




Article

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

Fernando Santa ^{1,*} , Roberto Henriques ¹ , Joaquín Torres-Sospedra ²  and Edzer Pebesma ³ 

¹ Nova Information Management School (NOVA IMS), Universidade Nova de Lisboa, 1070-312 Lisbon, Portugal; roberto@novaims.unl.pt

² Institute of New Imaging Technologies, Universitat Jaume I, 12071 Castellón, Spain; jtorres@uji.es

³ Institute for Geoinformatics, University of Münster, 48149 Münster, Germany; edzer.pebesma@uni-muenster.de

* Correspondence: fernando.santa@novaims.unl.pt; Tel.: +351-213-828-610

Received: 14 December 2018; Accepted: 20 January 2019; Published: 23 January 2019



Abstract: 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 behavioral 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 analyzing 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, i.e., groups of hours of the day that present a similar spatial distribution of human activities.

Keywords: human activity; spatio-temporal statistics; negative binomial regression; functional principal component analysis; multitype spatial point patterns

1. Introduction

Geospatial research comprises substantial efforts in studying the dynamics of cities [1–4]. In general terms, humans exhibit regularity in their spatial, temporal, and social behaviors [5–8]; nevertheless, large-scale urban social systems are complex and challenging to model or represent [9,10]. The accelerated urbanization process of current society and the forecasts for a significant augment in urban populations are expected to enhance this complexity [11]. To predict these systems requires a mathematical description of the patterns found in city data, forming the basis of models that can be used to anticipate trends, assess risks, and manage future events [12]. The lack of data in this context has historically been a substantial problem [13]; 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 [3,12]. Practitioners including geographers, social researchers, and data scientists consider social media to provide core source data for studying cities [14]. To be sure, such information remains sparse in the geographical space, is incomplete in a time interval [15–17], and might not be

representative [18]; still, this information is considered a complementary alternative to the gathered information through survey sampling techniques for analyzing human activity since it captures people's perceptions and spatio-temporal changes more accurately [1,19,20]. Such data have driven the development of specific methods in network analysis, data mining, and statistics in a new branch of knowledge called urban informatics or urban analytics [21,22].

However, the use of social networks and their data have not been explored in-depth in urban studies [23,24]. Currently, ubiquitous computing has permitted collecting a large amount of data shared by people about themselves [25] and their interaction with the physical world [26]. Those datasets are far from conventional in the sense of tabular or structured data, and data processing has not analyzed a significant amount of them because of the computational expense and the need for specific data analysis techniques [27]. 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 [28].

To this end, we suggest a spatio-temporal statistical approach to analyze 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 characterize 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.

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.

On the other hand, although the analysis of content data has been a hot topic into the field of Twitter Analytics and urban informatics, it raises several constraints that can limit its scope and we want to avoid in our proposal. First, Twitter streaming API provides human-generated content that is not only text, which reduces the amount of data for analyzing in procedures of sentiment analysis, and thus, the representativeness of its results, for example, a significant amount of geotagged tweets come from third parties such as Instagram and Foursquare, among others. Second, the computational expense of text mining methods, the problems with the processing of textual information due to UTF-8 characters, study more than one language, and hashtag parsing [29]. Finally, and perhaps the most critical, privacy concerns associated with the identification of the users [4,16,19].

Our approach demonstrates that spatio-temporal statistical analysis provides valuable tools to analyze 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 [30–32].

The current paper is prepared as follows. In Section 2 we summarize the literature review of related studies. Section 3 describes how to collect, organize, and prepare datasets to implement our data processing methodology. In Section 4, we provide the statistical framework for regression models of count data, multitype spatial point patterns, and clustering over the scores of FPCA. Section 5 presents the results of the method for the three studied urban environments. In Section 6 we discuss and interpret our findings regarding previous studies. Section 7 summarizes our conclusions and the limitations of our proposal.

2. Background and Related Work

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 [33]. On the other hand, Goodchild [34] 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 [19,30,35–37], human activity [1,2,17,20,38,39], population dynamics [24,28,40,41], and event detection and disaster management [4,29,31,32,42,43], 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, which are looking for identifying groups of spatio-temporal similar locations. From data mining, it stands out: self-organizing maps (SOM) [19,35], hierarchical SOM [44], independent component analysis (ICA) [17], density-based spatial clustering of Applications with Noise (DBSCAN) and ST-DBSCAN [29,43], and random forests [40]. Meanwhile, from the branch of statistical analysis, it emerges ordinary least-squares (OLS) [30], generalized additive models (GAM) [32], local indicators of spatial association (LISA) [24], space-time scan statistics (STSS) [42], Gaussian mixture models (GMM) [45], and kernel density estimation (KDE) [1,38]. 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. [30] normalized 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, while the statistical analysis of spatial and spatio-temporal point patterns estimates and uses autocorrelation structures to study the distribution of the events. Also, ST-DBSCAN depends on the selection of distance and tolerance parameters [29]. 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) [36]; then, for avoiding this issue, a better alternative is analyzing the locations and timestamps as points rather than spatial aggregations.

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 analyze significant volumes of human-generated data in cities, (2) a way to gain insight into human behavior 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 [46]. 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 [47]. 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 [48,49]. Furthermore, FPCA brings the possibility of reducing dimensionality, highlighting the relevant underlying characteristics in spatial summary measures [50].

3. Data

3.1. Collection

Our approach depends on collecting human-made social media activity over a period in an urban environment. Figure 1 summarizes 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** [51], the language and environment for statistical computing, and its package `tweet2r` [52] 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 [44,53–55], the gathered amount of data depends on the volume of usage of the social network in the city.

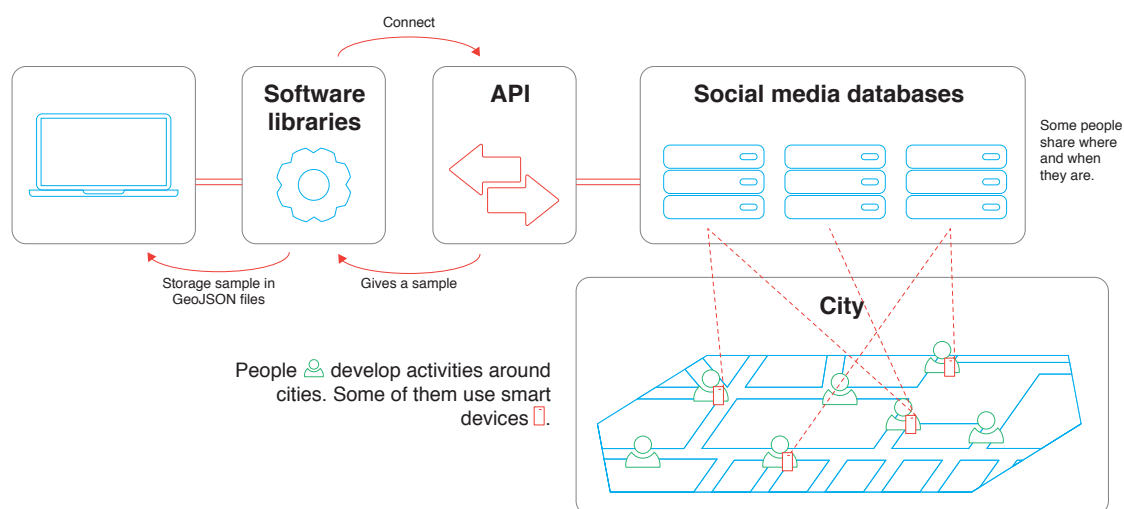


Figure 1. Schema of gathering geolocated social media data.

3.2. Preprocessing

Human-Generated Tweets

After data collection, it is necessary to conduct data preprocessing to identify the information generated exclusively by humans and to prepare the databases for the subsequent analyses. Figure 2 presents a schema with the main steps to implement. Initially, GeoJSON files are merged and converted into a table that organizes by rows each gathered tweet and by columns its metadata. Due to the significant number of recorded tweets, the gathered data contain noise that does not necessarily represent people's activities, and the information requires filtering to access only activity generated by humans before performing any analysis and to avoid bias in the results [56]. The cleaning and removal of the noises is a semiautomatic iterative task that evaluates several sources of perturbation. Tsou et al. [57] discussed system errors, commercial bot and automated tweets, and user tweeting frequency. Furthermore, Frias-Martinez et al. [19,54] described 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 in addition to 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 content associated with advertising, Tsou et al. [57] suggested reviewing the field called 'source' in the metadata of the tweets, which allows for the identification of 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 originating from an automated process by manual inspection. Finally, we analyze the user and location tweeting frequency by enumerating unique users and coordinates and then counting the number of tweets in each case, which permits identifying and eliminating those tweets related to users and places that appear with a high and unusual frequency.

3.3. Datasets Construction for Statistical Analysis

For performing the statistical analysis, our approach builds a new dataset that keeps only three fields, namely, 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 $0, 1, \dots, 23$.
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 h 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, $i = 1, 2, \dots, N$ 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 h; the elapsed time for that tweet is 159.31 h, and $t_{N_i} = 159$.

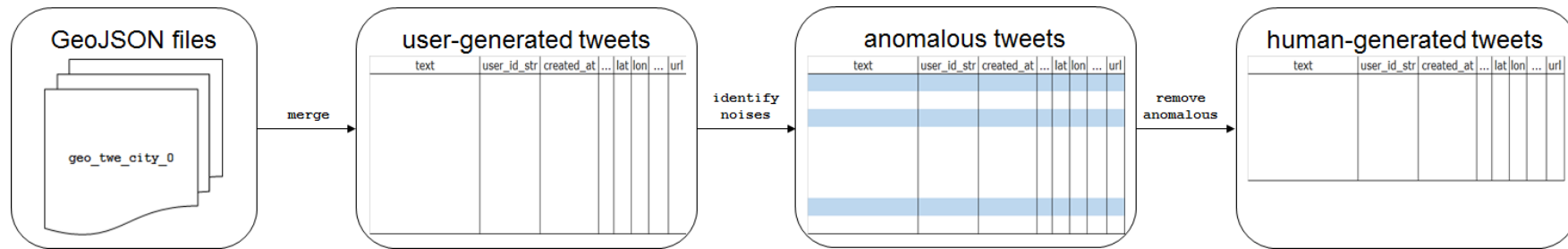


Figure 2. Process for preprocessing samples of geolocated tweets.

3.3.1. Temporal Dataset

Our temporal data analysis approach requires the creation of a table based on the field called t_N . Let n_h be the count of the number of obtained tweets at hour h , $h = 0, 1, \dots, T$. This procedure gives a discrete-time time series of counts $\{n_0, n_1, \dots, n_T\}$. 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 identifying autoregressive and seasonal autoregressive schemas, the five last hours ($n_{-1}, n_{-2}, \dots, n_{-5}$) and the same hours as the day before ($n_{-24}, n_{-25}, \dots, n_{-29}$). Following our previous example, where $t_0 = '2017-07-30 00:00:00'$ and $t_{T+1} = '2017-08-13 00:00:00'$, Table 1 shows an schema of a possible temporal dataset.

Table 1. Schema of a temporal dataset.

Date	t_N	n	Autoregressive			Seasonal Autoregressive			Day-of-The-Week			Hour-Of-The-Day		
			n_{-1}	...	n_{-5}	n_{-24}	...	n_{-29}	Tuesday	...	Sunday	00:00	...	23:00
2017-07-30 00:00	0	n_0	n_{-1}	...	n_{-5}	n_{-24}	...	n_{-29}	0	...	1	0	...	0
2017-07-30 01:00	1	n_1	n_0	...	n_{-4}	n_{-23}	...	n_{-28}	0	...	1	1	...	0
2017-07-30 02:00	2	n_2	n_1	...	n_{-3}	n_{-22}	...	n_{-27}	0	...	1	0	...	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2017-mm-dd hh:00	h	n_h	n_{h-1}	...	n_{h-5}	n_{h-24}	...	n_{h-29}	0	...	0	0	...	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2017-08-12 23:00	335	n_{335}	n_{334}	...	n_{330}	n_{311}	...	n_{306}	0	...	0	0	...	1

3.3.2. 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* [58]. Table 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 2. Schema of a spatio-temporal dataset.

East	North	Hour
x_{1_0}	y_{1_0}	0
x_{2_0}	y_{2_0}	0
⋮	⋮	⋮
x_{n_0}	y_{n_0}	0
x_{1_1}	y_{1_1}	1
x_{2_1}	y_{2_1}	1
⋮	⋮	⋮
x_{n_1}	y_{n_1}	1
⋮	⋮	⋮
$x_{1_{23}}$	$y_{1_{23}}$	23
$x_{2_{23}}$	$y_{2_{23}}$	23
⋮	⋮	⋮
$x_{n_{23}}$	$y_{n_{23}}$	23

3.4. 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 [24]. Yet, it is not possible to argue that LBSN data are a representative of actual activity in the cities [2] 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. Statistical Framework and Methods

Our data analysis approach focuses on three main aspects. First, we implement several statistical methods to analyze 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 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 behaviors. 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.

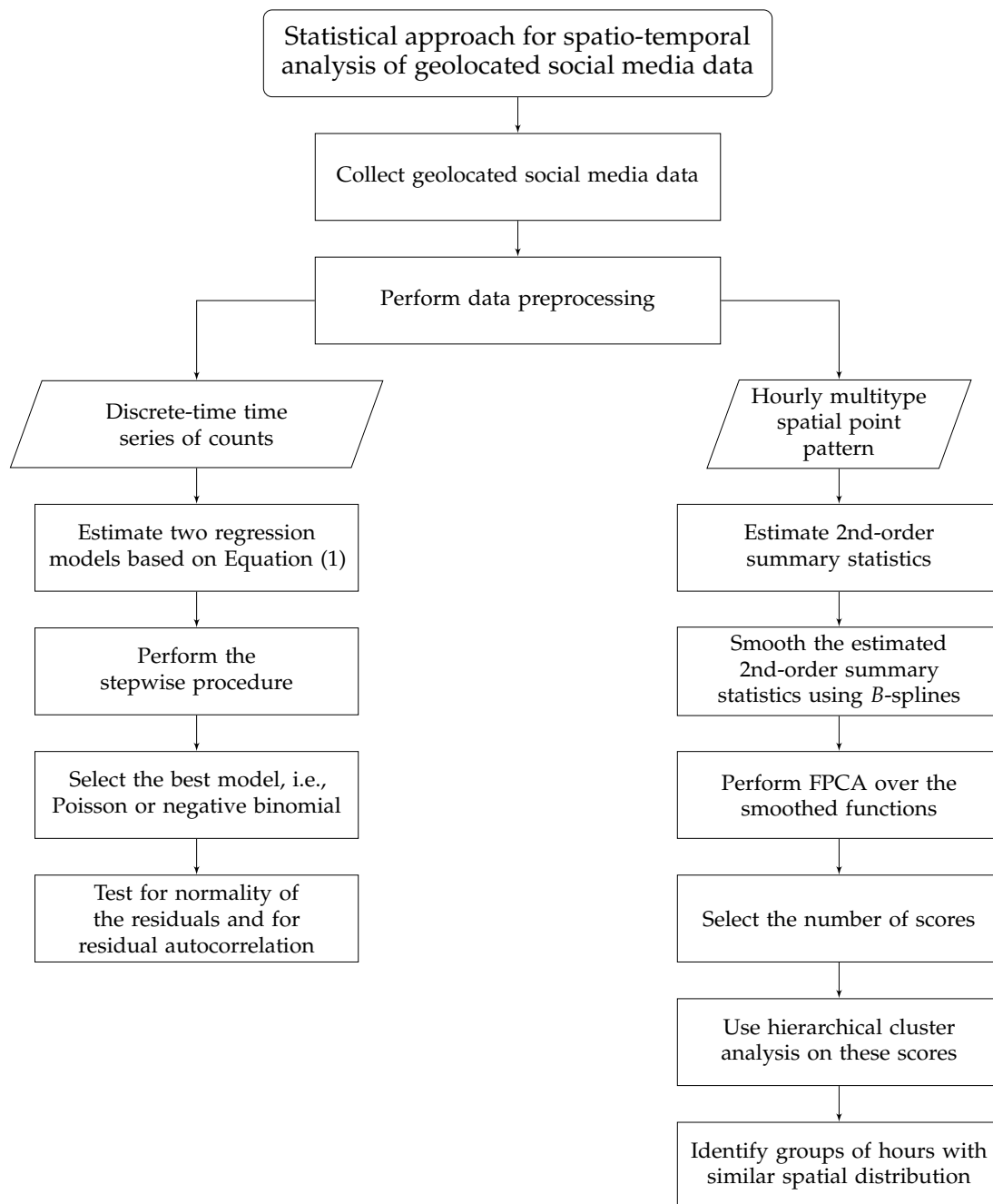


Figure 3. Methodological approach.

4.1. Regression Models for Count Data

We used regression techniques to explain the variations in the mean μ of a variable (called *response variable*) associated with a set of factors (called *explanatory variables* x_1, \dots, x_p) and to quantify the magnitude of their effect through a collection of values called *parameters of the model* $\beta_0, \beta_1, \dots, \beta_p$ [59]. Classical regression models rely on the assumption that the response variable follows a Gaussian distribution [60]. However, in our case, the response variable $n_t, t = 0, \dots, T$ represents the count of the number of geolocated tweets collected per hour, i.e., a non-negative, discrete variable. In this case, statistical modelling falls under the scope of generalized 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 [61]. To formulate a GLM requires the specification of three elements [62]. 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, $\eta = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$. Finally, a monotone, differentiable *link function* g that relates the mean of the response variable with the systematic part, $g(\mu) = \eta = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$.

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 (overdispersion) [63]. 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 [64]. The estimation of the parameters of the model is made by maximum likelihood through the iteratively weighted least-squares (IWLS) algorithm [65].

Let $n_t, t = 0, \dots, T$ 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:

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

where $\beta_j, j = 0, \dots, 40$ 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 (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 [66]. We carry out a stepwise process to select explanatory variables based on the Bayesian information criterion (BIC) [67]. The obtained models are compared to choose the best model regarding the probability distribution of the response variable by using a likelihood ratio contrast [68]. 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.

4.2. Spatio-Temporal Analysis

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 [69]. Thus, a set $\mathbf{S} = \{\mathbf{s}_1, \dots, \mathbf{s}_k\} : \mathbf{s}_i = (x_i \ y_i)^T, i = 1, \dots, k$, where $\mathbf{s}_i \in W \subset \mathbb{R}^2$ denotes the location of the i -th event, is called a *spatial point pattern* [70]. From the statistical point of view, two types of measures summarize the pattern: (1) *first-order statistics* characterize 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 $\lambda(\mathbf{s}_i, \mathbf{s}_j)$ [71]. 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 [48].

Second-order summary statistics are functions that express the degree of spatial relationship between the events of the pattern for several spatial scales [46]. 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 $g(r) = 1$ [49].

In many applications, the objective of the study is analyze 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 characterization 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 $\zeta_j, j = 1, \dots, l$ 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 [48].

The analysis implies the estimation of summary statistics for the formed pattern by each type of mark in several distances, e.g., $r_q, q = 1, \dots, m$. This estimation produces numerical realizations 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 [49]. Then, techniques such as the FDA provides tools for understanding the spatial behavior of the pattern since it considers that observations are functions or single units rather than consecutive measurements [72].

A dataset in the FDA is a sample of the following form [73]:

$$u_j(r_{q,j}) \in \mathbb{R}, \quad r_{q,j} \in [r_1, r_m], \quad j = 1, \dots, l, \quad q = 1, \dots, m \tag{2}$$

where we have l observed curves over the same interval $[r_1, r_m]$. The basic idea is that the objects of study are the smooth curves

$$\{u_j(r) : r \in [r_1, r_m], \quad j = 1, \dots, l\} \tag{3}$$

defined for all values of r but observed only at selected points $r_{q,j}$.

Then, the first step in the analysis involves converting the sample in Equation (2) to functions by using smoothing techniques, which involves determining a set of *functional blocks* or *basis functions* $\phi_f, f = 1, \dots, F$ and a set of coefficients $c_f, f = 1, \dots, F$ to define each function as a linear combination of these basis functions; thus, $u_j(r) = \sum_{f=1}^F c_{jf} \phi_f(r), j = 1, \dots, l$. 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 [74].

Conventionally, by using least-squares or localized least-squares fits, it is possible to estimate the coefficients $c_f, f = 1, \dots, F$. 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 *penalized smoothing* approach is preferred to minimize the effect of the random variability. This approach uses many basis functions and penalizes 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 [73].

Once a functional representation of the data is obtained, the FPCA is a valuable tool to explore and identify features in the curves and the number of types of them. As with the principal component analysis (PCA) used in classical multivariate methods, FPCA defines a new set of scalar variables $f_{j,p}, j = 1, \dots, l, p = 1, \dots, P$, called *scores*, as linear combinations of the smooth functions. Thus, [75],

$$f_{j,p} = \int \xi_p(r) u_j(r) dr \tag{4}$$

where $\xi_p(r)$ is a weight function that maximizes $l^{-1} \sum_{j=1}^l f_{j,p}^2$ subject to the constraint $\|\xi_p\|^2 = \int \xi_p(r)^2 dr = 1$. This process defines an eigenequation:

$$\int v(o, r) \xi_p(r) = \delta \xi_p(r) \tag{5}$$

with associated covariance function $v(o, r) = l^{-1} \sum_{j=1}^l u_j(o) u_j(r)$. The solution of Equation (5) gives the eigenvalues (δ) 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 $d \ll L$.

The format of the data (see Table 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 the day of a tweet shared in city W . We assume that these data constitute a full register of all events that happen within W at hour h . We 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:

$$\hat{K}_h(r) = \frac{|W|}{n_h(n_h - 1)} \sum_{i=1}^{n_h} \sum_{\substack{j=1 \\ j \neq i}}^{n_h} \mathbf{1}\{\|\mathbf{s}_{i_h} - \mathbf{s}_{j_h}\| \leq r\} e_{i_h i_h}(r) \quad (6)$$

where $|W|$ is the area of the city, n_h is the number of tweets at hour h , $\|\mathbf{s}_{i_h} - \mathbf{s}_{j_h}\|$ is the distance between the geolocation of two collected tweets at hour h , $\mathbf{1}\{\cdot\}$ takes the value 1 when the distance is less than or equal to r or 0 otherwise, and $e_{i_h i_h}(r)$ is an *edge correction weight* defined by the geometry of the window and the number of the events in the point pattern [46].

Based on the estimator (6), we calculate Besag's L -function $\hat{L}_h(r), h = 0, \dots, 23$ for distances $r_q, q = 1, \dots, m$, 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 smaller side length of the rectangle that circumscribes the window W [48]. From those estimations, we 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 [73]. We establish the number of the functional blocks by using the rule $F = m + 2$ [74]. 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 [76], which allows for the detection of groups of similar second-order summary statistics, i.e., groups of hours with a similar spatial distribution of tweet locations.

5. Case Study

To evaluate our data analysis approach, we collected geolocated tweets in a two-week period from 28 July 2017, at 12:22:00 UTC/GMT+1 h to 14 August 2017, 12:21:59 UTC/GMT+1 for the metropolitan areas of Lisbon, London, and New York City. Table 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 4373, 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 h, between 0 and 335. We then processed 3626, 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.

Table 3. Parameters of the query in the Twitter 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
	No geolocated	198,418	928,197	1,094,420
	Clean	11,817	87,448	119,802

Figures 4 and 5 show the bar charts of the temporal distribution of the collected tweets during the study period. We aggregated the tweets for the two-week period by hours of the day and days of the week. We found considerable differences between the amount of gathered information in each city, but their distribution throughout the day presented similar patterns of behavior. We discovered a sharp decrease in the usage of Twitter after midnight until early in the morning, followed by an increase that peaked in the evening. Those maximums did not occur at the same hour, lying between 19:00 and 21:00 in Lisbon, between 17:00 to 19:00 in London, and at 18:00 in Manhattan. Regarding the day of the week, the three cities showed marked variations: Lisbon had more social media activity from Tuesday to Thursday, London recorded more tweets on the weekends than on the weekdays, and there was no a considerable difference in the volume of human-generated data in Manhattan. Figure 6 displays the three-count time series for the 336 h of analysis, revealing a daily seasonal effect in the frequency of interaction with social networks. Lisbon’s time series exhibits unusual activity: a high number of tweets in the evenings on 1 August, 3 August, and 9 August 2017.

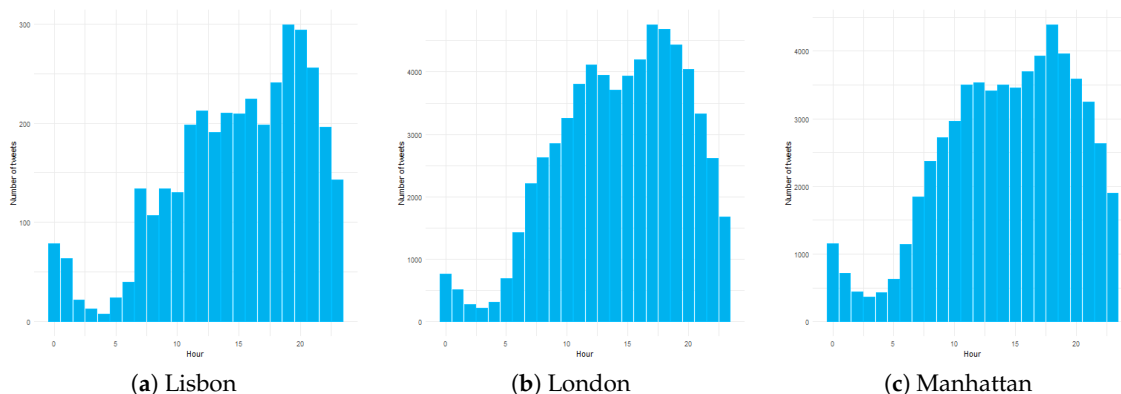


Figure 4. Hourly distribution of geolocated tweets.

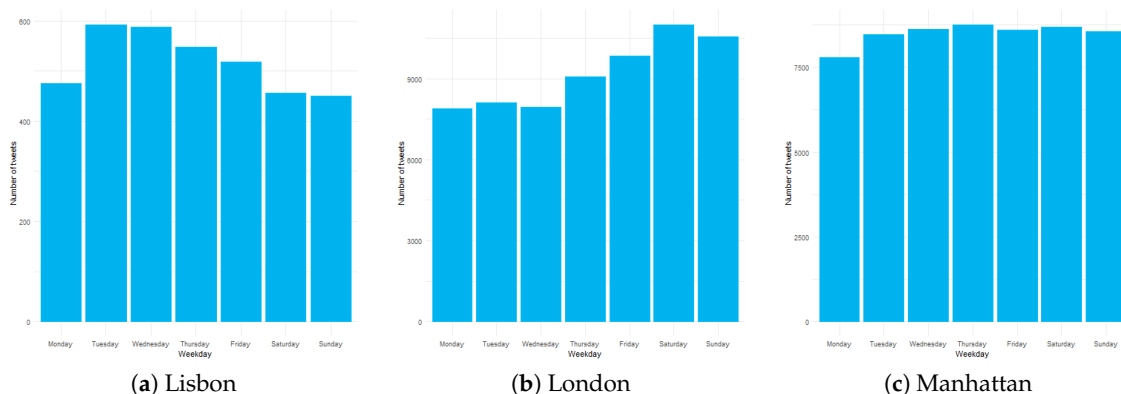


Figure 5. Weekly distribution of geolocated tweets.

For each city, we adjusted two count regression models based on Equation (1) by considering Poisson and negative binomial responses. Table 4 presents the statistics for the goodness of fit to the selected best model in each case. 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.

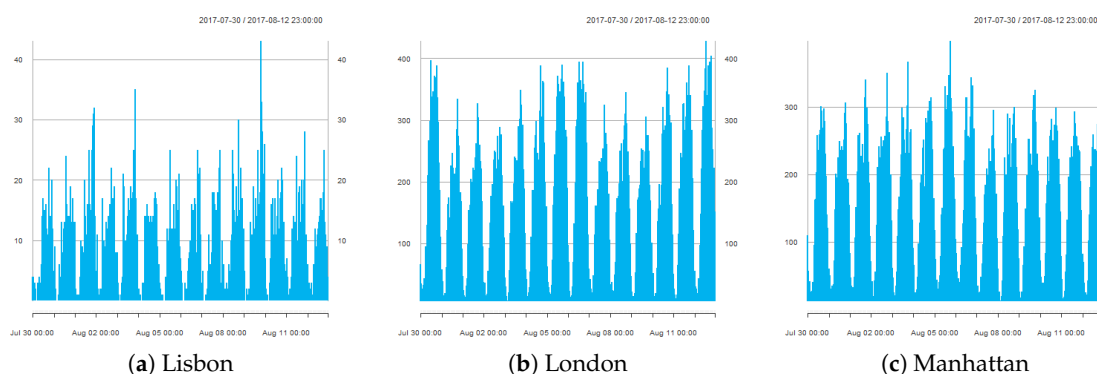


Figure 6. Time series of hourly geolocated tweets in three urban environments.

Table 4. Statistics of goodness of fit for estimated count regression models.

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

code: ***, **, *, ., ;; *p*-value: [0,0.001], (0.001,0.01], (0.01,0.05], (0.05,0.1), (0.1,1].

Table 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 from Tuesday to Thursday in Lisbon, from Thursday to Sunday in London, and with no influence in 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, with 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 h before is negatively correlated with the activity for the current hour. In the case of London, an increase in the social media activity 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.

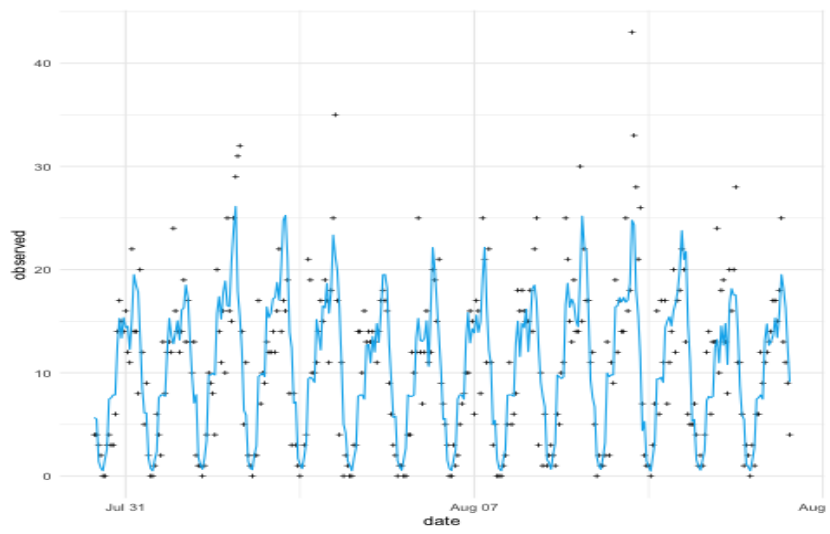
Table 5. Estimated regression coefficients and 95% confidence intervals in the fitted negative binomial regression models for the number of geolocated tweets per hour.

(a) Lisbon			(b) London			(c) Manhattan		
Parameter	Estimate	95% CI	Parameter	Estimate	95% CI	Parameter	Estimate	95% CI
Intercept	2.062	(1.952, 2.172)	Intercept	3.803	(3.721, 3.884)	Intercept	3.965	(3.823, 4.108)
Tuesday	0.237	(0.112, 0.362)	Thursday	0.07	(0.027, 0.112)	01:00	-0.321	(-0.455, -0.187)
Wednesday	0.239	(0.113, 0.364)	Friday	0.125	(0.08, 0.17)	02:00	-0.698	(-0.854, -0.542)
Thursday	0.197	(0.069, 0.325)	Saturday	0.174	(0.122, 0.226)	03:00	-0.813	(-0.984, -0.644)
02:00	-1.368	(-1.85, -0.934)	Sunday	0.156	(0.106, 0.206)	04:00	-0.618	(-0.789, -0.447)
03:00	-1.986	(-2.604, -1.458)	01:00	-0.291	(-0.414, -0.169)	05:00	-0.252	(-0.416, -0.089)
04:00	-2.536	(-3.333, -1.893)	02:00	-0.859	(-1.004, -0.717)	06:00	0.318	(0.166, 0.47)
05:00	-1.523	(-1.977, -1.115)	03:00	-1.055	(-1.214, -0.9)	07:00	0.697	(0.555, 0.839)
06:00	-1.033	(-1.393, -0.698)	04:00	-0.741	(-0.882, -0.603)	08:00	0.824	(0.693, 0.956)
11:00	0.531	(0.321, 0.741)	06:00	0.685	(0.578, 0.791)	09:00	0.837	(0.714, 0.959)
12:00	0.736	(0.53, 0.942)	07:00	1.019	(0.909, 1.129)	10:00	0.848	(0.728, 0.968)
13:00	0.585	(0.374, 0.795)	08:00	1.077	(0.958, 1.196)	11:00	0.955	(0.835, 1.076)
14:00	0.723	(0.515, 0.929)	09:00	1.078	(0.954, 1.203)	12:00	0.869	(0.739, 0.999)
15:00	0.712	(0.505, 0.919)	10:00	1.163	(1.037, 1.289)	13:00	0.791	(0.661, 0.922)
16:00	0.865	(0.653, 1.076)	11:00	1.289	(1.165, 1.412)	14:00	0.841	(0.714, 0.968)
17:00	0.751	(0.533, 0.97)	12:00	1.33	(1.204, 1.456)	15:00	0.82	(0.691, 0.95)
18:00	0.932	(0.725, 1.14)	13:00	1.251	(1.12, 1.382)	16:00	0.884	(0.756, 1.013)
19:00	1.161	(0.959, 1.365)	14:00	1.201	(1.073, 1.33)	17:00	0.919	(0.787, 1.052)
20:00	1.144	(0.941, 1.348)	15:00	1.292	(1.169, 1.414)	18:00	0.976	(0.837, 1.114)
21:00	1.036	(0.825, 1.249)	16:00	1.37	(1.249, 1.491)	19:00	0.795	(0.645, 0.945)
22:00	0.731	(0.513, 0.948)	17:00	1.496	(1.371, 1.621)	20:00	0.731	(0.586, 0.877)
23:00	0.471	(0.229, 0.711)	18:00	1.401	(1.263, 1.54)	21:00	0.711	(0.577, 0.846)
n_{-5}	-0.018	(-0.026, -0.01)	19:00	1.327	(1.188, 1.466)	22:00	0.572	(0.445, 0.699)
			20:00	1.248	(1.111, 1.384)	23:00	0.354	(0.233, 0.474)
			21:00	1.119	(0.991, 1.247)	n_{-1}	0.002	(0.001, 0.003)
			22:00	0.998	(0.883, 1.114)	n_{-2}	0.001	(0.000, 0.002)
			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)			

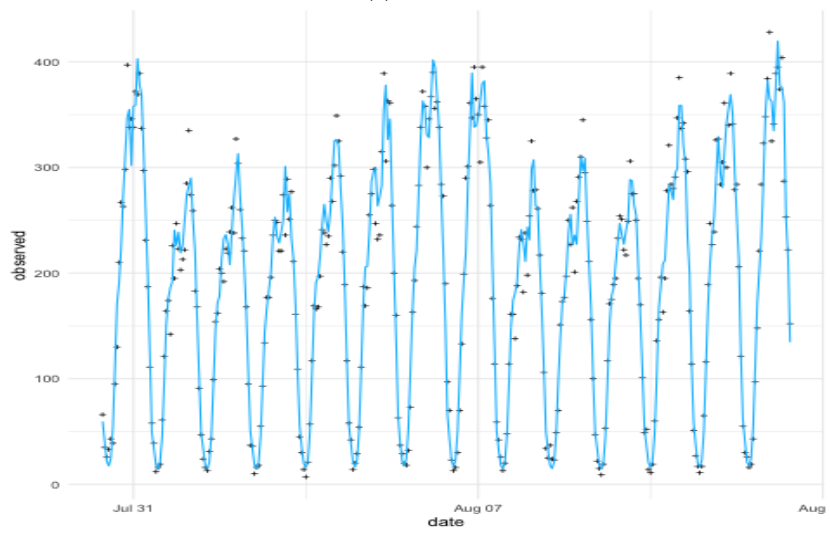
Figure 7 compares observed and fitted numbers of geolocated tweets over the observation period in the three cities. The predicted values obtained through the models adequately represent most of the trends that the data exhibit. We also evaluate the forecasts accuracy by summarizing the forecast errors of the estimated models for the two following days of the information, 13 and 14 August 2017, whose counts were not included in the estimation of the parameters of the models. Table 6 presents the statistics of the fit: Pearson coefficient correlation, root mean squared error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and symmetric mean absolute percentage error (sMAPE). The accuracy of forecasts with the estimated models shows that the fitted models for London and Manhattan are better than for Lisbon, despite the RMSE and MAE are lower in Lisbon, but those measures are absolute and cannot be compared between cities.

Table 6. Forecast accuracy evaluation of the fitted negative binomial regression models for the number of geolocated tweets per hour.

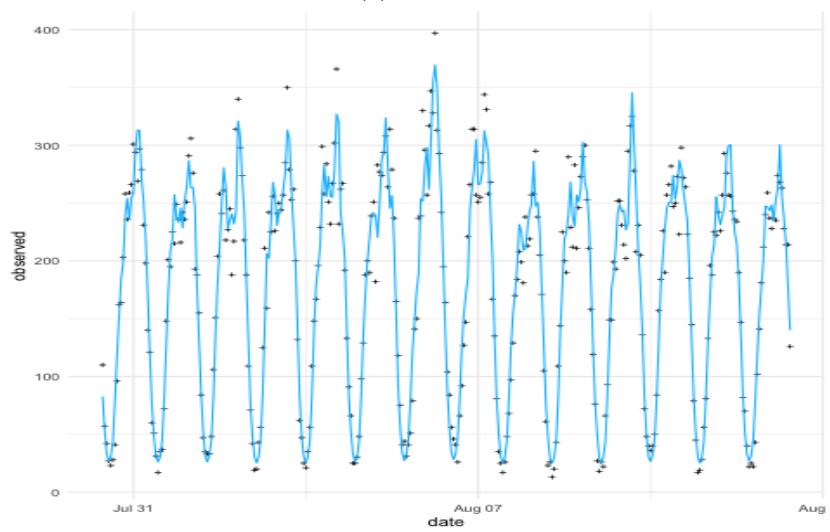
City	Pearson’s Correlation	RMSE	MAE	MAPE	sMAPE
Lisbon	0.83	3.95	3.07	79.06	65.30
London	0.97	32.71	22.23	19.89	18.89
Manhattan	0.98	20.17	15.54	19.15	16.76



(a) Lisbon



(b) London



(c) Manhattan

Figure 7. Observed temporal variation of geolocated tweets (black dots) together with the fitted variation from a negative binomial regression model (deep-sky-blue lines).

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 are as follows: Lisbon: $W = 0.996$, p -value = 0.65; London: $W = 0.995$, p -value = 0.34; and Manhattan: $W = 0.996$, p -value = 0.27; these results indicate 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 A1 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 and 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 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, 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 function. We worked with a sequence of values from 0 meters up to r_m with intervals of 25 meters. Table 7 shows a summary of the obtained ranges. We then estimated the centered version $\hat{L}(r) - r$ of the function, for every hourly formed spatial point pattern, over the set of those distances.

Table 7. Distance parameters for estimating the second-order summary statistics for the hourly multitype spatial point patterns of tweets in the three studied cities

City	Length of the Shorter Side	1/4 of the Length	r_m	m
Lisbon	11,530.11	2882.53	2875	115
London	44,819.03	11,204.76	11,200	449
Manhattan	30,153.90	7533.96	7525	302

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 of $\lambda = 0.00001$. We then calculated the FPCA over the smoothed curves and kept the first two principal scores since together they contributed 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 [74]. Finally, we obtained a dendrogram by applying agglomerative hierarchical clustering through Ward's method on the matrix of dissimilarities computed with the Euclidian distance between the scores.

Figures 8–10 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 the considered distances and hours of the day, except for the case of the events registered at 04:00, whose curve decreases rapidly and reaches negative values after 1.75 km. In addition, 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. As a functional principal component symbolizes variation over the average curve, the interpretation depends on this capability. Thus, since the first component takes negative values for distances up to 500 meters, 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 8 reveals that the spatial distribution, of the shared events at 04:00, is quite dissimilar in comparison with the behavior 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.

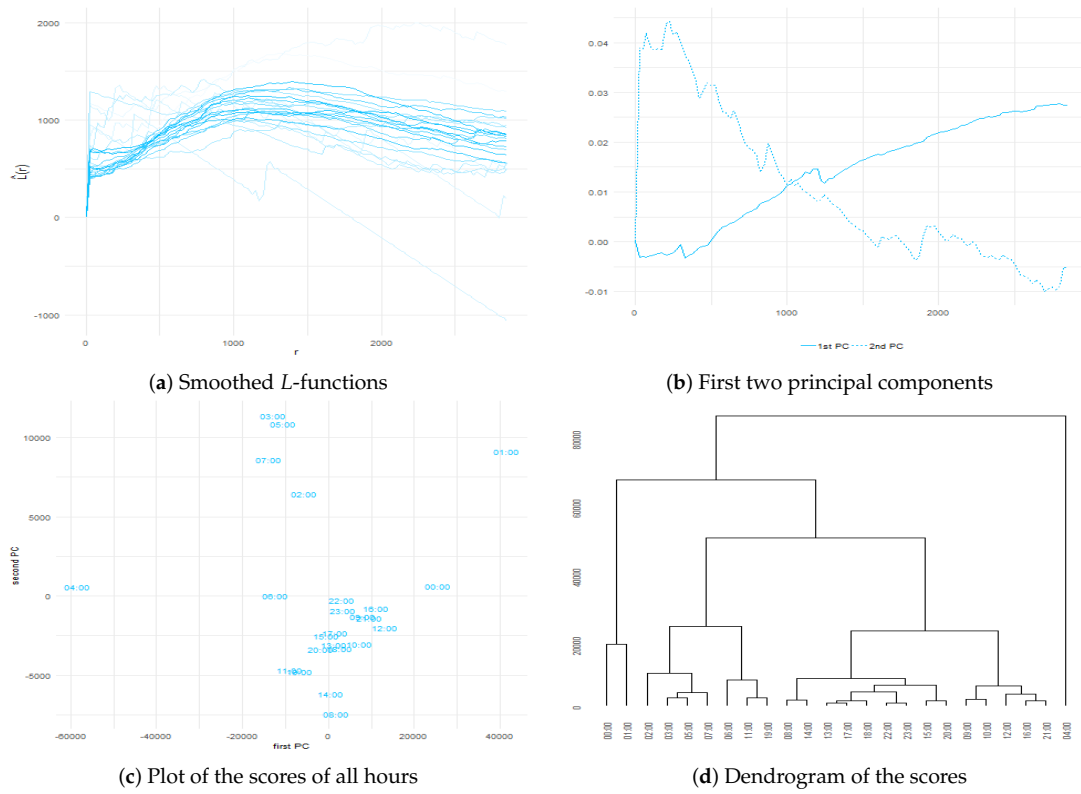


Figure 8. Results from an FPCA on L -functions in Lisbon.

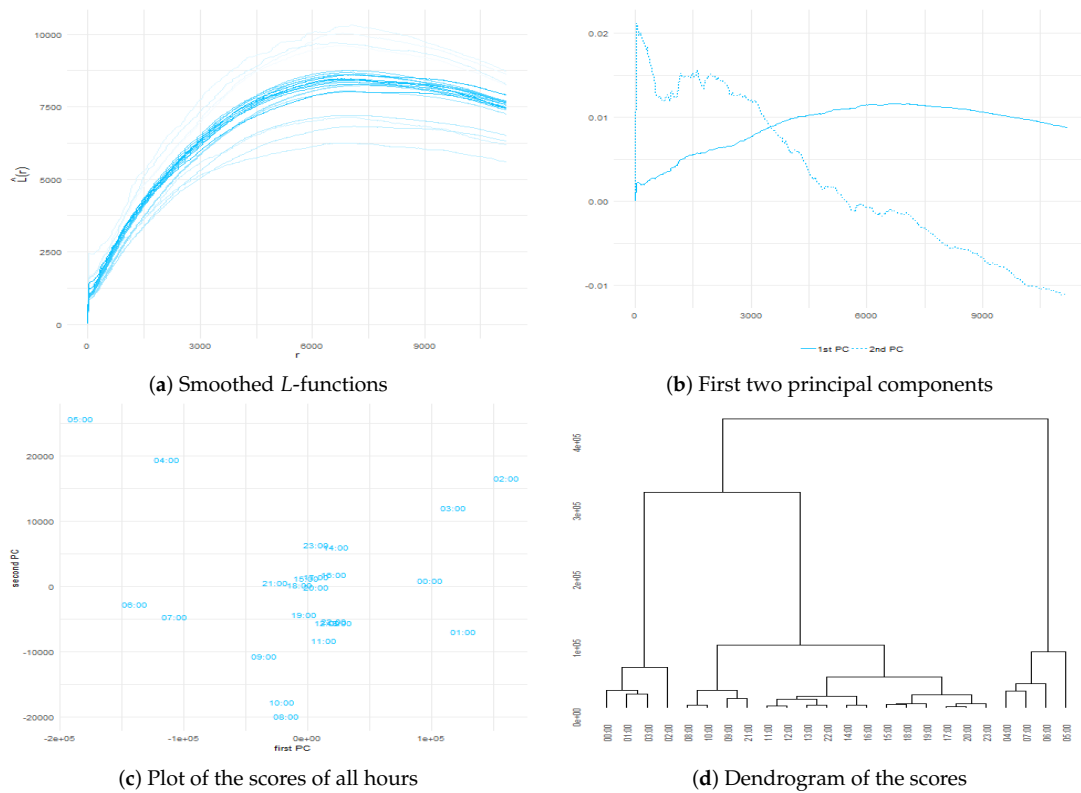


Figure 9. Results from an FPCA on L -functions in London.

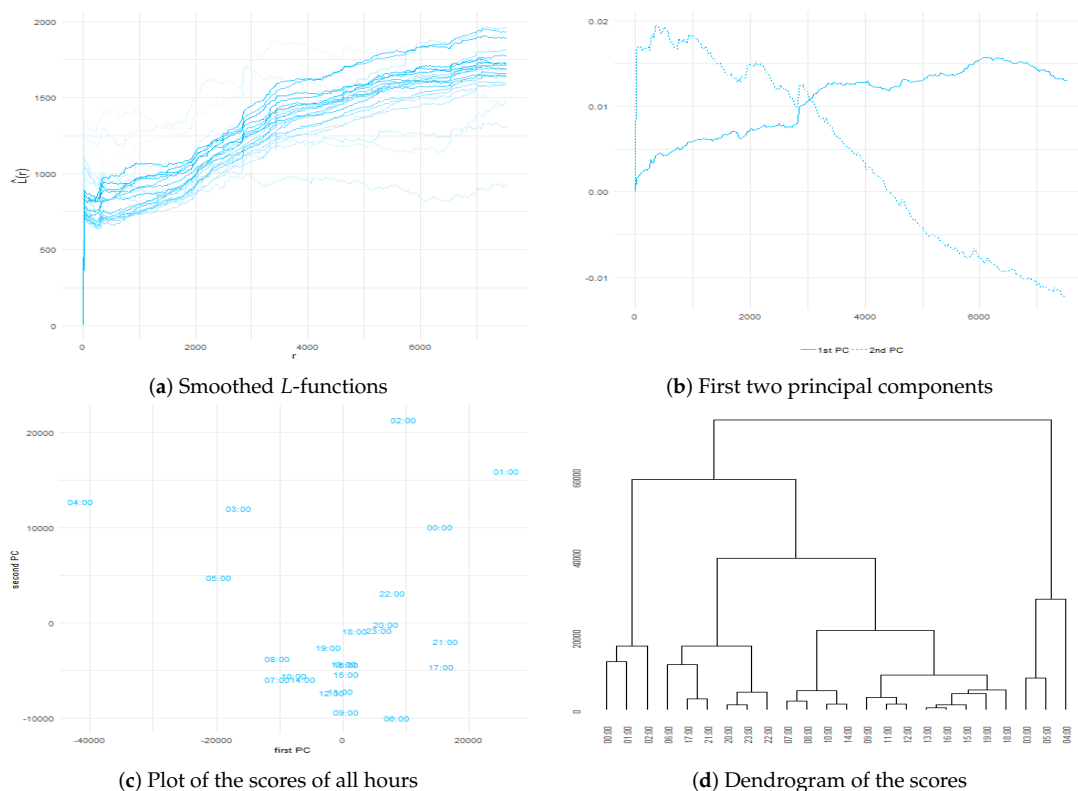


Figure 10. Results from an FPCA on *L*-functions in Manhattan.

The smoothed centered *L*-Besag functions for London show less irregularity than those 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. 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 the scatterplot of the first two functional principal components reveals that there are approximately four groups, three of them for early hours in the morning and the other for the rest of the day.

Similar to the other two PC 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 for all distances. In addition, the smoothed curves that belong to hours after midnight until 05:00 exhibit more irregularity than those at other hours of the day. The first two harmonics explain 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 higher than 2.8 km, whereas the second component portrays increments of the variations up to 4.2 km. The hourly data can be divided into four groups: one for activities performed from 00:00 to 02:00, another for 03:00 to 05:00, and the other two for later hours.

Finally, we obtained the intensity function of each mark of the multitype spatial point pattern by using bidimensional-KDE. We employed the quartic kernel and selected the bandwidth by using Scott’s method [77]. We then standardized all the estimated values and normalized them to a scale of 0–1 by subtracting their minimum and then dividing by their range. Figures 11–13 display the estimations. In Lisbon, there is a persistent accumulation of the human-generated data in the southern area lying along the Tagus River; this location contains the main attractions within the city. On the other hand, London’s metropolitan area concentrates social media activity in the surrounding boroughs of

the city of London, which contain tourist locations, large companies, and commercial areas. The island of Manhattan aggregates most of the Twitter activity in the direction southwest of Central Park to the limits with the Upper Bay and the Hudson River, in which Times Square, SoHo, and the Financial District, among other attractions, are located.

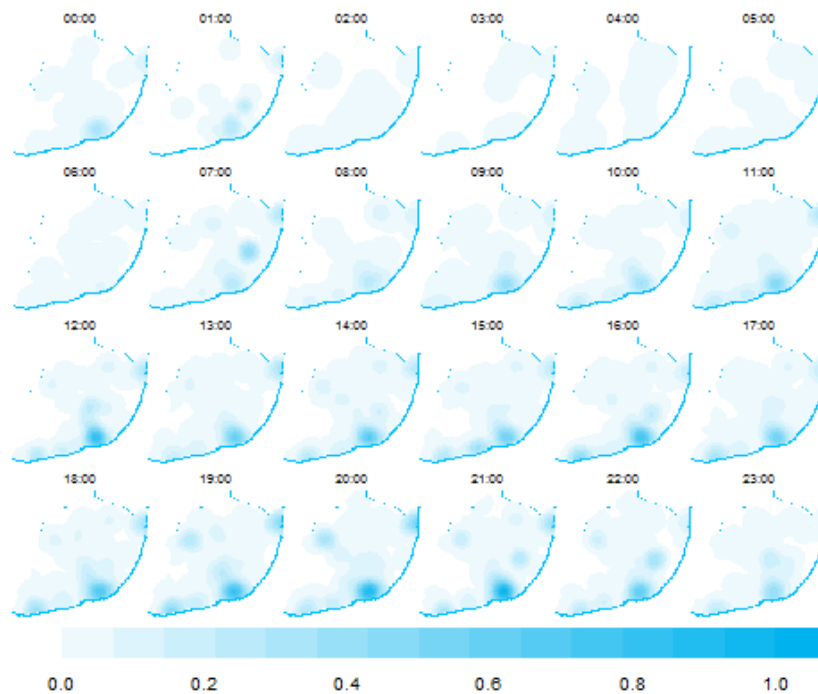


Figure 11. Estimated intensity function of the hourly multitype spatial point pattern in Lisbon.

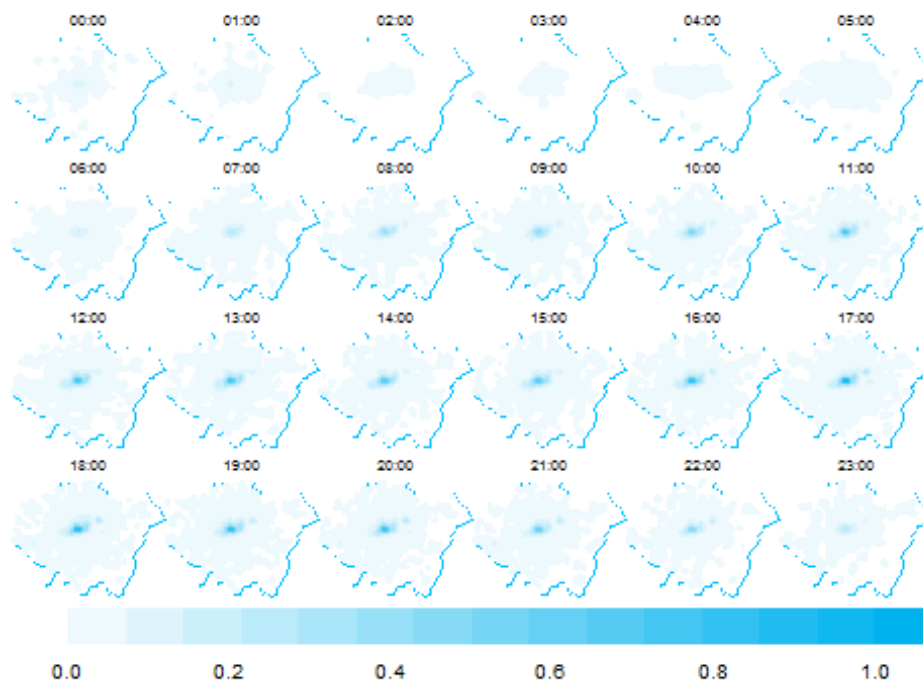


Figure 12. Estimated intensity function of the hourly multitype spatial point pattern in London.

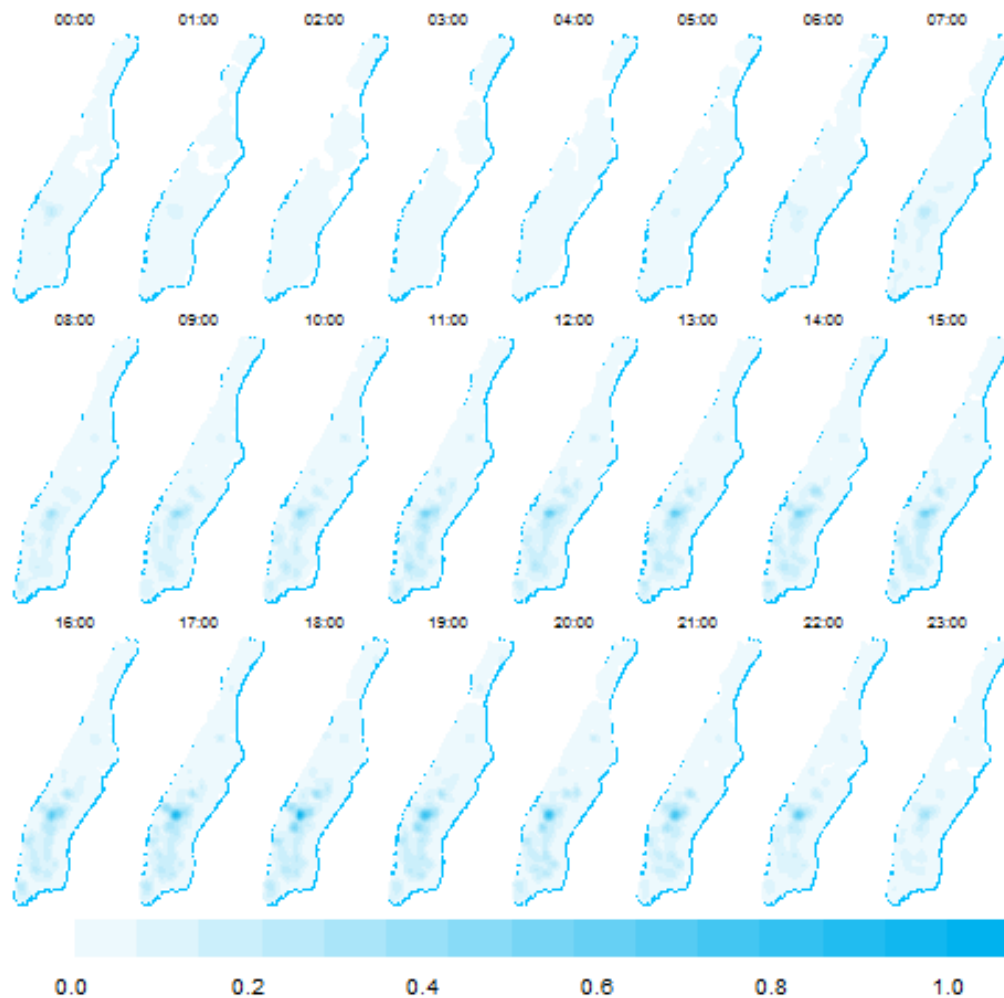


Figure 13. Estimated intensity function of the hourly multitype spatial point pattern in Manhattan.

6. 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 analyzed systems and agrees with as stated by [9]. 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. [1,35,44] studied social media data from London and Manhattan and identified areas with high social media activity and differences in behavior 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 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 9 August 2017, at 19:00 to the prematch tweets of the Portuguese local soccer league between Benfica and 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 that people tend to visit the same places at the same times [6,7,16]. 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 of spatial correlation through the time tend to be more homogeneous in short distances, at 500 m, 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 [5–7]. The behavior 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 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; as a consequence, this covered other relevant spatial effects that occur in different periods of the day.

We found some limitations in our research. One of the main issues was regarding the analysis of content data that 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. However, that was not our primary goal. We aimed 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. On the other hand, the identification and elimination of shared content by machines, bots, and other sources that are not people is a complex task that restricts the analysis. Also, 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 is just 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.

Our work has characterized the behavior of human-generated Twitter data in cities as a proxy for human activity. We have found temporal and spatial regularities, namely, autoregressive and long-term temporal trends, as well as, spatial clustering in various scales that can be used as a first step for developing more complex and reliable models to explain the underlying nature of urban dynamics, such as mechanistic models [46]. This approach would face the problem to estimate parameters of spatio-temporal models in large datasets in which the Bayesian techniques have taken in a boom since they allow developing complex models without imposing simplified structures [78]. Notably, the integrated nested Laplace approximations (INLA) has shown to be a fast and accurate algorithm in comparison with the conventional Markov chain Monte Carlo methods [79–81] that can analyze social media data.

On the other hand, 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 realizations of spatio-temporal processes with autoregressive behavior which do not come from planned experiments. Its observations, number of events, are not independent, and phenomena are only partially observed [82]. 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 behavior of the people in the cities at other times of the day. Therefore, we also suggest 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.

7. Conclusions

This research reviewed relevant literature in the field of human urban activity, social media analytics and summarized several components in its study, as well as reflections on the statistical methods to find urban patterns. It has been highlighted the potential opportunities and impact of the analytical approaches suggested for the analysis of Twitter data and the study of the spatio-temporal distributions related to the understanding of the human urban activity. These have mainly 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 dynamics patterns on a large-scale level.

We have proposed an alternative to studying the spatio-temporal distribution of geolocated tweets using statistical methods for answering several questions. It first focused on developing a meaningful approach to analyze 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 cities.

Additionally, this work aimed to characterize temporal and spatial structures in the data to capture the spatio-temporal behavior 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 characterize and compare behaviors across those urban environments which allowed to gain information involving human conduct in cities with different structures and dynamics.

We have found that our approach was able to identify almost the same behaviors 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 automatized and thus, analyze 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.

Author Contributions: Conceptualization: F.S.; methodology: F.S. and R.H.; validation and formal analysis: F.S.; supervision: R.H., J.T.-S., and E.P.; writing (original draft preparation): F.S.; and writing (review and editing): all authors.

Funding: The authors gratefully acknowledge funding from the European Union through the GEO-C project (H2020-MSCA-ITN-2014, grant agreement number 642332, <http://www.geo-c.eu/>).

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; and in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

FPCA	Functional principal component analysis
API	Application programming interface
LBSN	Location-based social networks
SOM	Self-organizing maps
ICA	Independent Component Analysis
DBSCAN	Density-based spatial clustering of Applications with Noise
OLS	Ordinary least-squares
GAM	Generalized additive models
LISA	Local indicators of spatial association
STSS	Space-time scan statistics
GMM	Gaussian mixture models
KDE	Kernel density estimation
LDA	Latent Dirichlet allocation
MAUP	modifiable area unit problem
CRS	Coordinate reference system
FDA	Functional data analysis
GLM	Generalized linear models
IWLS	Iteratively weighted least-squares
BIC	Bayesian information criterion
CSR	Complete spatial randomness
PCA	Principal components analysis
RMSE	Root mean squared error
MAE	Mean absolute error
MAPE	Mean absolute percentage error
sMAPE	Symmetric mean absolute percentage error
INLA	Integrated nested Laplace approximations

Appendix A

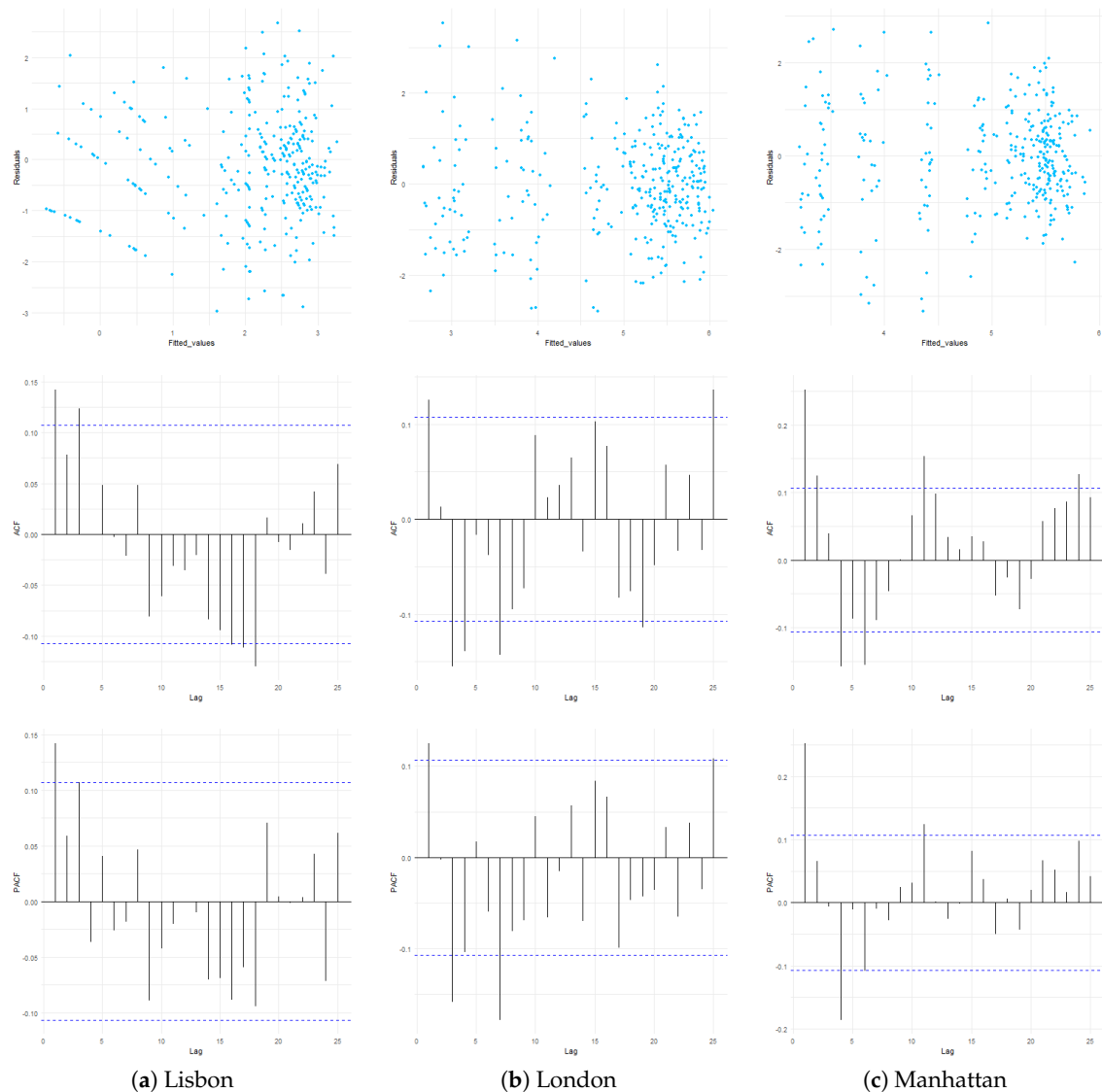


Figure A1. Residual plots for the selected regression models.

References

1. França, U.; Sayama, H.; Mcswigen, C.; Daneshvar, R.; Bar-Yam, Y. Visualizing the “heartbeat” of a city with tweets. *Complexity* **2015**, *21*, 280–287. [[CrossRef](#)]
2. Celikten, E.; Falher, G.L.; Mathioudakis, M. Modeling Urban Behavior by Mining Geotagged Social Data. *IEEE Trans. Big Data* **2017**, *3*, 220–233. [[CrossRef](#)]
3. Jiang, S.; Ferreira, J.; González, M.C. Clustering daily patterns of human activities in the city. *Data Min. Knowl. Discov.* **2012**, *25*, 478–510. [[CrossRef](#)]
4. Tasse, D.; Hong, J.I. Using social media data to understand cities. In Proceedings of the NSF Workshop on Big Data and Urban Informatics, Chicago, IL, USA, 11–12 August 2014; pp. 64–79.
5. Simini, F.; González, M.; Maritan, A.; Barabási, A.L. A universal model for mobility and migration patterns. *Nature* **2012**, *484*, 96–100. [[CrossRef](#)] [[PubMed](#)]
6. Song, C.; Qu, Z.; Blumm, N.; Barabási, A.L. Limits of predictability in human mobility. *Science* **2010**, *327*, 1018–1021. [[CrossRef](#)] [[PubMed](#)]

7. González, M.; Hidalgo, C.; Barabasi, A.L. Understanding individual human mobility patterns. *Nature* **2008**, *453*, 779–782. [[CrossRef](#)] [[PubMed](#)]
8. Brockmann, D.; Hufnagel, L.; Geisel, T. The scaling laws of human travel. *Nature* **2006**, *439*, 462–465. [[CrossRef](#)]
9. Batty, M. Cities as Complex Systems: Scaling, Interaction, Networks, Dynamics and Urban Morphologies. In *Encyclopedia of Complexity and Systems Science*; Meyers, R.A., Ed.; Springer: New York, NY, USA, 2009; pp. 1041–1071. [[CrossRef](#)]
10. Jackson, M.C. Social systems theory and practice: The need for a critical approach. *Int. J. Gen. Syst.* **1985**, *10*, 135–151. [[CrossRef](#)]
11. United Nations. *World Urbanization Prospects: The 2014 Revision, Highlights*; Technical Report ST/ESA/SER.A/352; Department of Economic and Social Affairs, Population Division: New York, NY, USA, 2014.
12. Vespignani, A. Predicting the behavior of techno-social systems. *Science* **2009**, *325*, 425–428. [[CrossRef](#)]
13. Thériault, M.; Des Rosiers, F. *Modeling Urban Dynamics*; John Wiley & Sons: Hoboken, NJ, USA, 2013.
14. Silva, T.H.; de Melo, P.O.S.V.; Almeida, J.M.; Loureiro, A.A.F. Social Media as a Source of Sensing to Study City Dynamics and Urban Social Behavior: Approaches, Models, and Opportunities. In *Ubiquitous Social Media Analysis*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 63–87. [[CrossRef](#)]
15. Huang, Q. Mining online footprints to predict user’s next location. *Int. J. Geogr. Inf. Sci.* **2016**, *31*, 523–541. [[CrossRef](#)]
16. Gao, H.; Liu, H. Data Analysis on Location-Based Social Networks. In *Mobile Social Networking*; Springer: New York, NY, USA, 2013; pp. 165–194. [[CrossRef](#)]
17. Ferrari, L.; Rosi, A.; Mamei, M.; Zambonelli, F. Extracting urban patterns from location-based social networks. In Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks, Chicago, IL, USA, 1 November 2011; ACM: New York, NY, USA, 2011; pp. 9–16.
18. Toole, J.; de Montjoye, Y.A.; González, M.; Pentland, A.S. Modeling and Understanding Intrinsic Characteristics of Human Mobility. In *Social Phenomena, Computational Social Sciences*; Gonçalves, B., Perra, N., Eds.; Springer: Cham, Switzerland, 2015; pp. 15–35.
19. Frias-Martinez, V.; Soto, V.; Hohwald, H.; Frias-Martinez, E. 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, Amsterdam, The Netherlands, 3–5 September 2012; IEEE Computer Society: Washington, DC, USA, 2012; pp. 239–248. [[CrossRef](#)]
20. Wakamiya, S.; Lee, R.; Sumiya, K. 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, Chicago, IL, USA, 1 November 2011; ACM: New York, NY, USA, 2011; pp. 77–84.
21. Stimmel, C.L. *Building Smart Cities: Analytics, ICT, and Design Thinking*; CRC Press: Boca Raton, FL, USA, 2015.
22. Zheng, Y.; Capra, L.; Wolfson, O.; Yang, H. Urban Computing: Concepts, Methodologies, and Applications. *ACM Trans. Intell. Syst. Technol.* **2014**, *5*, 38. [[CrossRef](#)]
23. Steiger, E.; de Albuquerque, J.P.; Zipf, A. An Advanced Systematic Literature Review on Spatiotemporal Analyses of Twitter Data. *Trans. GIS* **2015**, *19*, 809–834. [[CrossRef](#)]
24. Steiger, E.; Westerholt, R.; Resch, B.; Zipf, A. Twitter as an indicator for whereabouts of people? Correlating Twitter with UK census data. *Comput. Environ. Urban Syst.* **2015**, *54*, 255–265. [[CrossRef](#)]
25. Kaplan, A.M.; Haenlein, M. Users of the world, unite! The challenges and opportunities of Social Media. *Bus. Horiz.* **2010**, *53*, 59–68. [[CrossRef](#)]
26. Nummi, P. Social Media Data Analysis in Urban e-Planning. *Int. J. E-Plan. Res.* **2017**, *6*, 18–31. [[CrossRef](#)]
27. Gandomi, A.; Haider, M. Beyond the hype: Big data concepts, methods, and analytics. *Int. J. Inf. Manag.* **2015**, *35*, 137–144. [[CrossRef](#)]
28. Thakur, G.; Sims, K.; Mao, H.; Piburn, J.; Sparks, K.; Urban, M.; Stewart, R.; Weber, E.; Bhaduri, B. Utilizing Geo-located Sensors and Social Media for Studying Population Dynamics and Land Classification. In *Human Dynamics Research in Smart and Connected Communities*; Springer International Publishing: Cham, Switzerland, 2018; pp. 13–40. [[CrossRef](#)]
29. Huang, Y.; Li, Y.; Shan, J. Spatial-Temporal Event Detection from Geo-Tagged Tweets. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 150. [[CrossRef](#)]

30. García-Palomares, J.C.; Salas-Olmedo, M.H.; Moya-Gómez, B.; Condeço-Melhorado, A.; Gutiérrez, J. City dynamics through Twitter: Relationships between land use and spatiotemporal demographics. *Cities* **2018**, *72*, 310–319. [[CrossRef](#)]
31. Resch, B.; Usländer, F.; Havas, C. Combining machine-learning topic models and spatiotemporal analysis of social media data for disaster footprint and damage assessment. *Cartogr. Geogr. Inf. Sci.* **2017**, *45*, 362–376. [[CrossRef](#)]
32. De Albuquerque, J.P.; Herfort, B.; Brenning, A.; Zipf, A. A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management. *Int. J. Geogr. Inf. Sci.* **2015**, *29*, 667–689. [[CrossRef](#)]
33. Kim, E.; Helal, S.; Cook, D. Human activity recognition and pattern discovery. *IEEE Pervasive Comput./IEEE Comput. Soc. IEEE Commun. Soc.* **2010**, *9*, 48–53. [[CrossRef](#)] [[PubMed](#)]
34. Goodchild, M.F. Citizens as sensors: The world of volunteered geography. *GeoJournal* **2007**, *69*, 211–221. [[CrossRef](#)]
35. Frias-Martinez, V.; Frias-Martinez, E. Spectral clustering for sensing urban land use using Twitter activity. *Eng. Appl. Artif. Intell.* **2014**, *35*, 237–245. [[CrossRef](#)]
36. Soliman, A.; Soltani, K.; Yin, J.; Padmanabhan, A.; Wang, S. Social sensing of urban land use based on analysis of Twitter users' mobility patterns. *PLoS ONE* **2017**, *12*, e0181657. [[CrossRef](#)] [[PubMed](#)]
37. Resch, B.; Summa, A.; Zeile, P.; Strube, M. Citizen-Centric Urban Planning through Extracting Emotion Information from Twitter in an Interdisciplinary Space-Time-Linguistics Algorithm. *Urban Plan.* **2016**, *1*, 114. [[CrossRef](#)]
38. Hasan, S.; Zhan, X.; Ukkusuri, S.V. 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, Chicago, IL, USA, 11–14 August 2013; ACM: New York, NY, USA, 2013; p. 6.
39. Huang, W.; Li, S. Understanding human activity patterns based on space-time-semantics. *ISPRS J. Photogramm. Remote Sens.* **2016**, *121*, 1–10. [[CrossRef](#)]
40. Patel, N.N.; Stevens, F.R.; Huang, Z.; Gaughan, A.E.; Elyazar, I.; Tatem, A.J. Improving Large Area Population Mapping Using Geotweet Densities. *Trans. GIS* **2016**, *21*, 317–331. [[CrossRef](#)] [[PubMed](#)]
41. Huang, Q.; Wong, D.W.S. Activity patterns, socioeconomic status and urban spatial structure: What can social media data tell us? *Int. J. Geogr. Inf. Sci.* **2016**, *30*, 1873–1898. [[CrossRef](#)]
42. Cheng, T.; Wicks, T. Event detection using Twitter: A spatio-temporal approach. *PLoS ONE* **2014**, *9*, e97807. [[CrossRef](#)] [[PubMed](#)]
43. Shi, Y.; Deng, M.; Yang, X.; Liu, Q.; Zhao, L.; Lu, C.T. A Framework for Discovering Evolving Domain Related Spatio-Temporal Patterns in Twitter. *ISPRS Int. J. Geo-Inf.* **2016**, *5*, 193. [[CrossRef](#)]
44. Steiger, E.; Resch, B.; Zipf, A. Exploration of spatiotemporal and semantic clusters of Twitter data using unsupervised neural networks. *Int. J. Geogr. Inf. Sci.* **2015**, *30*, 1694–1716. [[CrossRef](#)]
45. Bakerman, J.; Pazdernik, K.; Wilson, A.; Fairchild, G.; Bahran, R. Twitter Geolocation. *ACM Trans. Knowl. Discov. Data* **2018**, *20*, 1–17. [[CrossRef](#)]
46. Diggle, P.J. *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*; CRC Press: Boca Raton, FL, USA, 2013.
47. Liboschik, T.; Fokianos, K.; Fried, R. tscount: An R Package for Analysis of Count Time Series Following Generalized Linear Models. *J. Stat. Softw.* **2017**, *82*. [[CrossRef](#)]
48. Baddeley, A.; Rubak, E.; Turner, R. *Spatial Point Patterns: Methodology and Applications with R*; CRC Press: Boca Raton, FL, USA, 2015.
49. Illian, J.; Penttinen, A.; Stoyan, H.; Stoyan, D. *Statistical Analysis and Modelling of Spatial Point Patterns*; John Wiley & Sons: Hoboken, NJ, USA, 2008; Volume 70.
50. Lee, D.J.; Zhu, Z.; Toscas, P. Spatio-temporal functional data analysis for wireless sensor networks data. *Environmetrics* **2015**, *26*, 354–362. [[CrossRef](#)] [[PubMed](#)]
51. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2018.
52. Aragón, P.; Juan, P.; Staab, J. tweet2r: Twitter Collector for R and Export to 'SQLite', 'postGIS' and 'GIS' Format, 2018. R Package Version 1.1. Available online: <https://cran.r-project.org/web/packages/tweet2r/tweet2r.pdf> (accessed on 15 November 2018).

53. Morstatter, F.; Pfeffer, J.; Liu, H.; Carley, K.M. Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. ICWSM, 2013. Available online: <https://arxiv.org/abs/1306.5204> (accessed on 15 November 2018).
54. Hawelka, B.; Sitko, I.; Beinat, E.; Sobolevsky, S.; Kazakopoulos, P.; Ratti, C. Geo-located Twitter as proxy for global mobility patterns. *Cartogr. Geogr. Inf. Sci.* **2014**, *41*, 260–271. [[CrossRef](#)] [[PubMed](#)]
55. Steinert-Threkeld, Z.C. *Twitter as Data*; Cambridge University Press: Cambridge, UK, 2018. [[CrossRef](#)]
56. Yin, J.; Gao, Y.; Du, Z.; Wang, S. Exploring multi-scale spatiotemporal twitter user mobility patterns with a visual-analytics approach. *ISPRS Int. J. Geo-Inf.* **2016**, *5*, 187. [[CrossRef](#)]
57. Tsou, M.H.; Zhang, H.; Jung, C.T. Identifying Data Noises, User Biases, and System Errors in Geo-tagged Twitter Messages (Tweets). *arXiv* **2017**, arXiv:1712.02433.
58. Pebesma, E.J.; Bivand, R.S. Classes and methods for spatial data in R. *R News* **2005**, *5*, 9–13.
59. Montgomery, D.C.; Peck, E.A.; Vining, G.G. *Introduction to Linear Regression Analysis*; John Wiley & Sons: Hoboken, NJ, USA, 2012; Volume 821.
60. Myers, R.H.; Montgomery, D.C.; Vining, G.G.; Robinson, T.J. *Generalized Linear Models: With Applications in Engineering and the Sciences*; John Wiley & Sons: Hoboken, NJ, USA, 2012; Volume 791.
61. Nelder, J.A.; Wedderburn, R.W.M. Generalized Linear Models. *J. R. Stat. Soc. Ser. A* **1972**, *135*, 370. [[CrossRef](#)]
62. Dobson, A.J.; Barnett, A.G. *An Introduction to Generalized Linear Models*, 4th ed.; Chapman & Hall/CRC Texts in Statistical Science; Chapman and Hall/CRC: London, UK, 2008.
63. Hilbe, J.M. Log negative binomial regression as a generalized linear model. *Grad. Coll. Comm. Stat.* **1993**, *1024*, 1–16.
64. McCullagh, P.; Nelder, J.A. *Generalized Linear Models*; CRC Press: Boca Raton, FL, USA, 1989; Volume 37.
65. Hardin, J.W.; Hilbe, J.M. *Generalized Linear Models and Extensions*; Stata Press: College Station, TX, USA, 2012.
66. Katsouyanni, K.; Schwartz, J.; Spix, C.; Touloumi, G.; Zmirou, D.; Zanobetti, A.; Wojtyniak, B.; Vonk, J.; Tobias, A.; Pönkä, A.; et al. Short term effects of air pollution on health: A European approach using epidemiologic time series data: The APHEA protocol. *J. Epidemiol. Community Health* **1996**, *50*, S12–S18. [[CrossRef](#)] [[PubMed](#)]
67. Venables, W.N.; Ripley, B.D. *Modern Applied Statistics with S*; Springer: New York, NY, USA, 2002. [[CrossRef](#)]
68. Cameron, A.C.; Trivedi, P.K. Econometric models based on count data. Comparisons and applications of some estimators and tests. *J. Appl. Econom.* **1986**, *1*, 29–53. [[CrossRef](#)]
69. Cressie, N. *Statistics for Spatial Data*; Wiley series in probability and mathematical statistics: Applied probability and statistics; John Wiley & Sons: Hoboken, NJ, USA, 1993.
70. O'Sullivan, D.; Unwin, D. *Geographic Information Analysis*; Wiley: Hoboken, NJ, USA, 2014.
71. Dale, M.; Fortin, M. *Spatial Analysis: A Guide For Ecologists*; Cambridge University Press: Cambridge, UK, 2014.
72. Illian, J.; Benson, E.; Crawford, J.; Staines, H. 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; Springer: New York, NY, USA, 2006; pp. 135–150.
73. Kokoszka, P.; Reimherr, M. *Introduction to Functional Data Analysis*; Chapman & Hall/CRC Texts in Statistical Science; CRC Press: Boca Raton, FL, USA, 2017.
74. Ramsay, J.; Hooker, G.; Graves, S. *Functional Data Analysis with R and MATLAB*; Springer: New York, NY, USA, 2009.
75. Ramsay, J.; Silverman, B. *Functional Data Analysis*; Springer Series in Statistics; Springer: New York, NY, USA, 2005.
76. Husson, F.; Lê, S.; Pags, J. *Exploratory Multivariate Analysis by Example Using R*; CRC Press: Boca Raton, FL, USA, 2017.
77. Scott, D.W. *Multivariate Density Estimation: Theory, Practice, and Visualization*; John Wiley & Sons: Hoboken, NJ, USA, 2015.
78. Blangiardo, M.; Cameletti, M. *Spatial and Spatio-Temporal Bayesian Models with R-INLA*; John Wiley & Sons: Hoboken, NJ, USA, 2015.
79. Rue, H.; Martino, S.; Chopin, N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. B* **2009**, *71*, 319–392. [[CrossRef](#)]
80. Illian, J.B.; Sørbye, S.H.; Rue, H. A toolbox for fitting complex spatial point process models using integrated nested Laplace approximation (INLA). *Ann. Appl. Stat.* **2012**, *6*, 1499–1530. [[CrossRef](#)]

81. Bivand, R.S.; Gómez-Rubio, V.; Rue, H. Spatial Data Analysis with R-INLA with Some Extensions. *J. Stat. Softw.* **2015**, *63*. [[CrossRef](#)]
82. Meyer, S.; Held, L.; Höhle, M. Spatio-Temporal Analysis of Epidemic Phenomena Using the R Package surveillance. *J. Stat. Softw.* **2017**, *77*. [[CrossRef](#)]



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).