases*, 9(10):1197.

Buscher, V., Doody, L., Webb, M., and Aoun, C. (2014). Urban mobility in the smart city age. Technical report, ARUP, http://publications.arup.com/Publications/U/.

Cameron, A. C. and Trivedi, P. K. (1986). Econometric models based on count data. comparisons and applications of some estimators and tests. *Journal of applied econometrics*, 1(1):29–53.

Castro, P., Zhang, D., and Li, S. (2012). Urban traffic modelling and prediction using large scale taxi gps traces. In Kay, J., Lukowicz, P., Tokuda, H., Olivier, P., and Krüger, A., editors, *Pervasive Computing*, pages 57–72. Springer.

Cats, O., Wang, Q., and Zhao, Y. (2015). Identification and classification of public transport activity centres in stockholm using passenger flows data. *Journal of Transport Geography*, 48:10–22.

Cebelak, M. (2013). Location–based social networking data: doubly–constrained gravity model origin–destination estimation of the urban travel demand for austin, tx. Master's thesis, The University of Texas at Austin.

Celikten, E., Falher, G. L., and Mathioudakis, M. (2017). Modeling urban behavior by mining geotagged social data. *IEEE Transactions on Big Data*, 3(2):220–233.

Chen, C.-C., Chiang, M.-F., and Peng, W.-C. (2015). Mining and clustering mobility evolution patterns from social media for urban informatics. *Knowledge and Information Systems*, pages 1–23.

Chen, K. and Müller, H.-G. (2012). Modeling repeated functional observations. *Journal of the American Statistical Association*, 107(500):1599–1609.

Chen, S., Claramunt, C., and Ray, C. (2014). A spatio–temporal modelling approach for the study of the connectivity and accessibility of the guangzhou metropolitan network. *Journal of Transport Geography*, 36:12–23.

Cheng, T. and Wicks, T. (2014). Event detection using twitter: a spatio-temporal approach. *PloS one*, 9(6):e97807.

Cho, E., Myers, S. A., and Leskovec, J. (2011). 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*, pages 1082–1090. ACM.

Chourabi, H., Nam, T., Walker, S., Gil-Garcia, J. R., Mellouli, S., Nahon, K., Pardo, T., and Scholl, H. (2012). Understanding smart cities: An integrative framework. In *2012 45th Hawaii International Conference on System Sciences*. Institute of Electrical & Electronics Engineers (IEEE).

Chua, A., Marcheggiani, E., Servillo, L., and Moere, A. V. (2015). Flowsampler: Visual analysis of urban flows in geolocated social media data. In Aiello, L. and McFarland, D., editors, *Social Informatics*, pages 5–17. Springer.

Colizza, V., Barrat, A., Barthélemy, M., and Vespignani, A. (2006). The role of the airline transportation network in the prediction and predictability of global epidemics. *Proceedings of the National Academy of Sciences of the United States of America*, 103(7):2015–2020.

Coscia, M., Rinzivillo, S., Giannotti, F., and Pedreschi, D. (2014). Spatial and temporal evaluation of network–based analysis of human mobility. In Can, F., Özyer, T., and Polat, F., editors, *State of the Art Applications of Social Network Analysis*, pages 269–293. Springer.

Cottrill, C., Pereira, F., Zhao, F., Dias, I., Lim, H., Ben-Akiva, M., and Zegras, P. (2013). Future mobility survey. *Transportation Research Record: Journal of the Transportation Research Board*, 2354:59–67.

Cox, W. (2014). Traffic congestion in the world: 10 worst and best cities. *New Geography*.

Cranshaw, J., Toch, E., Hong, J., Kittur, A., and Sadeh, N. (2010). Bridging the gap between physical location and online social networks. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pages 119–128. ACM.

Cressie, N. (1993). *Statistics for spatial data*. Wiley series in probability and mathematical statistics: Applied probability and statistics. J. Wiley.

Dale, M. and Fortin, M. (2014). *Spatial Analysis: A Guide For Ecologists*. Cambridge University Press.

de Albuquerque, J. P., Herfort, B., Brenning, A., and Zipf, A. (2015). A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management. *International Journal of Geographical Information Science*, 29(4):667–689.

De Domenico, M., Lima, A., González, M., and Arenas, A. (2015). Personalized routing for multitudes in smart cities. *EPJ Data Science*, 4(1):1–11.

de Montjoye, Y.-A., Hidalgo, C., Verleysen, M., and Blondel, V. (2013). Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, 3.

Delicado, P., Giraldo, R., Comas, C., and Mateu, J. (2010). Statistics for spatial functional data: some recent contributions. *Environmetrics: The official journal of the International Environmetrics Society*, 21(3-4):224–239.

Demšar, U. and Virrantaus, K. (2010). Space–time density of trajectories: exploring spatio–temporal patterns in movement data. *International Journal of Geographical Information Science*, 24(10):1527–1542.

Diab, A. and Mitschele-Thiel, A. (2014). Human mobility patterns. In Guo, B., Riboni, D., and Hu, P., editors, *Creating Personal, Social, and Urban Awareness through Pervasive Computing*, pages 245–273. IGI Global.

Diggle, P. J. (2013). *Statistical analysis of spatial and spatio-temporal point patterns*. CRC Press.

Dobson, A. J. and Barnett, A. G. (2008). *An Introduction to Generalized Linear Models*. Chapman & Hall/CRC Texts in Statistical Science, fourth edition edition.

Enemark, A. and Kneeshaw, S. (2013). Cities of tomorrow ??? action today. urbact ii capitalisation. how cities can motivate mobility mindsets. Technical report, URBAC.

Espín-Noboa, L., Lemmerich, F., Singer, P., and Strohmaier, M. (2016). Discovering and explaining mobility patterns in urban spaces: A study of manhattan taxi data. In *Proceedings of the 6th International Workshop on Location and the Web*, pages 186–194. ACM.

Ewing, R. and Cervero, R. (2001). Travel and the built environment: a synthesis. *Transportation Research Record: Journal of the Transportation Research Board*, 1780(1):87–114.

Ferrari, L., Rosi, A., Mamei, M., and Zambonelli, F. (2011). Extracting urban patterns from location–based social networks. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location–Based Social Networks*, pages 9–16. ACM.

Fischer, M. and Getis, A. (2010). *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications*. Springer.

Fischer, M. and Griffith, D. (2008). Modeling spatial autocorrelation in spatial interaction data: An application to patent citation data in the european union. *Journal of Regional Science*, 48(5):969–989.

Fischer, M., Reismann, M., and Scherngell, T. (2010). Spatial interaction and spatial autocorrelation. In Anselin, L. and Rey, S., editors, *Perspectives on spatial data analysis*, pages 61–79. Springer.

Fischer, M. and Wang, J. (2011). *Spatial data analysis: models, methods and techniques*. Springer Science + Business Media.

Forghani, M. and Karimipour, F. (2014). Extracting human behavioral patterns by mining geo–social networks. *ISPRS–International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 1:115–120.

França, U., Sayama, H., Mcswiggen, C., Daneshvar, R., and Bar-Yam, Y. (2015). Visualizing the "heartbeat" of a city with tweets. *Complexity*, 21(6):280–287.

Frias-Martinez, V. and Frias-Martinez, E. (2014). Spectral clustering for sensing urban land use using twitter activity. *Engineering Applications of Artificial Intelligence*, 35:237–245.

Frias-Martinez, V., Soto, V., Hohwald, H., and Frias-Martinez, E. (2012). Characterizing urban landscapes using geolocated tweets. In *Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust*, SOCIALCOM-PASSAT '12, pages 239–248, Washington, DC, USA. IEEE Computer Society.

Frühwirth-Schnatter, S. and Wagner, H. (2004). Data augmentation and gibbs sampling for regression models of small counts. *IFAS Research Paper Series*, 4.

Gabrielli, L., Rinzivillo, S., Ronzano, F., and Villatoro, D. (2014). From tweets to semantic trajectories: mining anomalous urban mobility patterns. In Nin, J. and Villatoro, D., editors, *Citizen in Sensor Networks*, pages 26–35. Springer.

Gandomi, A. and Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2):137–144.

Gao, H. and Liu, H. (2014). Data analysis on location-based social networks. In *Mobile Social Networking*, pages 165–194. Springer New York.

Gao, H. and Liu, H. (2015). Mining human mobility in location–based social networks. *Synthesis Lectures on Data Mining and Knowledge Discovery*, 7(2):1–115.

Gao, S. (2015). Spatio–temporal analytics for exploring human mobility patterns and urban dynamics in the mobile age. *Spatial Cognition & Computation*, 15(2):86–114.

García-Palomares, J. C., Salas-Olmedo, M. H., Moya-Gómez, B., Condeço-Melhorado, A., and Gutiérrez, J. (2018). City dynamics through twitter: Relationships between land use and spatiotemporal demographics. *Cities*, 72:310–319.

Giannotti, F., Pappalardo, L., Pedreschi, D., and Wang, D. (2013). A complexity science perspective on human mobility. In Renso, C., Spaccapietra, S., and Zimányi, E., editors, *Mobility Data: Modeling, Management, and Understanding*, pages 297–314. Cambridge University Press.

González, M., Hidalgo, C., and Barabasi, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, 453(7196):779–782.

Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4):211–221.

Gorzelany, J. (2013). The world's most traffic–congested cities. *Forbes.*

Griffith, D. (1992). What is spatial autocorrelation? reflections on the past 25 years of spatial statistics. *Espace géographique*, 21(3):265–280.

Griffith, D. and Chun, Y. (2013). Spatial autocorrelation and spatial filtering. In Fischer, M. and Nijkamp, P., editors, *Handbook of regional science*, pages 1477–1507. Springer Heidelberg.

Griffith, D. and Fischer, M. (2013). Constrained variants of the gravity model and spatial dependence: model specification and estimation issues. *Journal of Geographical Systems*, 15(3):291–317.

Guo, D. (2009). Flow mapping and multivariate visualization of large spatial interaction data. *IEEE Transactions on Visualization and Computer Graphics*, 15(6).

Hanson, S. and Huff, O. J. (1988). Systematic variability in repetitious travel. *Transportation*, 15(1-2):111–135.

Hardin, J. W. and Hilbe, J. M. (2012). *Generalized linear models and extensions.* Stata press.

Hasan, S., Schneider, C., Ukkusuri, S., and González, M. (2012). Spatiotemporal patterns of urban human mobility. *Journal of Statistical Physics*, 151:304–318.

Hasan, S., Zhan, X., and Ukkusuri, S. V. (2013). Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. In *Proceedings of the 2nd ACM SIGKDD international workshop on urban computing*, page 6. ACM.

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., and Ratti, C. (2014). Geo–located twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, 41(3):260–271.

Hawelka, B., Sitko, I., Kazakopoulos, P., and Beinat, E. (2015). Collective prediction of individual mobility traces with exponential weights. *arXiv preprint arXiv:1510.06582.*

Held, L., Höhle, M., and Hofmann, M. (2005). A statistical framework for the analysis of multivariate infectious disease surveillance counts. *Statistical modelling*, 5(3):187–199.

Herrera-Yagüe, C., Schneider, C. M., Couronné, T., Smoreda, Z., Benito, R. M., Zufiria, P. J., and González, M. C. (2015). The anatomy of urban social networks and its implications in the searchability problem. *Scientific reports*, 5.

Hilbe, J. M. (1993). Log negative binomial regression as a generalized linear model. *Graduate College Committee on Statistics*, 1024.

Huang, Q. (2016). Mining online footprints to predict user's next location. *International Journal of Geographical Information Science*, 31(3):523–541.

Huang, Q. and Wong, D. W. S. (2016). Activity patterns, socioeconomic status and urban spatial structure: what can social media data tell us? *International Journal of Geographical Information Science*, 30(9):1873–1898.

Huang, W. and Li, S. (2016). Understanding human activity patterns based on space-time-semantics. *ISPRS Journal of Photogrammetry and Remote Sensing*, 121:1–10.

Huang, Y., Li, Y., and Shan, J. (2018). Spatial-temporal event detection from geo-tagged tweets. *ISPRS International Journal of Geo-Information*, 7(4):150.

Husson, F., Lê, S., and Pags, J. (2017). *Exploratory Multivariate Analysis by Example Using R*. CHAPMAN & HALL/CRC COMPUTER SC. CRC Press.

Hyman, G. M. (1969). The calibration of trip distribution models. *Environment and Planning*, 1(3):105–112.

Hyndman, R. J. and Booth, H. (2008). Stochastic population forecasts using functional data models for mortality, fertility and migration. *International Journal of Forecasting*, 24(3):323–342.

Hyndman, R. J. and Shang, H. L. (2018). *ftsa: Functional Time Series Analysis*. R package version 4.9.

Hyndman, R. J. and Ullah, M. S. (2007). Robust forecasting of mortality and fertility rates: a functional data approach. *Computational Statistics & Data Analysis*, 51(10):4942–4956.

Illian, J., Benson, E., Crawford, J., and Staines, H. (2006). Principal component analysis for spatial point processes - assessing the appropriateness of the approach in an ecological context. In *Case Studies in Spatial Point Process Modeling*, Lecture Notes in Statistics, pages 135–150. Springer New York.

Illian, J., Penttinen, A., Stoyan, H., and Stoyan, D. (2008). *Statistical analysis and modelling of spatial point patterns*, volume 70. John Wiley & Sons.

Jackson, M. C. (1985). Social systems theory and practice: The need for a critical approach. *International Journal Of General System*, 10(2-3):135–151.

Jiang, S., Ferreira, J., and González, M. C. (2012). Clustering daily patterns of human activities in the city. *Data Mining and Knowledge Discovery*, 25(3):478–510.

Kaplan, A. M. and Haenlein, M. (2010). Users of the world, unite! the challenges and opportunities of social media. *Business horizons*, 53(1):59–68.

Karamshuk, D., Boldrini, C., Conti, M., and Passarella, A. (2011). Human mobility models for opportunistic networks. *Communications Magazine, IEEE*, 49(12):157–165.

Karamshuk, D., Boldrini, C., Conti, M., and Passarella, A. (2014). Spot: Representing the social, spatial, and temporal dimensions of human mobility with a unifying framework. *Pervasive and Mobile Computing*, 11:19–40.

Katsouyanni, K., Schwartz, J., Spix, C., Touloumi, G., Zmirou, D., Zanobetti, A., Wojtyniak, B., Vonk, J., Tobias, A., Pönkä, A., Medina, S., Bachárová, L., and Anderson, H. R. (1996). Short term effects of air pollution on health: a european approach using epidemiologic time series data: the aphea protocol. *Journal of Epidemiology & Community Health*, 50(Suppl 1):S12–S18.

Kim, E., Helal, S., and Cook, D. (2010). Human activity recognition and pattern discovery. *IEEE Pervasive Computing/IEEE Computer Society [and] IEEE Communications Society*, 9(1):48–53.

Kirkpatrick, N. (2015). The world's most congested cities, by the numbers. *The Washington Post*.

Kokoszka, P. and Reimherr, M. (2017). *Introduction to Functional Data Analysis*. Chapman & Hall/CRC Texts in Statistical Science. CRC Press.

Lee, D.-J., Zhu, Z., and Toscas, P. (2015). Spatio-temporal functional data analysis for wireless sensor networks data. *Environmetrics*, 26(5):354–362.

Lenormand, M., Louail, T., Cantú-Ros, O., Picornell, M., Herranz, R., Arias, J., Barthelemy, M., San Miguel, M., and Ramasco, J. (2015). Influence of sociodemographic characteristics on human mobility. *Nature Scientific Reports*.

LeSage, J., Fischer, M., and Scherngell, T. (2007). Knowledge spillovers across europe: Evidence from a poisson spatial interaction model with spatial effects. *Papers Regional Science*, 86(3):393–421.

LeSage, J. and Pace, R. K. (2008). Spatial econometric modeling of origin–destination flows. *Journal of Regional Science*, 48(5):941–967.

LeSage, J. and Pace, R. K. (2009). *Introduction to spatial econometrics*. Chapman &Hall/CRC.

Liang, X., Zhao, J., Dong, L., and Xu, K. (2013). Unraveling the origin of exponential law in intra–urban human mobility. *Scientific reports*, 3.

Liboschik, T., Fokianos, K., and Fried, R. (2017). tscount: An r package for analysis of count time series following generalized linear models. *Journal of Statistical Software*, 82(5).

Llorente, A., Garcia-Herranz, M., Cebrian, M., and Moro, E. (2015). Social media fingerprints of unemployment. *PLOS ONE*, 10(5):e0128692.

Lloyd, C. D. (2014). *Exploring spatial scale in geography*. John Wiley & Sons.

Louail, T., Lenormand, M., Picornell, M., Cantú, O., Herranz, R., Frias-Martinez, E., Ramasco, J., and Barthelemy, M. (2015). Uncovering the spatial structure of mobility networks. *Nature communications*, 6.

Louail, T., Lenormand, M., Ros, O., Picornell, M., Herranz, R., Frias-Martinez, E., Ramasco, J., and Barthelemy, M. (2014). From mobile phone data to the spatial structure of cities. *Scientific reports*, 4.

Lu, X., Wetter, E., Bharti, N., Tatem, A., and Bengtsson, L. (2013). Approaching the limit of predictability in human mobility. *Scientific reports*, 3.

Maat, K., Van Wee, B., and Stead, D. (2005). Land use and travel behaviour: expected effects from the perspective of utility theory and activity-based theories. *Environment and Planning B: Planning and Design*, 32(1):33–46.

Marshall, D. and Staeheli, L. (2015). Mapping civil society with social network analysis: Methodological possibilities and limitations. *Geoforum*, 61:56–66.

Martínez-Camblor, P. and Corral, N. (2011). Repeated measures analysis for functional data. *Computational Statistics & Data Analysis*, 55(12):3244–3256.

Mateu, J. and Romano, E. (2016). Advances in spatial functional statistics. *Stochastic Environmental Research and Risk Assessment*, 31(1):1–6.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*, volume 37. CRC press.

Meyer, S. and Held, L. (2014). Power-law models for infectious disease spread. *The Annals of Applied Statistics*, 8(3):1612–1639.

Meyer, S., Held, L., and Höhle, M. (2017). Spatio-temporal analysis of epidemic phenomena using the r package surveillance. *Journal of Statistical Software*, 77(11).

Montgomery, D. C., Peck, E. A., and Vining, G. G. (2012). *Introduction to linear regression analysis*, volume 821. John Wiley & Sons.

Morstatter, F., Pfeffer, J., Liu, H., and Carley, K. M. (2013). Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. In *ICWSM*.

Murgante, B. and Borruso, G. (2012). Analyzing migration phenomena with spatial autocorrelation techniques. In Murgante, B., Gervasi, O., Misra, S., Nedjah, N., Rocha, A., Taniar, D., and Apduhan, B., editors, *Computational Science and Its Applications–ICCSA 2012*, volume 2, pages 249–262. Springer.

Myers, R. H., Montgomery, D. C., Vining, G. G., and Robinson, T. J. (2012). *Generalized linear models: with applications in engineering and the sciences*, volume 791. John Wiley & Sons.

Nanni, M., Trasarti, R., Furletti, B., Gabrielli, L., Van Der Mede, P., De Bruijn, J., De Romph, E., and Bruil, G. (2014). Transportation planning based on gsm traces: A case study on ivory coast. In Nin, J. and Villatoro, D., editors, *Citizen in Sensor Networks*, pages 15–25. Springer.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370.

Nguyen, T. and Szymanski, B. (2012). Using location–based social networks to validate human mobility and relationships models. In *Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on*, pages 1215–1221. IEEE.

Nin, J., Carrera, D., and Villatoro, D. (2014). On the use of social trajectory–based clustering methods for public transport optimization. In Nin, J. and Villatoro, D., editors, *Citizen in Sensor Networks*, pages 59–70. Springer.

Noulas, A. (2013). *Human urban mobility in location–based social networks: analysis, models and applications.* PhD thesis, University of Cambridge.

Noulas, A., Scellato, S., Mascolo, C., and Pontil, M. (2011). An empirical study of geographic user activity patterns in foursquare. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pages 570–573. Association for the Advancement of Artificial Intelligence.

Nummi, P. (2017). Social media data analysis in urban e-planning. *International Journal of E-Planning Research*, 6(4):18–31.

O'Sullivan, D. and Unwin, D. (2014). *Geographic Information Analysis.* Wiley.

Otte, E. and Rousseau, R. (2002). Social network analysis: a powerful strategy, also for the information sciences. *Journal of information Science*, 28(6):441–453.

Palchykov, V., Mitrović, M., Jo, H.-H., Saramäki, J., and Pan, R. (2014). Inferring human mobility using communication patterns. *Scientific reports*, 4.

Pan, G., Qi, G., Zhang, W., Li, S., Wu, Z., and Yang, L. (2013a). Trace analysis and mining for smart cities: issues, methods, and applications. *IEEE Communications Magazine*, 121.

Pan, W., Ghoshal, G., Krumme, C., Cebrian, M., and Pentland, A. (2013b). Urban characteristics attributable to density–driven tie formation. *Nature communications*, 4.

Pappalardo, L., Simini, F., Rinzivillo, S., Pedreschi, D., Giannotti, F., and Barabási, A.-L. (2015). Returners and explorers dichotomy in human mobility. *Nature communications*, 6.

Park, S. Y. and Staicu, A.-M. (2015). Longitudinal functional data analysis. *Stat*, 4(1):212–226.

Patel, N. N., Stevens, F. R., Huang, Z., Gaughan, A. E., Elyazar, I., and Tatem, A. J. (2016). Improving large area population mapping using geotweet densities. *Transactions in GIS*, 21(2):317–331.

Patuelli, R. and Arbia, G. (2013). Editorial: Advances in the statistical modelling of spatial interaction data. *Journal of Geographical Systems*, 15(3):229–231.

Paul, M., Held, L., and Toschke, A. M. (2008). Multivariate modelling of infectious disease surveillance data. *Statistics in medicine*, 27(29):6250–6267.

Pebesma, E. J. and Bivand, R. S. (2005). Classes and methods for spatial data in R. *R News*, 5(2):9–13.

Phithakkitnukoon, S., Horanont, T., Di Lorenzo, G., Shibasaki, R., and Ratti, C. (2010). Activity–aware map: Identifying human daily activity pattern using mobile phone data. In *Human Behavior Understanding*, pages 14–25. Springer.

Pinheiro, C. (2014). Revealing human mobility behavior and predicting amount of trips based on mobile data records. In *Proceedings of the 14th SAS Global Forum*. SAS.

Pirozmand, P., Wu, G., Jedari, B., and Xia, F. (2014). Human mobility in opportunistic networks: Characteristics, models and prediction methods. *Journal of Network and Computer Applications*, 42:45–58.

Prasetyo, P. K., Achananuparp, P., and Lim, E.-P. (2016). On analyzing geotagged tweets for location-based patterns. In *Proceedings of the 17th International Conference on Distributed Computing and Networking - ICDCN '16*. ACM Press.

Pressl, R. and Köllinger, C. (2012). Design and implementation of sustainable mobility campaigns. Technical report, Forschungsgesellschaft Mobilität – Austrian Mobility Research FGM–AMOR.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rae, A. (2009). From spatial interaction data to spatial interaction information? geovisualisation and spatial structures of migration from the 2001 UK census. *Computers, Environment and Urban Systems*, 33(3):161–178.

Ramsay, J., Hooker, G., and Graves, S. (2009). *Functional Data Analysis with R and MATLAB*. Use R! Springer New York.

Ramsay, J. and Silverman, B. (2005). *Functional Data Analysis*. Springer Series in Statistics. Springer New York.

Ranjan, G., Zang, H., Zhang, Z.-L., and Bolot, J. (2012). Are call detail records biased for sampling human mobility? *ACM SIGMOBILE Mobile Computing and Communications Review*, 16(3):33–44.

Ravenstein, E. G. (1885). The laws of migration. *Journal of the Statistical Society of London*, 48(2):167–235.

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

Ren, Y., Ercsey-Ravasz, M., Wang, P., González, M., and Toroczkai, Z. (2014). Predicting commuter flows in spatial networks using a radiation model based on temporal ranges. *Nature communications*, 5.

Resch, B., Summa, A., Zeile, P., and Strube, M. (2016). Citizen-centric urban planning through extracting emotion information from twitter in an interdisciplinary space-time-linguistics algorithm. *Urban Planning*, 1(2):114.

Resch, B., Usländer, F., and Havas, C. (2017). Combining machine-learning topic models and spatiotemporal analysis of social media data for disaster footprint and damage assessment. *Cartography and Geographic Information Science*, 45(4):362–376.

Robertson, C., Nelson, T. A., MacNab, Y. C., and Lawson, A. B. (2010). Review of methods for space–time disease surveillance. *Spatial and spatio-temporal epidemiology*, 1(2-3):105–116.

Robins, G. (2013). A tutorial on methods for the modeling and analysis of social network data. *Journal of Mathematical Psychology*, 57(6):261–274.

Roy, J. and Thill, J.-C. (2004). Spatial interaction modelling. In Florax, R. and Plane, D., editors, *Fifty Years of Regional Science*, pages 339–361. Springer Science + Business Media.

Saberi, M., Mahmassani, H. S., Brockmann, D., and Hosseini, A. (2016). A complex network perspective for characterizing urban travel demand patterns: graph theoretical analysis of large–scale origin–destination demand networks. *Transportation*, pages 1–20.

Sagl, G., Loidl, M., and Beinat, E. (2012). A visual analytics approach for extracting spatio–temporal urban mobility information from mobile network traffic. *ISPRS International Journal of Geo–Information*, 1(3):256–271.

Salmon, M., Schumacher, D., and Höhle, M. (2016). Monitoring count time series inR: Aberration detection in public health surveillance. *Journal of Statistical Software*, 70(10).

Scellato, S., Noulas, A., and Mascolo, C. (2011). Exploiting place features in link prediction on location–based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1046–1054. ACM.

Schlich, R. and Axhausen, K. (2003). Habitual travel behaviour: evidence from a six-week travel diary. *Transportation*, 30(1):13–36.

Scott, D. W. (2015). *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons.

Scott, J. (2012). *Social network analysis*. Sage, 2nd edition.

Shaw, S.-L., Tsou, M.-H., and Ye, X. (2016). Human dynamics in the mobile and big data era. *International Journal of Geographical Information Science*, 30(9):1687–1693.

Shi, Y., Deng, M., Yang, X., Liu, Q., Zhao, L., and Lu, C.-T. (2016). A framework for discovering evolving domain related spatio-temporal patterns in twitter. *ISPRS International Journal of Geo-Information*, 5(10):193.

Shittu, A., Shah, M., and Chiroma, M. (2015). Perception based determinants of mobility dilemma in ilorin metropolis. *Open Journal of Social Sciences*, 3(4):61–70.

Silva, T. H., de Melo, P. O. S. V., Almeida, J. M., and Loureiro, A. A. F. (2013). Social media as a source of sensing to study city dynamics and urban social behavior: Approaches, models, and opportunities. In *Ubiquitous Social Media Analysis*, pages 63–87. Springer Berlin Heidelberg.

Simini, F., González, M., Maritan, A., and Barabási, A.-L. (2012). A universal model for mobility and migration patterns. *Nature*, 484(7392):96–100.

Snijders, T. (2011). Statistical models for social networks. *Annual Review of Sociology*, 37:131–153.

Sohn, T., Varshavsky, A., LaMarca, A., Chen, M., Choudhury, T., Smith, I., Consolvo, S., Hightower, J., Griswold, W., and De Lara, E. (2006). Mobility detection using everyday gsm traces. In Dourish, P. and Friday, A., editors, *UbiComp 2006: Ubiquitous Computing*, pages 212–224. Springer.

Soliman, A., Soltani, K., Yin, J., Padmanabhan, A., and Wang, S. (2017). Social sensing of urban land use based on analysis of twitter users' mobility patterns. *PLOS ONE*, 12(7):e0181657.

Song, C., Koren, T., Wang, P., and Barabási, A.-L. (2010a). Modelling the scaling properties of human mobility. *Nature Physics*, 6(10):818–823.

Song, C., Qu, Z., Blumm, N., and Barabási, A.-L. (2010b). Limits of predictability in human mobility. *Science*, 327(5968):1018–1021.

Steenbruggen, J., Tranos, E., and Nijkamp, P. (2015). Data from mobile phone operators: A tool for smarter cities? *Telecommunications Policy*, 39(3-4):335–346.

Steiger, E., Resch, B., and Zipf, A. (2016). Exploration of spatiotemporal and semantic clusters of twitter data using unsupervised neural networks. *International Journal of Geographical Information Science*, 30(9):1694–1716.

Steiger, E., Westerholt, R., Resch, B., and Zipf, A. (2015). Twitter as an indicator for whereabouts of people? correlating twitter with UK census data. *Computers, Environment and Urban Systems*, 54:255–265.

Steinert-Threkeld, Z. C. (2018). *Twitter as Data*. Cambridge University Press.

Stimmel, C. L. (2015). *Building smart cities: analytics, ICT, and design thinking*. CRC Press.

Sun, J., Wang, Y., Si, H., Yuan, J., and Shan, X. (2010). Aggregate human mobility modeling using principal component analysis. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications*, 1(2/3):83–95.

Sun, L., Chen, C., and Zhang, D. (2014). Understanding urban dynamics from taxi gps traces. In Guo, B., Riboni, D., and Hu, P., editors, *Creating Personal, Social, and Urban Awareness through Pervasive Computing*, pages 299–317. IGI Global.

Szell, M., Sinatra, R., Petri, G., Thurner, S., and Latora, V. (2012). Understanding mobility in a social petri dish. *Scientific reports*, 2.

Tasse, D. and Hong, J. I. (2014). Using social media data to understand cities. In *Proceedings of NSF Workshop on Big Data and Urban Informatics*, pages 64–79. NSF Chicago, IL.

Thakur, G., Sims, K., Mao, H., Piburn, J., Sparks, K., Urban, M., Stewart, R., Weber, E., and Bhaduri, B. (2018). Utilizing geo-located sensors and social media for studying population dynamics and land classification. In *Human Dynamics Research in Smart and Connected Communities*, pages 13–40. Springer International Publishing.

Thériault, M. and Des Rosiers, F. (2013). *Modeling Urban Dynamics*. John Wiley & Sons.

Thompson, D. (1974). Spatial interaction data. *Annals of the Association of American Geographers*, 64(4):560–575.

Toole, J., de Montjoye, Y.-A., González, M., and Pentland, A. S. (2015). Modeling and understanding intrinsic characteristics of human mobility. In Gonçalves, B. and Perra, N., editors, *Social Phenomena, Computational Social Sciences*, pages 15–35. Springer.

Tsou, M.-H., Zhang, H., and Jung, C.-T. (2017). Identifying data noises, user biases, and system errors in geo-tagged twitter messages (tweets). *arXiv preprint arXiv:1712.02433*.

United Nations (2014). World urbanization prospects: The 2014 revision, highlights. Technical Report ST/ESA/SER.A/352, Department of Economic and Social Affairs, Population Division.

Vaca-Ruiz, C., Quercia, D., Aiello, L., and Fraternali, P. (2014). Tracking human migration from online attention. In Nin, J. and Villatoro, D., editors, *Citizen in Sensor Networks*, pages 73–83. Springer.

Van Audenhove, F.-J., Korniichuk, O., Dauby, L., and Pourbaix, J. (2014). The future of the urban mobility 2.0. Technical report, Arthur D. Little, http://www.adlittle.com/downloads.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer New York.

Vespignani, A. (2009). Predicting the behavior of techno-social systems. *Science*, 325(5939):425–428.

Vilhelmson, B. (1999). Daily mobility and the use of time for different activities. the case of sweden. *GeoJournal*, 48(3):177–185.

Wachowicz, M. and Liu, T. (2016). Finding spatial outliers in collective mobility patterns coupled with social ties. *International Journal of Geographical Information Science*, pages 1–26.

Wakamiya, S., Lee, R., and Sumiya, K. (2011). Crowd-based urban characterization: extracting crowd behavioral patterns in urban areas from twitter. In *Proceedings of the 3rd ACM SIGSPATIAL international workshop on location-based social networks*, pages 77–84. ACM.

Wang, D., Pedreschi, D., Song, C., Giannotti, F., and Barabasi, A.-L. (2011). Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1100–1108. ACM.

Wesolowski, A., Eagle, N., Tatem, A., Smith, D., Noor, A., Snow, R., and Buckee, C. O. (2012). Quantifying the impact of human mobility on malaria. *Science*, 338(6104):267–270.

Wu, F., Wang, H., Li, Z., Lee, W.-C., and Huang, Z. (2015). Semmobi: A semantic annotation system for mobility data. In *Proceedings of the 24th International Conference on World Wide Web Companion*, pages 255–258. International World Wide Web Conferences Steering Committee.

Yan, X.-Y., Han, X.-P., Wang, B.-H., and Zhou, T. (2013). Diversity of individual mobility patterns and emergence of aggregated scaling laws. *Scientific reports*, 3.

Yin, J., Gao, Y., Du, Z., and Wang, S. (2016). Exploring multi-scale spatiotemporal twitter user mobility patterns with a visual-analytics approach. *ISPRS International Journal of Geo-Information*, 5(10):187.

Yuan, J., Zheng, Y., and Xie, X. (2012). Discovering regions of different functions in a city using human mobility and pois. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 186–194. ACM.

Yuan, Y. and Raubal, M. (2012). Extracting dynamic urban mobility patterns from mobile phone data. In *Geographic information science*, pages 354–367. Springer.

Zhang, J., Wang, F.-Y., Wang, K., Lin, W.-H., Xu, X., and Chen, C. (2011). Data-driven intelligent transportation systems: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 12(4):1624–1639.

Zhao, Y.-L., Chen, Q., Yan, S., Zhang, D., and Chua, T.-S. (2014). Community understanding in location–based social networks. In Fu, Y., editor, *Human–Centered Social Media Analytics*. Springer.

Zheng, Y., Capra, L., Wolfson, O., and Yang, H. (2014). Urban computing: Concepts, methodologies, and applications. *ACM Transaction on Intelligent Systems and Technology*.