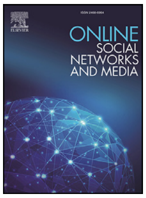
Online Social Networks and Media 1 (2017) 56–69

Contents lists available at ScienceDirect

Online Social Networks and Media

journal homepage: www.elsevier.com/locate/osnem



Using social media to characterize urban mobility patterns: State-of-the-art survey and case-study

Matteo Manca[a], Ludovico Boratto[a], Victor Morell Roman[b], Oriol Martori i Gallissà[b], Andreas Kaltenbrunner[a,*]

[a] Eurecat – Technology Centre of Catalonia, Av. Diagonal 177, 08018 Barcelona, Spain
[b] URBANing, Montserrat, Terrassa 95 - 08221, Spain

ARTICLE INFO

Article history:
Received 22 December 2016
Revised 6 April 2017
Accepted 6 April 2017

Keywords:
Urban informatics
Pattern mining
User behavior

ABSTRACT

The knowledge of the urban mobility is a crucial aspect for city planners and administrators. The huge amount of geo-spatial data, generated by the combination of social media systems and the wide use of smart devices, is creating new challenges and opportunities to satisfy this thirst of knowledge. In this work, we explore how social media data can be used to infer knowledge about urban dynamics and mobility patterns in a urban area. Specifically, in order to highlight the main advantages, limitations, and open issues, we focus on mobility patterns by presenting a survey of the state of the art and a case-study based on the city of Barcelona.

1. Introduction

Recent years observed an increasing trend to move from rural (non-urban) to urban areas. Indeed, more than half of the word populati on lives in cities [1] and this tendency will keep growing over the following years. It has been estimated that in the near future about the 9% of the world population will live in 41 very big cities.[1] The above mentioned urbanization is changing a lot of people's lives, often improving those, but at the same time cities are becoming more and more complex and dynamic. Therefore, new challenges, such as air pollution, traffic congestion, resource allocation, and mass tourism are continuously arising. To tackle these challenges and to improve the user's city experiences, administrators and city planners need to deeply know dynamics of the city and how they interact. Indeed, several actors, such as the user habits, mobility patterns, and most visited POIs (Points Of Interests), play a role in these dynamics.

During the past years, most of the studies to understand the trends of a city were based on citizen surveys [2]. Recently, other worthy sources of data have also been considered, for instance wireless sensors and mobile network data. In the following, we will review some details of the above mentioned sources of data.

Surveys. Surveys are able to provide accurate information about user residence, mobility patterns, and habits in general. However, this methodology presents several limitations, since it has high costs, which leads to surveys being usually applied to a small sample of the population. Moreover, the collected data is limited in space and time and, usually, surveys are updated with a very low frequency, thus complicating the job of the administrators in trying to make decisions that improve the quality of life in a specific city.

Wireless sensors. Wireless sensors and traffic cameras represent a valuable and alternative source of data to infer information about user behavior in a given area. Many research works exploit wireless sensor logs to obtain knowledge about urban user behavior and mobility patterns. Song et al. [3] developed and evaluated several location predictors using a dataset containing two years of traces collected from the Dartmouth College's wireless network. The authors of [4] use video surveillance cameras to study the citizens' behavior and social dynamics in St. Petersburg. Although some of the limitations related to classical surveys can be solved with wireless sensors and cameras (like the low update frequency and limitation in time), some others still keep existing like the high cost, due to the installation and management of the sensors [5], and the spatial limitation.

Mobile phone networks. Given the strong impact that mobile devices have had on our lives, another opportunity to gain a deeper knowledge about the city dynamics is given by mobile phone networks. Moreover, differently from citizens surveys, mobile networks allow to perform large scale studies. For instance, in [2]

[*] Corresponding author.
E-mail address: kaltenbrunner@gmail.com (A. Kaltenbrunner).

[1] http://www.economist.com/node/21642053

http://dx.doi.org/10.1016/j.osnem.2017.04.002
2468-6964/

authors implemented a model to extract mobility information from a dataset of about 1 million mobile phones in East Massachusetts.

Another study [6] analyses a six-month mobile phone dataset finding that human mobility is characterised by a high degree of temporal and spatial regularity. These results are confirmed in [7] where the authors exploit mobile phone data to highlight the lack of variability in mobility predictions. Pappalardo et al. [8] use mobile phone and GPS data to study user mobility, discovering two main classes of users: returners, who focus their mobility to a few locations and explorers, whose mobility is not limited to few locations.

Berlingerio et al. [9] implemented a system that uses the location of mobile phone data to identify travel patterns in a city with the aim to help decision makers to improve the public transport systems. In [10], Gabrielli et al. use mobile phone data to study the mobility behavior of visitors in a urban area. In [11], De Nadai et al. exploited the mobile phone data coming from an Italian service provider (Telecom Italia Mobile) to extract user activity and used it to test the four conditions proposed by Jane Jacobs in her book, “The Death and Life of Great American Cities”, in six Italian cities. From the above mentioned works it is obvious that mobile phone networks provide a useful tool to mine mobility patterns and user behavior in a city, thus allowing large scale studies and data updates with a very high frequency or almost in real time. Nevertheless, the main limitation (that also affects the previously mentioned sources of data) regards the not free and public availability of the data due to privacy, security, and proprietary reasons. Indeed, usually this data belongs to service providers and, to have access to the data, an agreement with the company is necessary [12]. Other sources of data like GPS traces and government data would also be worthy in understanding cities phenomena, but also this cases would be characterized by several restrictions in terms of data access.

Our contribution. Given the limitations highlighted for the previously described sources of data, this work focuses on a different possible one. Nowadays, the ever growing popularity of social media systems and the ubiquitous use of smart devices is generating a huge amount of data that is freely available and that covers all aspects of user behavior and life, such as the behavior in social media systems and Internet in general, economic activities, visited places, and user preferences and opinions. In such a scenario, users represent sensors that continuously generate a stream of data that can be exploited by everyone to infer information about collective user behaviors and to analyze what the users want to disclose about the so-called *spatial self* [13]. Therefore, this data represents a new challenge and a new opportunity for urban scientists, city planners, and administrators. Indeed, the use of social media data has several advantages, for instance it does not have temporal or spatial limitations, it allows large-scale studies, it is accessible (almost) in real time and without the need to sign any agreement with companies or service providers; the use of the API (Application Programming Interface) provided by the social media company is enough.

To highlight the capabilities of social media data in this context, this work investigates the following research question:

Research question 1. To what extent social media data can be exploited to gain knowledge about urban dynamics and mobility patterns in a city or in a urban area in general?

To highlight the main advantages, limitations, and open issues, we focus on mobility patterns by presenting a survey of the state of the art. It is worth to highlight that other surveys have been presented in the literature, focusing on spatio-temporal analysis of Twitter data [14], data analysis on location-based social networks [15,16], and urban computing in general [17]. However, none of the previously mentioned studies is devoted at analyzing how social media can be employed to mine urban mobility patterns. Therefore, our study represents the most extensive and up-to-date analysis of the literature in this area. This analysis of the literature also allowed us to define a workflow that can be followed in order to mine these patterns. To illustrate this workflow, we conducted a case-study based on the city of Barcelona, whose main goal was to make a comparison between mobility patterns of local citizens with respect to those of tourists. The city of Barcelona was chosen because of its massive flow of tourists, which makes the extraction of deep knowledge of movement dynamics essential for city administrators.

The scientific contributions of this work can be summarized as follows:

- this study represents the first literature survey that focuses on the employment of social media data to mine urban mobility patterns;
- we identify a workflow that shows how social media can become a source to extract these patterns;
- we apply this workflow into a real-world case-study on the city of Barcelona;
- we present open issues and future research challenges in this area.

Section 2 presents a state-of-the-art survey following all steps of the above mentioned workflow, Section 3 highlights the main insights coming from the case study, Section 4 contains open issues, future research challenges and concluding remarks.

2. Workflow for mining urban mobility patterns

In this section, we tackle the problem of mining urban mobility patterns using social media data. Based on an analysis of the literature in the field, we identified a set of tasks that is usually performed to solve the problem.

Thanks to this sequence of tasks, we identified a workflow that makes it possible to mine urban mobility patterns by employing social media data. The workflow is illustrated in Fig. 1. The first of these tasks deals with the data collection and preprocessing (an analysis of how it can be performed is presented in Section 2.1). After the data has been collected, each user can be profiled (an overview of the user profiling process is given in Section 2.2), each geolocated data point can be classified according to the mining purpose (the analysis of this task is presented in Section 2.3). Finally, the user profile and the classified geolocated data points can be employed to define the paths followed by the users and mine the mobility patterns (Section 2.4).

The presentation of each task of the workflow is structured as follows: we first present the problem definition, then introduce our proposal; after that, we show the applied proposal in a case-study of the city of Barcelona, and conclude with a survey on how other approaches in the literature solve this task.

2.1. Data collection and preprocessing

Data collection and preprocessing is a crucial and not negligible aspect of each data mining process. It has been estimated that the 80% of the whole data mining process consists of data preparation [18]. In this section this task is analyzed in the context of urban mobility pattern mining.

Most of the social media platforms use APIs to provide access to their data. Usually the APIs make multiple functions available based on a set of parameters that allow to perform several activities, such as to download a stream of data in real time, to specify a time window, to specify a set of keywords, or to specify a bounding box. The use of these functions represents a valuable

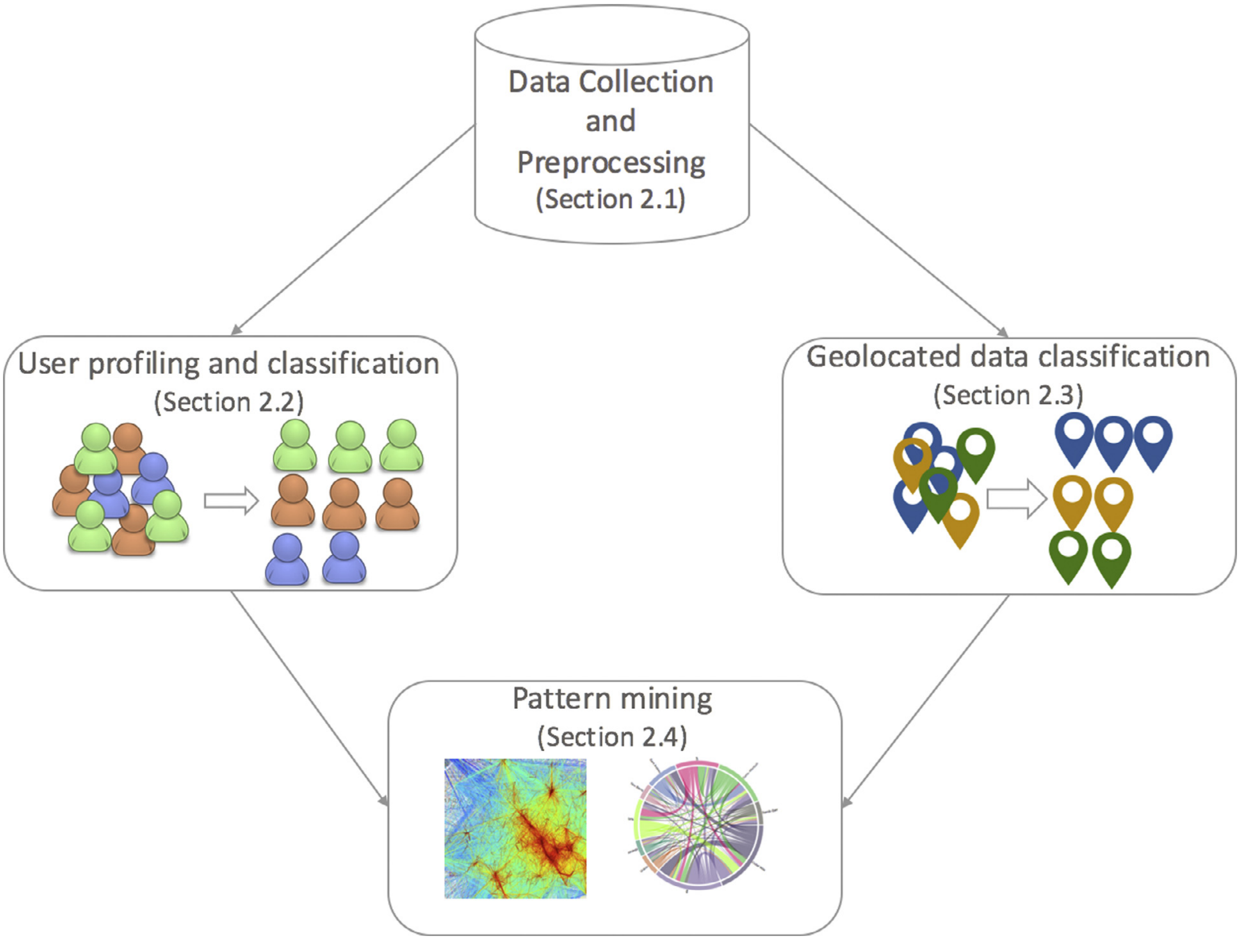

Data Collection and Preprocessing (Section 2.1)
User profiling and classification (Section 2.2)
Geolocated data classification (Section 2.3)
Pattern mining (Section 2.4)

Fig. 1. Mobility pattern mining workflow: each box indicates the section explaining details about the corresponding step.

method to create a collection of geolocated data coming from social media.

Before analyzing the approaches to data collection and preprocessing, we should present the different sources of data considered in the literature, for example *Twitter*, where a short message can be geolocated, or *Foursquare, Gowalla*, and *Brightkite*, where a check-in location is associated to a category of place (e.g., restaurant), and *Instagram* and *Flickr*, where a picture can be geolocated.

In the case of Twitter the most common approach to collect data for a specific area implies using Twitter's public Streaming APIs [2] defining a bounding box. These APIs retrieve all tweets that have the geo-coordinate field populated and that fall within the specified bounding box as well as all tweets without a value in the coordinate field, but with a value in the place field that corresponds to a geographic region which intersects with defined bounding box. All the remaining tweets will not be retrieved by the streaming APIs. However, the data collection process can be affected by a sampling bias since the amount of data retrieved through Twitter's public streaming APIs cannot exceed 1% of all tweets being tweeted at a specific moment. This means that if the result of the query consists of more than this threshold, a random sub-sample of these tweets will be retrieved corresponding to less than this 1%. The exceeding part will be lost to the data collector. A strategy to avoid reaching this limit is the combination of smaller bounding boxes and several accesses to the APIs in parallel from different machines.

An interesting aspect to notice related to data collection is that in several cases a social media site can become a source to collect the data of another social media site. This is especially true with Twitter, because of the simplicity with which the data can be collected. Indeed, the platform has been employed by several studies to collect Foursquare check-ins [5,19–24], both Instagram photos and Foursquare check-ins [25], or Foursquare and Gowalla check-ins [26]. The reason behind the use of Twitter for the data collection process is mostly related to the fact that, at the time in which these studies were conducted, the original platforms did not offer the data publicly. We can formulate the associated data collection task as follows.

Problem 1. How can we collect a set of geolocated objects coming from social media systems? And given this collection of data, how can a preprocessing task transform raw data into consistent data that can be analyzed? Common problems that characterize raw data are: presence of errors and outliers, missing values, and inconsistencies in the data.

2.1.1. Our proposed solution

Being the goal of the analysis the study of user mobility patterns, all objects of the dataset should contain geolocation information, i.e., latitude and longitude. So, a first needed preprocessing task consists of filtering out all data containing missing values for these fields or filling the missing values. Other filtering and preprocessing tasks are strictly related to the kind of analysis that is going to be performed. For instance if we were interested in analyzing objects within a specific geographical area, a solution based on the use of a shape file for that area of interest can be applied. In

[2] https://dev.twitter.com/streaming/overview

Table 1
Dataset description.

Number of geolocated tweets in Barcelona	1,120,216
Proportion of tourists' tweets	19%
Proportion of locals' tweets	81%
Number of unique users	93 946
Proportion of tourists	57.5%
Proportion of locals	42.5%

this case, a geometric merge between the point (latitude and longitude) of the data objects and the polygons of the geographical area is performed. Other preprocessing tasks could involve the selection of a specific time windows, removal of inactive users, data transformation and normalization, and any other operation that makes data consistent for the specific analysis.

2.1.2. Case-study

The initial dataset used in the proposed case-study was built by means of the Twitter Streaming APIs.

In a first stage, our data collection process filters data by location specifying a bounding box for Catalunya, i.e., two comma-separated pairs of longitude and latitude representing the coordinates of the bottom-left point and the coordinates of the top-right point: [0.1592, 40.523, 3.3326, 42.8615].

The obtained dataset, created specifying the Catalunya bounding box, contains $12,873,348$ tweets posted during the year 2015. The proportion of tweets has been far below the rate limit of 1% of the streaming APIs during the entire data collection period. We apply some preprocessing operations to select only tweets needed for our subsequent analyses. First of all, from the initial dataset we filter out tweets that do not contain geolocation information, but that were retrieved using the place field value (reducing the size of the dataset to $3,288,440$ tweets). Moreover, since our analysis focuses on tweets published in Barcelona, we need to filter out all tweets that do not fall within the area covered by the city. To achieve this and to be as accurate as possible we did not use a bounding box but we employed a shape file of Barcelona and selected all tweets that fell within those boundaries. Filtering out the tweets that were outside of the city of Barcelona, we obtain a final set of $1,225,199$ geolocated tweets (see Table 1). Moreover, we removed tweets from a user who published $104,983$ tweets from the same coordinates (latitude and longitude) as this was likely a bot account, which is no longer active and has probably been suspended. This left us with a dataset of $1,120,216$ tweets.

2.1.3. Other solutions and similar approaches

Data collection and preprocessing is an essential step that is usually driven by the type of mobility pattern that has to be mined. In order to analyze the everyday lives of people, Fuchs et al. [27] considered all the tweets of users who stayed in the Seattle area for at least 10 days during a two-month period, and outside for less than 10 days; the data was preprocessed to remove tweets that contained Foursquare logins (i.e., they only considered geolocated tweets). Similarly, another form of data collection that allows the approach to capture information about the life of the users focuses on the frequency with which a user posts geolocated data, and [22] considers only users who used Foursquare at least three times per day over one month. In [28], the area in which the patterns have to be mined can be specified through a query, so that the system can automatically retrieve the area of interest with a quadtree-based approach, and collect the data through the Twitter API. Other approaches focused on the data collection on a specific area, such as [29] that collected all the tweets in Manhattan for a 1-year period. Similarly, in [30] the authors aim to verify if the city of Louisville (Kentucky) is actually divided into east and west, according to the notion of the '9th Street Divide'; for this reason, they collect two datasets from Twitter of 703 users from the east and 662 from the west. Another approach that involves a location-based data collection is that of Girardin et al. [31], which collected Flickr data for two years in the province of Florence; moreover, data was collected before and after the users entered or left the province, in order to monitor inbound and outboud activities.

Apart from selecting the desired users and geolocated data, the preprocessing step is also aimed to remove noise in the data. In [26, 32] two collected consecutive locations are removed if the speed was higher than 1000 km/h (i.e., faster than a passenger plane). In order to preprocess the data, Jurdak et al. [33] measure the displacement distribution, known as *spatial dispersal kernel* $P(d)$, where d is the distance between a user's two consecutive reported locations; the function that characterizes the displacement distribution $P(d)$ is then used to remove all the displacements lower than 10 m.

2.2. User profiling and classification

Now we address the problem of identifying different classes of users in a specific area. As stated in the Introduction, city planners and administrator are interested in studying different aspects of urban areas. As a consequence, there often exists the need to identify different classes or types of users. We can formulate this as follows:

Problem 2. Given a set of social media users, how can we identify a discrete number of categories, based on specific criteria, and profile each user by assigning her to a given category? The initial set of users should be segmented in multiple classes or groups (for instance all city users could be grouped in locals and tourists).

2.2.1. Our proposed solution

Based on the domain of application, different user profiling and classification techniques can be applied. For instance, a user could be classified as "active" or "passive" based on the number or frequency of her social media posts. Another possible user profiling method consists of classifying the user as "local" or "tourist" by analyzing the number of user's active days (for instance by looking at the dates of the last and the first posts) in a specific area or the user's default location. Precisely, to separate the initial set of users into "locals" and "tourists", we propose the heuristic presented in Algorithm 1. Given a user u, the algorithm checks the *userLocation* field of her tweets. If this field has a value among those contained in a set S of predefined user locations, then the user is classified as *Local*. If the *userLocation* field does not have a value or it has any other value, the algorithm computes the number of days between the first and the last tweet posted by the user in Barcelona. If this number is higher than 20, the user is considered a *Local*, otherwise she is classified as a *Tourist*.

2.2.2. Case-study

In the Barcelona case-study we separate the initial set of users into two subsets, *Locals* and *Tourists*, by applying the Algorithm 1.

As first step we need to define: I. the set S of specific user locations that make a user a *local* user, and II. the minimum number of days Δ to consider a user as a *local*. The set S was defined as $S =$ [*"bcn", "barcelona", "badalona", "hospitalet"*]; so, if a user had one these values in her user location, she is considered a *local*. Although at the state of the art a user is usually considered a tourist when she publishes all her objects within 30 days [31,34], we consider this value too high for our case-study, because according to a recent study[3], the average number of nights a tourist stays in

[3] http://professional.barcelonaturisme.com/imgfiles/estad/Est2015.pdf

Algorithm 1 User profiling heuristic.

```
Let $T = \{t_i\}_{i=1}^{m}$ be a set of all geolocated data objects;
Let $U = \{u_i\}_{i=1}^{n}$ be the set of all users that published the objects in $T$;
Let $S$ be a set of specific user locations;
for $i = 1, \ldots, n$ do
    $u = U[i]$                                        ▷ User
    if $userLocation(u)$ in $S$ then
        $u = Local$
    else
        Let $t_{last}$ be the last objects of $T$ published by the user $u$;
        Let $t_{first}$ be the first objects of $T$ published by the user $u$;
        Let $\delta_u = t_{last} - t_{first}$ be the difference in days between the last and first object of $u$;
        if $\delta_u > \Delta$ then
            $u = Local$
        else
            $u = Tourist$
        end if
    end if
end for
```

Barcelona accommodations is much lower. We therefore decrease the number of days from 30 to 20 (i.e., $\Delta = 20$).

After applying the preprocessing tasks and the user profiling algorithm, we obtained the dataset reported in Table 1.

2.2.3. Other solutions

Given the geolocated data of a user, a user profile in the form of a vector that characterizes her preferences can be formed. In [35], Jin et al. build a vector whose elements are the points that Foursquare awarded to the user in the considered week.[4] In [27], the authors mine mobility patterns associated to the lifestyle of the users and define 22 categories represented by keywords (such as food, family, etc.) and profile each user based on the relevance of each category for her (i.e., the relative frequency with which the keyword occurred).

The separation between locals and tourists has been done by several approaches, for different purposes. Some of them make this split in order to analyze migration patterns, such as [36], in which Cao et al. analyze the spread of the Influenza like Illness (ILI) infection by monitoring Twitter posts; in their model, the authors also consider if the user is a resident or a visitor. In order to study global mobility patterns, the approach in [32] assigns as the country of residence of a user the one in which she tweeted the most, and in all the other countries she is considered a visitor. Other approaches, instead, make the separation since they are interested in analyzing either the locals or the tourists. In order to provide space–time visual analytics of where the Seattle locals tweet and what they talk about, in [37] the authors profile a user as local or visitor, by counting the days in which a user tweeted inside and outside Seattle. In order to mine the mobility patterns of tourists in Florence, in [31] all the users who posted geolocated Flickr photos for less than 30 days in the province are considered.

2.3. Geolocated data classification

In the previous section, the problem of profiling each user according to specific criteria was addressed. Now we focus on the profiling of each geolocated data object, obtaining in this way a segmentation of the initial dataset. More precisely, we address the following problem:

Table 2
Number of tweets per district.

Ciutat Vella	333,183
Eixample	245,517
Gràcia	63,775
Horta-Guinardó	55,490
Les Corts	74,238
Nou Barris	49,626
Sant Andreu	55,936
Sant Martí	136,201
Sants-Montjuïc	137,328
Sarrià-Sant Gervasi	73,905

Problem 3. Given a set of geolocated data objects, how can we find a set of categories and identify to which category each object belongs? The main objective of a solution to this problem would be to segment the initial set of geolocated objects in to multiple groups.

2.3.1. Our proposed solution

Most of social media data objects contain information about the date and the time they were published. This information might be exploited to classify data objects with respect to different criteria, such as day of the week (Monday, Tuesday, etc.), period of the day they were published (morning, evening, night), period of the week (working day or weekend), month name, or season. With respect to the position, the classification task might assign as class the district (neighborhood, zone, street, etc.) name where the object was published from.

2.3.2. Case-study

According to the needs of our analysis, the data of the Barcelona case-study was classified according to time and location. Precisely, we classify each tweet either as “weekend” or “working day” tweet based on the day it was posted. The obtained dataset is composed by 370,942 tweets, classified as “weekend” and 854,257 as “working day”. Moreover, based on the latitude and longitude fields we add a label to each tweet indicating the district name it was posted from. The result of this classification task is presented in Table 2 where the number of tweets for each district is shown.

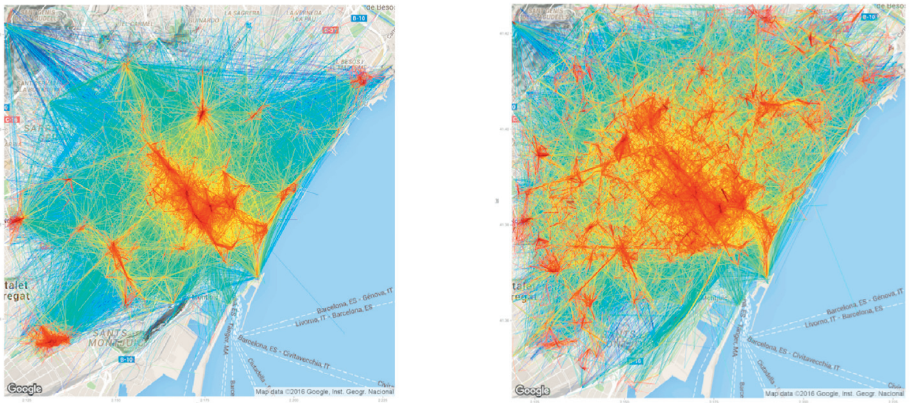
2.3.3. Other solutions

Most approaches classify a social media data object as belonging to a geographic area, in order to analyze the mobility patterns in that area. In [38], all the tweets associated to an area defined by a Voronoi diagram[5] are considered. In [39], each geolocated tweet is given as input to the k-means clustering algorithm, which defines “Regions of interest” (RoIs), i.e., close places with the same tweeting activity (therefore, each tweet is assigned to a cluster). Hasan et al. [40] assign each Foursquare check-in to a 200 m × 200 m square into which a city is divided and the squares are ranked by popularity for the subsequent pattern mining step.

Other approaches present a topic-based classification of geolocated tweets. Given the 22 categories associated to the lifestyle of the users, defined in [27] as keywords, each tweet was categorized according to the keywords it contained. In order to evaluate the happiness associated to a pattern, Frank et al. [41] measure the degree of happiness with respect to the covered distance in a travel; therefore, the happiness of each tweet in the path is computed by comparing its content with a 10,000-word dictionary, in

[4] Foursquare used to award points, usually based on the users status related to the venue (e.g., a user was a *mayor* of a venue if she was the person who checked into that venue on more days than anyone else in the past 60 days).

[5] Given a seed point, an area of a Voronoi diagram is represented by the points closer to that point than to any other seed point.



(a) Tourist paths

(b) Local paths

Fig. 2. Paths performed by users in Barcelona. Shorter paths are visualized through warmer colors, the longer a paths the colder the color tone.

which each word contains a score from 1 (sad) to 9 (happy). In order to analyze the spread of the ILI influenza, in [36] a tweet is “flu-flagged” if it contains a set of keywords, such as flu, cough, sneeze, and fever.

Another case is represented by the association of a check-in to a friend’s house; indeed, if the location of the check-in of a user A is close to the house of her friend’s B, the check-in is classified as a visit of a user to B for the analysis of the social relations in mobility patterns [42].

2.4. Pattern mining

Given the geolocated data points of each user, in order to mine the mobility patterns it is necessary to form a *path* or *trajectory*, that puts together in an ordered way all the places visited by a user. Once all the information about each user has been collected (both in terms of paths and of category to whom she belongs), and the places she visited have been classified, the mobility patterns can finally be mined. The goal of the current section is to investigate the following research problem:

Problem 4. Given a dataset of social media users and their related geolocated data objects (posts), how can we extract the user paths avoiding those that are too short, that span over too long time periods, or paths that involve more than one day (where the date of the source point is different with respect to that of the destination point)?

2.4.1. Our proposed solution

We propose to extract user *paths* or *trajectories* using two different approaches depending on whether the analysis focuses only on one-hop paths, i.e., paths that only involve two points (a source and a destination point), or multi-hop paths, i.e., paths that involve more than two points. The proposed approaches are based on the following definitions:

Definition 1. Let (t_1, t_2) be a pair of consecutive posts by the same user u. We consider (t_1, t_2) a user path (trajectory) *iff* t_1 and t_2 have been published during the same day, the distance in meters between t_1 and t_2 is higher than a threshold d and the difference in hours between t_1 and t_2 is lower than δ_t.

Definition 2. Let $(t_1, \ldots, t_n)$ be a set of consecutive posts of the same user u. We consider $(t_1, \ldots, t_n)$ a user’s path (trajectory) *iff* the ratio between the distance in kilometers and the time expressed in hours between two consecutive points t_i and t_{i+1} is in the range $[min\_speed, \max\_speed]$.

2.4.2. Case-study

Given the set of geolocated data described in Table 1, we exploit Definition 1 to extract user paths with parameters $d = 150$ m and $\delta_t = 10$ h. The measure of 150 m was chosen because of the structure of Barcelona. Indeed, Barcelona is structured as a set of squared blocks and we estimated that the distance between two parallel consecutive streets is approximately 150 m. So a pair of tweets is considered a trajectory if those tweets have been published in different blocks. Precisely, for each user we collect all her tweets, sort those by timestamp and extract each possible pair of tweets that respects Definition 1. As result, we obtain a total of $165,998$ user paths, $41,626$ performed by tourists and the remaining $124,372$ by local citizens. To provide a complete overview of the mobility patterns, we plot all user paths in two different maps of Barcelona, one for the tourists and the other for the locals (see Fig. 2a and b). These maps have been created using ggmap [43] (a R library for the visualization of spatial data)[6].

The maps in Fig. 2 show the different mobility behavior between tourists and locals, based on one-hop paths. In general tourists tend to center in specific zone of the city characterized by well known points of interest (i.e., Sagrada Familia, Camp Nou, Font Montjuïc, Parc Güell) and in central neighborhoods (like Gotic, Born, and Raval). While, as shown in Fig. 2b, the paths of the local users are more spread all around the city, coinciding in many cases with the main traffic arteries.

In a second step, we study paths that also involve more than two points focusing just on walking paths. So, considering that the average walking speed of a normal-weight adult is about 5 km/h [44] and allowing some fluctuations around this value, we apply Definition 2 setting $min\_speed = 1$ and $max\_speed = 7$. The obtained paths (a total of $45,638$ paths, $34,382$ for local users and $11,256$ for tourists) consists of (two or more) consecutive points that have been covered with a speed that falls in this speed range.

2.4.3. Other solutions

According to the type of approach, each point in a path might include just the place visited by the user, or take the form of a tuple that includes other information such as the time, the category of the venue (in case of check-in data), or the content of the tweet.

However, most of the approaches represent a path as a sequence of $<$ *location, timestamp*$>$ pairs. There are studies that, instead of considering the timestamp, consider a time window, such as [45].

[6] The longer paths have been represented with cool colors and the shorter ones with warmer colors.

A path can also be represented as a graph, ignoring the specific timestamp in which the geolocalization occurred, but simply modeling the sequence of events. In [5,21,23] each location is represented as a node and there is a direct edge from a node to another if the two locations where visited subsequently.

When the mobility pattern is mined considering the content of a social media data object, the text itself is part of the path, since each point in it is represented as a tuple of the type < *location, timestamp, content*> . An example of approach that includes also the content in the tuple, is presented in [36], where patterns are mined to detect influence spread. In the following, different pattern mining approaches are presented.

Clustering-based approaches. A set of approaches cluster the individual user paths, to discover how an area has been used by the users. The most-widely known clustering algorithm in the literature is k-means [46]. The objective of the algorithm is to split a set of n objects (vectors) into k groups (clusters), such that the objects that belong to a cluster are closer to its center than to the center of any other cluster. The algorithm has been employed in [38], to detect four clusters in Manhattan, by showing different tweeting behavior in each area, according to the time in the day or the day of the week. A 72-element vector for each geographic area and each day is built, where each component contains the number of tweets generated in that area in a 20-minute interval. The vectors generated during weekdays and weekend days have been separated and the vectors that belong to the same category have been averaged. The average weekday vector and weekend vector for each area are then merged into a 144-element vector that represents the activity in that area and is given as input to the clustering algorithm. Other clustering algorithms are also used to detect the patterns. In [45], the paths are grouped thanks to a spatial clustering algorithm, and visualised into 3D maps that show both the spatial and temporal dimension in which an area has been used. Cranshaw et al. [20] represent each venue as a vector, each element of the vector is associated to a user and it contains the number of check-ins of that user for that venue. Venues are clustered with a spectral clustering approach, in order to find which areas of a city are characterized by the same dynamics. Jin et al. [35] also employ a clustering algorithm (more specifically, Non-Negative Matrix Factorization), in order to capture the temporal and spatial characteristics of a users consecutive Foursquare check-in activities, according to the score-based user profile previously presented.

Model-based approaches. Model-based approaches usually build models that consider a set of observed geolocated data points in a path and assign a category to which this set of points belong (i.e., a pattern). An important part of the literature is characterized by adaptations of the Latent Dirichlet Allocation (LDA) algorithm [47] for the detection of mobility patterns. LDA is a topic modeling algorithm, which takes into account a sequence of words that form a document and analyzes the corpus formed by all the documents, in order to detect the topics that characterize the corpus; this is done in an unsupervised way, thanks to a Bayesian Network. Long et al. [48] discover local geographic topics from Foursquare check-ins, using LDA model applied to the user paths (each document is formed by a sequence of Foursquare check-ins in a day); the discovered patterns showed that venues that appear together in many users paths can be taken as geographic topics. Similarly, Ferrari et al. [29] applied LDA to paths formed by analyzing sequence of *crowd-footprints* in a day (a *crowd-footprint* represents the most crowded place at a given time) and discovered 30 geographical topics that characterize Manhattan. In order to detect the patterns followed by the users affected by the ILI (Influenza like Illness) influence, [36] presents a *spatiotemporal data cube model*, which considers the temporal, spatial, and human dimensions of the paths. Cho et al. [42] developed a model that considers the spatial locations that a user regularly visits, the temporal movement between these locations, and a model of movement that is influenced by the ties of the social network; results show that users have a strong periodic behavior throughout certain periods of the day, alternating between home and work during weekdays, and between home and social network driven locations on weekends. Liu et al. [49] evaluate if check-in data can be well fitted by the gravity model (which studies the interactions between two places and is used to estimate traffic and migration flows), and found out that the observed spatial interactions are governed by a power law distance decay effect; moreover, the authors found out that both inter- and intra-urban displacements follow an exponential distribution without the heavy-tail property. In [24], the authors present a model that views an activity pattern as a multinomial distribution of activity labels (each activity label is a Foursquare check-in) and individual activities as a mixture of activity patterns. Moreover, the model is extended in two ways, in order to capture user-specific patterns and identify the top users that contribute to a specific pattern, and to account for missing activities. Yuan et al. [50] present a probabilistic generative model, named W^4 (Who, where, when and what). Thanks to it, it is possible to mine mobility patterns from four factors, i.e., user, geographic information, time, and activity.

Path-distribution-based approaches. Other approaches study the distribution of the data points in a path, in order to analyze the mobility patterns. In [19], Noulas et al. study the complementary cumulative distribution function of Foursquare check-ins and find out that 20% of them cover a distance of 1 km, 60% are between 1 and 10 km, 20% take place at distances over 10 km, and around 5% go beyond 100 km. Cheng et al. [26], by analyzing check-ins in Foursquare and Gowalla, found that the distance-based displacement of consecutive check-ins made by users follows a power-law distribution, and that the return probability is strongest for places the users had visited most recently. In order to study if the city of Louisville (Kentucky) is characterized by the so-called “9th Street Divide”, in [30] the authors studied mobility patterns, observing that the two neighborhoods of the city can be considered as fluid, as people move from east to the west side of the city (and viceversa). Girardin et al. [31] consider the paths built with Flickr photos and build inbound and outbound maps that show how tourists move, and analyzed the most frequent flows. The results show that Americans follow a specific graph constituted by the nodes of Florence, Siena, Pisa, Genova and Perugia, while Italians are more adventurous. The authors of [40] mine patterns considering Foursquare check-ins (both the venue and the category) and the popularity of the area; results show a correlation between the popularity of a place and the probability to select this place as a destination.

3. Case-study outcomes

This section reports the main insights of the Barcelona case-study presented in Section 2. Fig. 3 shows the distribution of tourists and locals in all districts of Barcelona, emphasizing a stronger presence of tourists in the central districts (Eixample and Ciutat Vella), while the presence of locals is more spread all around the city.

Moreover, we analyze how users move inside and throughout the districts of the city during working days and during the weekend. Table 3 reports the proportion of paths inside the same district and across multiple districts, performed by tourists and locals. The reported statistics highlight a similar behavior of locals and tourists during working days and during weekends, i.e., about 60% of all paths involve more than one district To deepen this aspect of

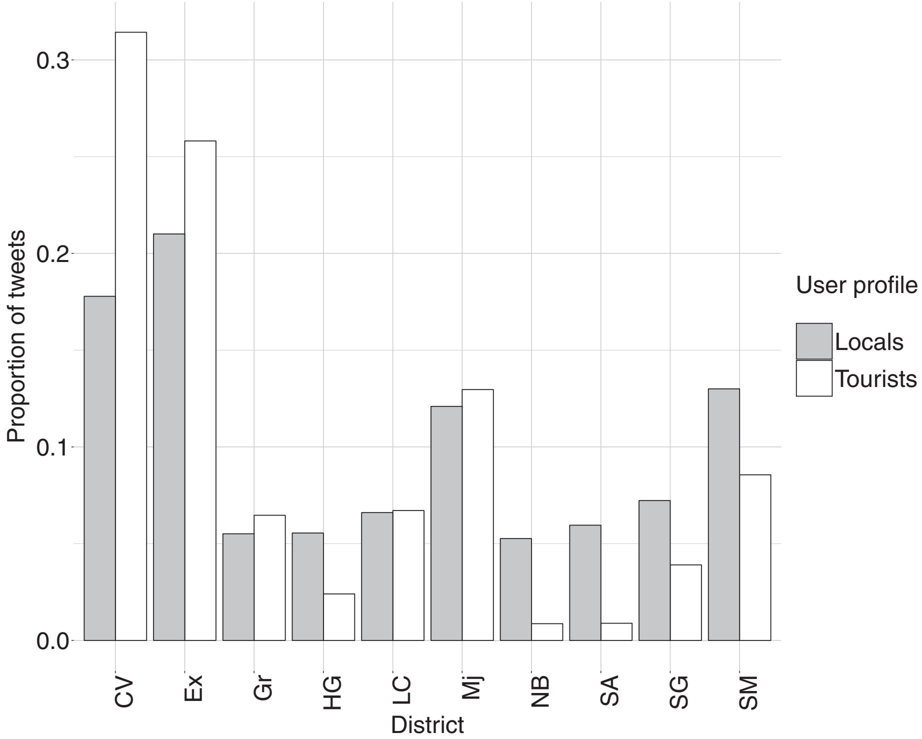


Fig. 3. Tweets distribution per district of Barcelona in 2015. CV: Ciutat Vella, Ex: Eixample, Gr: Gràcia, HG: Horta-Guinardó, LC: Les Corts, NB: Nou Barris, SA: Sant Andreu, SM: Sant Martí, Mj: Sants-Montjuïc, SG: Sarrià-Sant Gervasi.

Table 3
Paths inside and across districts.

	Inside	Across
Locals weekend	0.36	0.64
Locals working days	0.38	0.62
Tourists weekend	0.39	0.61
Tourists working days	0.4	0.6

user paths and behavior, we study separately paths inside districts and paths that involve multiple districts. Fig. 4 shows the proportion of single district paths per district segregated by locals and tourists. We observe that almost 50% of the tourist's single district paths were performed inside of Ciutat Vella, i.e., the most central district of Barcelona (and also among its most touristic ones). As expected, when analyzing the overall proportion of tweets per district (Fig. 3), there are only very few single district paths in more residential districts like Nou Barris and Sant Andreu. Furthermore, we analyse the proportion of incoming paths, outgoing paths and single district paths normalized per district. (Fig. 5). The obtained results show that in general the proportion of incoming paths and outgoing paths are very similar and usually higher than the proportion of single district paths. An exception of this pattern is represented by the behavior of the tourists in the district of Ciutat Vella, where the proportion of single district paths inside the district is much higher than in the other districts and very close to the proportion of incoming and outgoing paths.

After this first analysis we want to gain more knowledge about paths across districts. Fig. 6 shows the mobility behavior of both classes of users, i.e., locals and tourists, during working days and during weekends with respect to the districts of the city. These figures report only the paths where the source tweet and destination tweet belong to different districts.

Observing the chord diagrams in Fig. 6, several aspects leap out. Fig. 6a and b show that, although almost all districts are populated during weekend and during working days, the most central ones (like Ciutat Vella) are more visited during weekends to the detriment of others like Nou Barris. As expected, differently from local users, tourists have the same behavior during working days and during weekends. Fig. 6c and d report the mobility of tourists, respectively during working days and during weekend days, throughout the Barcelona districts. We can also notice that most of the tourists paths involve the two most touristic districts of Barcelona, i.e., Ciutat Vella and Eixample, while other districts like Sant Andreu and Nou Barris are almost ignored by these users.

Another aspect we analyse is the length of the travels. Fig. 7a, shows the paths distribution for locals and tourists by means of two boxplots which report minimum, first quartile, median, third quartile, and maximum. Observing the figure, we can notice that both, locals and tourists, have approximately a median of 2000 m and that the distances of tourist paths are less spread with respect to those of locals.

Fig. 7b reports the cumulative distribution function and allows to better understand the different behavior of locals and tourists. The plot points out that locals are more likely to cover short or long distances, while tourists are more common to cover intermediate distances between approximately 500 and 5000 m.

Until now we have only considered one-hop paths, i.e., paths composed by only two points: a source point and a destination point. Fig. 8 shows the average distance per hop with respect to the number of path hops. The figure highlights that the average distance per hop is inversely proportional to the number of path hops, i.e., the average distance decrease as the number of path hops grows. However this trend is weaker for tourists. Moreover, independently of the number of path hops, tourists are inclined to perform on average paths that involve longer hops in comparison to those of the locals. Future research might also investigate how to exploit these spatio-temporal features to cluster paths and discover possible groups of users that move together.

To gain more knowledge about patterns in user behavior, we also analyze if users tend to tweet frequently in a small set of specific locations. To this end we round each geographic co-

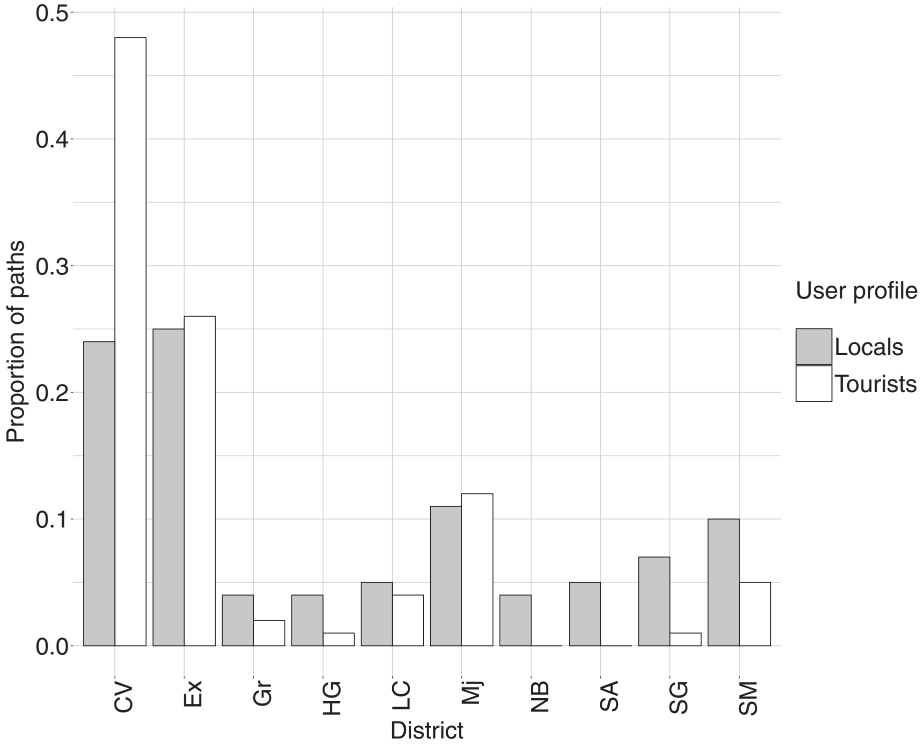


Fig. 4. Distribution of single district paths per district for locals (gray bars) and tourists (white bars). CV:Ciutat Vella, Ex: Eixample, Gr: Gràcia, HG: Horta-Guinardó, LC: Les Corts, NB: Nou Barris, SA: Sant Andreu, SM: Sant Martí, Mj: Sants-Montjuïc, SG: Sarrià-Sant Gervasi.

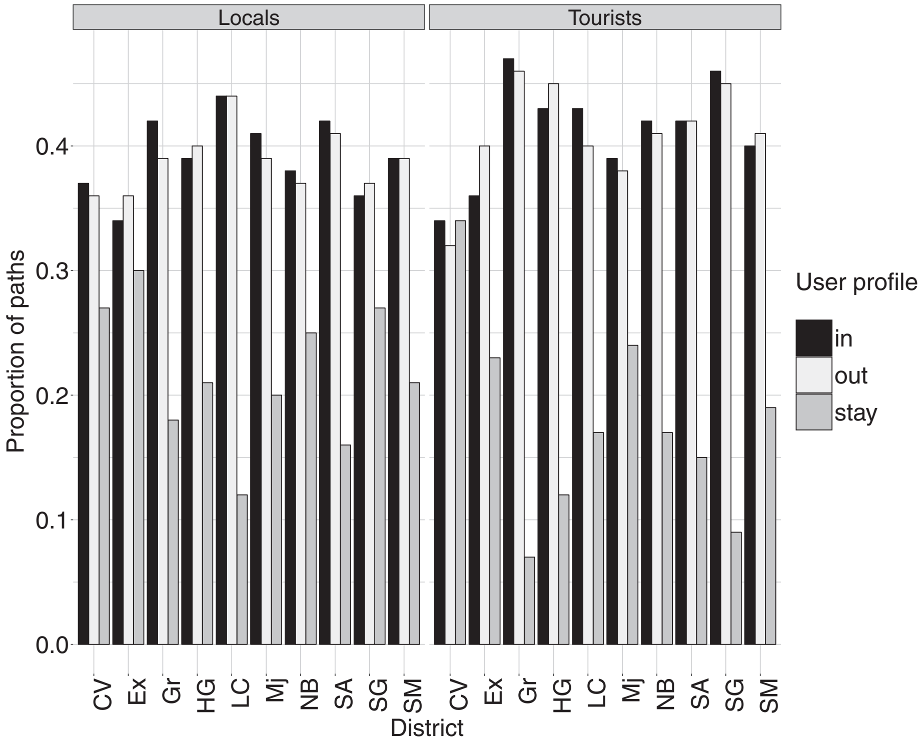
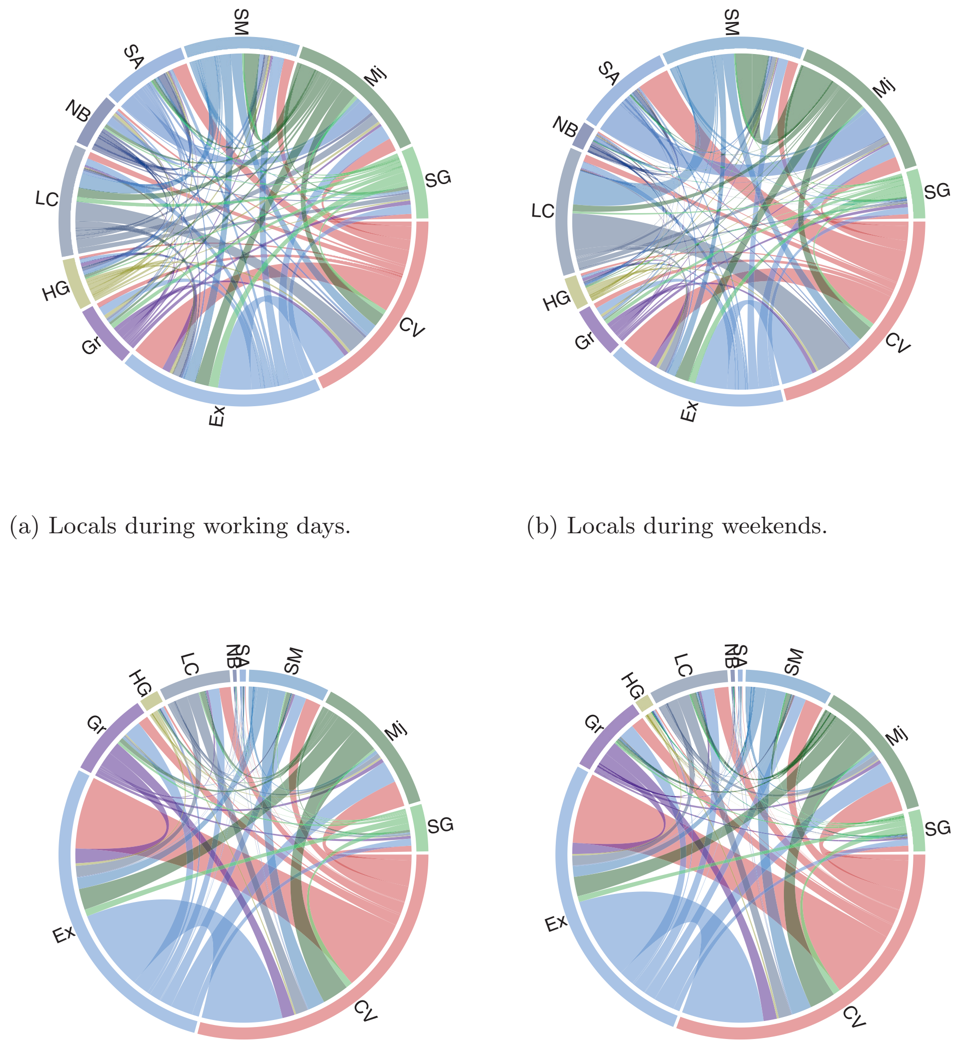


Fig. 5. Proportion of incoming paths, outgoing paths and inside paths for each district (CV:Ciutat Vella, Ex: Eixample, Gr: Gràcia, HG: Horta-Guinardó, LC: Les Corts, NB: Nou Barris, SA: Sant Andreu, SM: Sant Martí, Mj: Sants-Montjuïc, SG: Sarrià-Sant Gervasi.).

ordinate to three decimal digits, which splits our area of interests (the city of Barcelona) into small rectangular grids of approximately 110 m × 110 m. The centroids of each of these sub-areas are the rounded coordinates and each point in a sub-area is closer to the corresponding centroid than to any other centroid of the other sub-areas (i.e a rectangular Voronoi diagram).

Fig. 9 reports the probability (solid lines) that active users (locals and tourists) visited exactly L of these locations. The set of active users has been obtained after removing user with less than 20 tweets. The shaded areas represent the proportions of users that



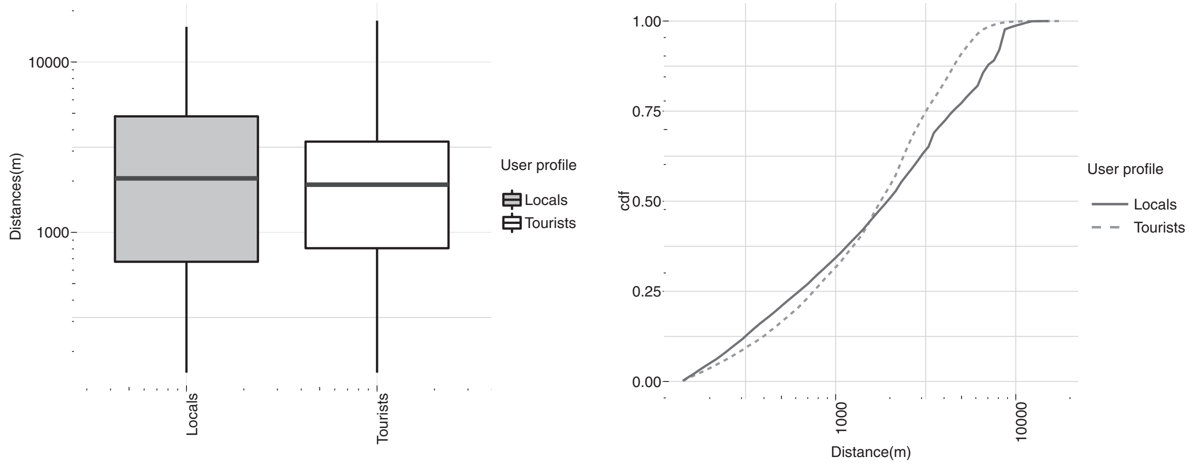
(a) Locals during working days.

(b) Locals during weekends.

(c) Tourists during working days.

(d) Tourists during weekends.

Fig. 6. Flow diagrams of locals and tourists between districts comparing working days and weekends. CV: Ciutat Vella, Ex: Eixample, Gr: Gràcia, HG: Horta-Guinardó, LC: Les Corts, NB: Nou Barris, SA: Sant Andreu, SM: Sant Martí, Mj: Sants-Montjuïc, SG: Sarrià-Sant Gervasi.



(a) Boxplots with paths distribution

(b) Cummulative distribution function

Fig. 7. Path statistics for tourists and locals (boxplots show minimum, first quartile, median, third quartile, and maximum of path distances).

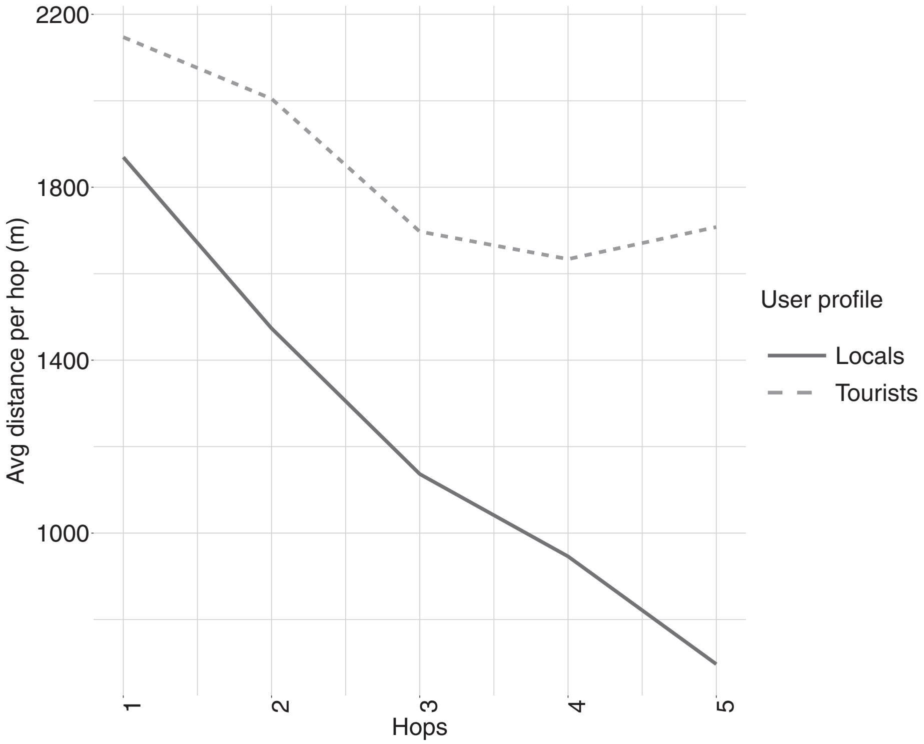


Fig. 8. Average distance per hop.

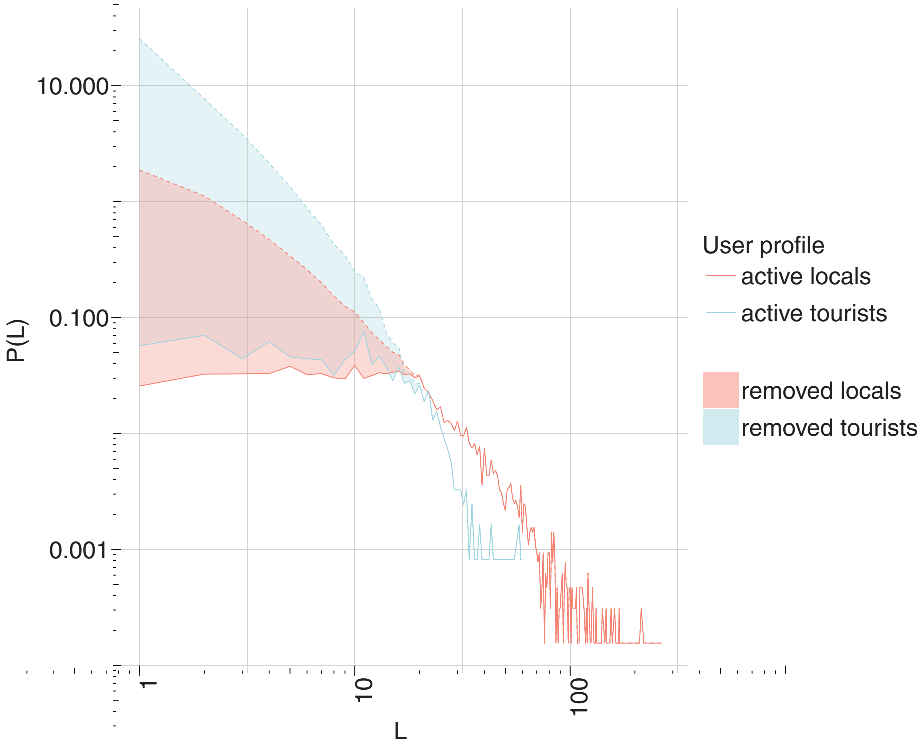


Fig. 9. Probability that a user visits L locations. The solid lines are the probabilities computed after removing (inactive) users with less than 20 tweets, the shaded areas represent the relative proportions of users who have been removed in relation to the active users.

have been removed in relation to the active users (note that these areas are not probabilities). The figure highlights that the probability that the users visit a small number of location is slightly higher for tourists. A limitation of our analysis is that we do not consider the *time* variable, i.e., the time span during which the users are active which might explain this result.

After this initial analysis of the number of visited areas per user we ranked for the active users the areas from which they tweeted based on the corresponding tweet-frequency. We compute the average frequency of each of the top-L ranked locations for all users who tweet from at least, $L = 5,$ 10, 30 or 50 different locations. Our findings (reported in Fig. 10) confirm in most cases the results

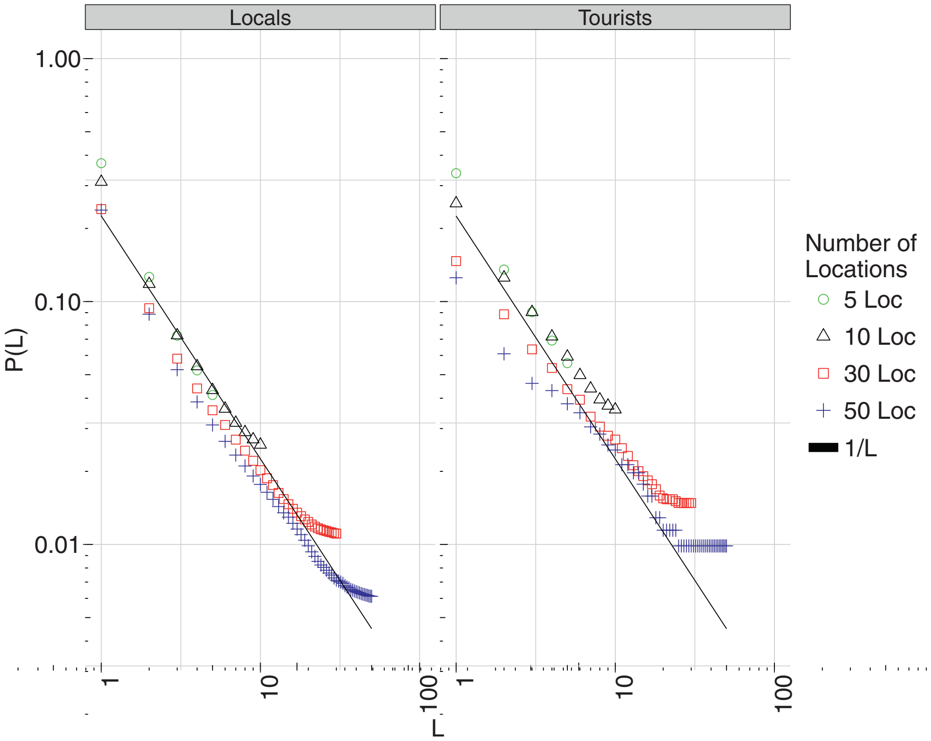


Fig. 10. Frequency-ranking of top-L most frequent tweet locations for users who tweet from at least for 5, 10, 30 or 50 different locations. Solid line represent a $1/L$ approximation.

presented in [6], i.e., that the probability $P(L)$ to find a user in a location with ranking L can be approximated by the function $1/L$ (solid line in Fig. 10). This is especially true for local users; however, the figure highlights as well that the proportion of tweets from the the most frequent areas of tourists who tweet from many different areas (more than 30) is lower than the one observed for locals. In this case the approximation with the function $1/L$ overestimates this probabilities which suggest that tourism activity may alter this otherwise quite universal patter of human behavior.

4. Discussion and conclusions

As this study shows, social media systems can be fruitful sources of knowledge to extract urban mobility patterns. Indeed, they offer a global coverage at a low cost, the collected data can be exploited for several mining purposes and it allows to classify all the actors involved in the problem, i.e., the users, the geolocated data points, the geographic areas, and the content that the users share (e.g., the textual part of a tweet).

Even though the number of advantages and possible solutions that can be developed thanks to social media is large, a set of open issues and research challenges still exists. In what follows we present them to highlight possible new frontiers in this research area.

Low frequency of data sharing. The sharing frequency of geolocated data by the users is unequal and is usually low (i.e., users tend to hide the location from which they are posting [14,45]). For instance, in Twitter less than 1% of published posts are geotagged [51]. This poses challenges on the analysis and interpretation of the data [19] and on the ability to collect individual paths [15,31].

To address this problem, different studies, have tried to infer geolocation information exploiting different features of social media data. For instance, some works use the user's social network to infer her geolocation [52,53]. Another type of approaches addresses this problem through text-based geolocation inference techniques [54]. As pointed out in [55], a valuable method is represented by hybrid approaches that exploit both social and text features of social media data. Moreover, incentive mechanisms, such as micro-payments, are possible solutions being investigated for this issue [5].

Data sampling. Social media users are a specific sample of the user population, since the users need to have an Internet connection and some technological expertise. This means that geolocated data collected from social media might be related to a younger population [33]. While it is widely-known that a sampling bias exists for technologies that capture mobile dynamics [31,33], studying how social media's sampling bias affects the mobility patterns of geolocated data is still an open issue [33].

Data collection from third parties. When working with third-party services, the data might not be public [5]. Given the large amount of information that these services collect (e.g., the Waze application), this might limit the urban mobility pattern mining process.

Privacy issues. Even though people are aware of sharing personal information publicly [28], privacy concerns about collecting data without the users' consent exist [16,23,31]. To overcome this issue, approaches such as [22] anonymize user and venues ids. However, aggregating the individual data points to extract information without the users' consent might still violate privacy requirements.

Big data issues. Even though the individual users might share geolocated data with a low frequency, social media systems generate data at a very high rate, leading to the widely-known *big data* problem. This might create challenges for real time storing, processing, and indexing of the data [5], which can have an impact on the mining of up-to-date mobility patterns.

To face this problem, our proposed workflow (composed of four main steps: I. Data collection and preprocessing, II. User profiling and classification, III. Data classification, and IV. Pattern mining) can be easily implemented in a distributed architecture by parallelizing the *user profiling and classification* and *data classification* steps.

The presented study outlined multiple take-home messages. Primarily, when collecting data using the Twitter streaming APIs for a specific area, two main aspects have to be considered: I. the limitations on the amount of data retrieved through Twitter's public streaming APIs, and II. the use of a bounding box might not be precise enough, a subsequent preprocessing based on the employment of shapefiles is recommended. Secondly, in Section 2.2 we have seen that the user classification is a non-trivial task and it may not always be possible to find an obvious clear separation between the classes. A good practice in these cases consist in studying heuristics which allow criteria and parameters that can be different for different places and domains.

The analysis of the state of the art and the conducted case-study have highlighted that social media may be a valuable source of data for city planners, administrators, and urban scientist since they allow to overcome many of the limitations that characterize other mentioned sources of data. At the same time, this relatively new source is generating many challenges which will be the objective of future research.

References

[1] K. Pelechrinis, D. Quercia, Urban informatics and the web, in: Proceedings of the Twenty-Fourth International Conference on World Wide Web, ACM, New York, NY, USA, 2015, p. 1547, doi:10.1145/2740908.2741983.

[2] F. Calabrese, M. Diao, G.D. Lorenzo, J.F. Jr., C. Ratti, Understanding individual mobility patterns from urban sensing data: a mobile phone trace example, Transp. Res. Part C: Emerg. Technol. 26 (2013) 301–313. http://dx.doi.org/10.1016/j.trc.2012.09.009.

[3] L. Song, D. Kotz, R. Jain, X. He, Evaluating location predictors with extensive wi-fi mobility data, SIGMOBILE Mob. Comput. Commun. Rev. 7 (4) (2003) 64–65, doi:10.1145/965732.965747.

[4] A.V. Kurilkin, O.O. Vyatkina, S.A. Mityagin, S.V. Ivanov, Evaluation of urban mobility using surveillance cameras, Proc. Comput. Sci. 66 (2015) 364–371. http://dx.doi.org/10.1016/j.procs.2015.11.042.

[5] T.H. Silva, P.O.S.V. de Melo, J.M. Almeida, A.A.F. Loureiro, Large-scale study of city dynamics and urban social behavior using participatory sensing, IEEE Wirel. Commun. 21 (1) (2014) 42–51, doi:10.1109/MWC.2014.6757896.

[6] M.C. Gonzalez, C.A. Hidalgo, A.-L. Barabasi, Understanding individual human mobility patterns, Nature 453 (7196) (2008) 779–782, doi:10.1038/nature06958.

[7] C. Song, Z. Qu, N. Blumm, A.-L. BarabÃ!'si, Limits of predictability in human mobility, Science 327 (5968) (2010) 1018–1021, doi:10.1126/science.1177170.

[8] L. Pappalardo, F. Simini, S. Rinzivillo, D. Pedreschi, F. Giannotti, A.-L. Barabási, Returners and explorers dichotomy in human mobility, Nat. Commun. 6 (2015) 8166, doi:10.1038/ncomms9166.

[9] M. Berlingerio, F. Calabrese, G. Di Lorenzo, R. Nair, F. Pinelli, M.L. Sbodio, AllAboard: A System for Exploring Urban Mobility and Optimizing Public Transport Using Cellphone Data, Springer, Berlin, Heidelberg, pp. 663–666. 10.1007/978-3-642-40994-3_50

[10] L. Gabrielli, B. Furletti, F. Giannotti, M. Nanni, S. Rinzivillo, Use of mobile phone data to estimate visitors mobility flows, in: Proceedings of MoKMaSD, 2014.

[11] M. De Nadai, J. Staiano, R. Larcher, N. Sebe, D. Quercia, B. Lepri, The death and life of great italian cities: a mobile phone data perspective, in: Proceedings of the Twenty-Fifth International Conference on World Wide Web, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 2016, pp. 413–423, doi:10.1145/2872427.2883084.

[12] D. Arribas-Bel, Accidental, open and everywhere: emerging data sources for the understanding of cities, Appl. Geogr. 49 (2014) 45–53. http://dx.doi.org/10.1016/j.apgeog.2013.09.012. The New Urban World

[13] R. Schwartz, G.R. Halegoua, The spatial self: location-based identity performance on social media, New Media Soc. 17 (10) (2015) 1643–1660, doi:10.1177/1461444814531364.

[14] E. Steiger, J.P. de Albuquerque, A. Zipf, An advanced systematic literature review on spatiotemporal analyses of twitter data, Trans. GIS 19 (6) (2015) 809–834, doi:10.1111/tgis.12132.

[15] H. Gao, H. Liu, Data Analysis on Location-Based Social Networks, Springer, New York, NY, pp. 165–194. 10.1007/978-1-4614-8579-7_8

[16] O. Roick, S. Heuser, Location based social networks – definition, current state of the art and research agenda, Trans. GIS 17 (5) (2013) 763–784, doi:10.1111/tgis.12032.

[17] Y. Zheng, L. Capra, O. Wolfson, H. Yang, Urban computing: concepts, methodologies, and applications, ACM TIST 5 (3) (2014) 38:1–38:55, doi:10.1145/2629592.

[18] S. Zhang, C. Zhang, Q. Yang, Data preparation for data mining, Appl. Artif. Intell. 17 (2003) 375–381.

[19] A. Noulas, S. Scellato, C. Mascolo, M. Pontil, An empirical study of geographic user activity patterns in foursquare, in: L.A. Adamic, R.A. Baeza-Yates, S. Counts (Eds.), Proceedings of the Fifth International Conference on Weblogs and Social Media, The AAAI Press, Barcelona, Catalonia, Spain, 2011.

[20] J. Cranshaw, R. Schwartz, J.I. Hong, N.M. Sadeh, The livehoods project: utilizing social media to understand the dynamics of a city, in: J.G. Breslin, N.B. Ellison, J.G. Shanahan, Z. Tufekci (Eds.), Proceedings of the Sixth International Conference on Weblogs and Social Media, The AAAI Press, Dublin, Ireland, 2012.

[21] T.H. Silva, P.O.S.V. de Melo, J.M. Almeida, J.F.S. Salles, A.A.F. Loureiro, Visualizing the invisible image of cities, in: Proceedings of the IEEE International Conference on Green Computing and Communications, Conference on Internet of Things, and Conference on Cyber, Physical and Social Computing, GreenCom/iThings/CPSCom, IEEE Computer Society, Besancon, France, 2012, pp. 382–389, doi:10.1109/GreenCom.2012.62.

[22] D. Preotiuc-Pietro, T. Cohn, Mining user behaviours: a study of check-in patterns in location based social networks, in: H.C. Davis, H. Halpin, A. Pentland, M. Bernstein, L.A. Adamic (Eds.), Proceedings of the Web Science (co-located with ECRC), ACM, Paris, France, 2013, pp. 306–315, doi:10.1145/2464464.2464479.

[23] T.H. Silva, P.O.S.V. de Melo, J.M. Almeida, J.F.S. Salles, A.A.F. Loureiro, Revealing the city that we cannot see, ACM Trans. Internet Technol. 14 (4) (2014) 26:1–26:23, doi:10.1145/2677208.

[24] Urban activity pattern classification using topic models from online geo-location data, Transp. Res. Part C: Emerg. Technol. 44 (2014) 363–381. http://dx.doi.org/10.1016/j.trc.2014.04.003.

[25] T.H. Silva, P.O.S. Vaz de Melo, J.M. Almeida, J. Salles, A.A.F. Loureiro, A comparison of foursquare and Instagram to the study of city dynamics and urban social behavior, in: Proceedings of the Second ACM SIGKDD International Workshop on Urban Computing, ACM, New York, NY, USA, 2013, pp. 4:1–4:8, doi:10.1145/2505821.2505836.

[26] Z. Cheng, J. Caverlee, K. Lee, D.Z. Sui, Exploring millions of footprints in location sharing services, in: L.A. Adamic, R.A. Baeza-Yates, S. Counts (Eds.), Proceedings of the Fifth International Conference on Weblogs and Social Media, The AAAI Press, Barcelona, Catalonia, Spain, 2011.

[27] G. Fuchs, G. Andrienko, N. Andrienko, P. Jankowski, Extracting personal behavioral patterns from geo-referenced tweets, in: Proceedings of the Sixteenth AGILE Conference on Geographic Information Science, 2013.

[28] T. Fujisaka, R. Lee, K. Sumiya, Exploring urban characteristics using movement history of mass mobile microbloggers, in: A. Dalton, R. Want (Eds.), Proceedings of the Eleventh Workshop on Mobile Computing Systems and Applications, HotMobile '10, ACM, Annapolis, Maryland, USA, 2010, pp. 13–18, doi:10.1145/1734583.1734588.

[29] L. Ferrari, A. Rosi, M. Mamei, F. Zambonelli, Extracting urban patterns from location-based social networks, in: Proceedings of the International Workshop on Location Based Social Networks, Chicago, IL, USA, 2011, pp. 9–16, doi:10.1145/2063212.2063226.

[30] T. Shelton, A. Poorthuis, M. Zook, Social media and the city: rethinking urban socio-spatial inequality using user-generated geographic information, Landsc. Urb. Plan. 142 (2015) 198–211. http://dx.doi.org/10.1016/j.landurbplan.2015.02.020. Special Issue: Critical Approaches to Landscape Visualization

[31] F. Girardin, F. Dal Fiore, J. Blat, C. Ratti, Understanding of tourist dynamics from explicitly disclosed location information, in: Proceedings of the Symposium on LBS and Telecartography, 58, 2007.

[32] B. Hawelka, I. Sitko, E. Beinat, S. Sobolevsky, P. Kazakopoulos, C. Ratti, Geo-located twitter as the proxy for global mobility patterns, Cartogr Geogr Inf Sci 41 (3) (2014) 260–271.

[33] R. Jurdak, K. Zhao, J. Liu, M. AbouJaoude, M. Cameron, D. Newth, Understanding human mobility from twitter, PLoS ONE 10 (7) (2015) e0131469.

[34] W.F. Theobald, Global Tourism, third ed, Elsevier Butterworth-Heinemann, Maryland Heights, MO, 2005. "Copyright c2005"

[35] L. Jin, X. Long, K. Zhang, Y. Lin, J.B.D. Joshi, Characterizing users' check-in activities using their scores in a location-based social network, Multimed. Syst. 22 (1) (2016) 87–98, doi:10.1007/s00530-014-0395-8.

[36] G. Cao, S. Wang, M. Hwang, A. Padmanabhan, Z. Zhang, K. Soltani, A scalable framework for spatiotemporal analysis of location-based social media data, Comput. Environ. Urb. Syst. 51 (2015) 70–82, doi:10.1016/j.compenvurbsys.2015.01.002.

[37] G.L. Andrienko, N.V. Andrienko, H. Bosch, T. Ertl, G. Fuchs, P. Jankowski, D. Thom, Thematic patterns in georeferenced tweets through space–time visual analytics, Comput. Sci. Eng. 15 (3) (2013) 72–82, doi:10.1109/MCSE.2013.70.

[38] V. Frías-Martínez, V. Soto, H. Hohwald, E. Frías-Martínez, Characterizing urban landscapes using geolocated tweets, in: Proceedings of the International Conference on Privacy, Security, Risk and Trust, PASSAT and International Confernece on Social Computing, SocialCom, IEEE, Amsterdam, Netherlands, 2012, pp. 239–248, doi:10.1109/SocialCom-PASSAT.2012.19.

[39] R. Lee, K. Sumiya, Measuring geographical regularities of crowd behaviors for twitter-based geo-social event detection, in: X. Zhou, W. Lee, W. Peng, X. Xie (Eds.), Proceedings of the International Workshop on Location Based Social Networks, ACM, San Jose, CA, USA, 2010, pp. 1–10, doi:10.1145/1867699.1867701.

[40] S. Hasan, X. Zhan, S.V. Ukkusuri, Understanding urban human activity and mobility patterns using large-scale location-based data from online social media, in: Proceedings of the Second ACM SIGKDD International Workshop on Urban Computing, Chicago, Illinois, USA, 2013, pp. 6:1–6:8, doi:10.1145/2505821.2505823.

[41] M.R. Frank, L. Mitchell, P.S. Dodds, C.M. Danforth, Happiness and the patterns of life: a study of geolocated tweets, Sci Rep 3 (2013) 2625.

[42] E. Cho, S.A. Myers, J. Leskovec, Friendship and mobility: user movement in location-based social networks, in: C. Apté, J. Ghosh, P. Smyth (Eds.), Proceedings of the Seventeenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, San Diego, CA, USA, 2011, pp. 1082–1090, doi:10.1145/2020408.2020579.
[43] D. Kahle, H. Wickham, ggmap: spatial visualization with ggplot2, R J. 5 (1) (2013) 144–161.
[44] R.C. Browning, E.A. Baker, J.A. Herron, R. Kram, Effects of obesity and sex on the energetic cost and preferred speed of walking, J. Appl. Physiol. 100 (2) (2006) 390–398, doi:10.1152/japplphysiol.00767.2005.
[45] Q. Huang, D.W.S. Wong, Modeling and visualizing regular human mobility patterns with uncertainty: an example using twitter data, Ann. Assoc. Am. Geogr. 105 (6) (2015) 1179–1197, doi:10.1080/00045608.2015.1081120.
[46] J.B. MacQueen, Some methods for classification and analysis of multivariate observations, in: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1, University of California Press, 1967, pp. 281–297.
[47] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[48] X. Long, L. Jin, J. Joshi, Exploring trajectory-driven local geographic topics in foursquare, in: A.K. Dey, H. Chu, G.R. Hayes (Eds.), Proceedings of the ACM Conference on Ubiquitous Computing, ACM, Pittsburgh, PA, USA, 2012, pp. 927–934, doi:10.1145/2370216.2370423.
[49] Y. Liu, Z. Sui, C. Kang, Y. Gao, Uncovering patterns of inter-urban trips and spatial interactions from check-in data, PLoS ONE 9 (1) (2014) e86026.
[50] Q. Yuan, G. Cong, K. Zhao, Z. Ma, A. Sun, Who, where, when, and what: a non-parametric Bayesian approach to context-aware recommendation and search for twitter users, ACM Trans. Inf. Syst. 33 (1) (2015) 2:1–2:33, doi:10.1145/2699667.
[51] J. Mahmud, J. Nichols, C. Drews, Home location identification of twitter users, ACM Trans. Intell. Syst. Technol. 5 (3) (2014) 47:1–47:21, doi:10.1145/2528548.
[52] L. Backstrom, E. Sun, C. Marlow, Find me if you can: improving geographical prediction with social and spatial proximity, in: Proceedings of the Ninteenth International Conference on World Wide Web, ACM, New York, NY, USA, 2010, pp. 61–70, doi:10.1145/1772690.1772698.
[53] L. Kong, Z. Liu, Y. Huang, Spot: locating social media users based on social network context, Proc. VLDB Endow. 7 (13) (2014) 1681–1684, doi:10.14778/2733004.2733060.
[54] B. Han, P. Cook, T. Baldwin, Text-based twitter user geolocation prediction, J. Artif. Int. Res. 49 (1) (2014) 451–500.
[55] D. Jurgens, T. Finethy, J. McCorriston, Y.T. Xu, D. Ruths, Geolocation prediction in twitter using social networks: a critical analysis and review of current practice, in: M. Cha, C. Mascolo, C. Sandvig (Eds.), Proceedings of the Ninth International Conference on Web and Social Media, University of Oxford, Oxford, UK, 2015, pp. 188–197.

Matteo Manca is a Data Scientist in the Digital Humanities group at Eurecat,which he joined in September 2015; his main research activities are related to social media mining and computational social science. Currently Matteo's efforts are focused on applying computational methods to digital trace data in order to gain knowledge about human behaviour and social phenomena. In 2014, he earned his PhD in computer science from the University of Cagliari with a thesis focused on the study and implementation of social recommendation approaches for the social media domain. In 2009, Matteo got his master degree in computer science from the same university. Before joining Eurecat, he had worked in a rare diseases center as computer scientist and he gained experiences in different software companies. Moreover, he has worked as intern in CureMetrix (working remotely), San Diego (CA) - USA and in the Web mining group at Yahoo! Researcher Barcelona.

Ludovico Boratto is a research scientist in the Digital Humanities Department at Eurecat. His research interests focus on Data Mining and Machine Learning approaches, mostly applied to recommender systems and social network analysis. The results of his research have been published in top-tier journals, such as Information Sciences (Elsevier) and IEEE Intelligent Systems. His research activity also brought him to give talks and tutorials at top-tier conferences (RecSys 2016) and research centers (Yahoo! Research). He is currently guest editor of three journal's special issues. In 2012 he got a Ph.D. at the University of Cagliari (2012), where he was a research assistant until May 2016. In 2010 and 2014 he spent 10 months at Yahoo! Research in Barcelona as a visiting researcher.

Víctor Morell Román is architect at Urbaning and graduated from the Polytechnic University of Catalonia in the Superior Technical School of Architecture of the Vallés. He completed his studies in architecture with a Master in real estate valuations at the Polytechnic University of Catalonia. He has extensive experience collaborating in offices of architecture and urbanism. He is currently pursuing a degree in computer science.

Oriol Martori i Gallissà is Managing Director at Urbaning and a landscape architect and urban planning technician, with more than 12 years of experience developing projects of urban strategy, planning and management. His training and experience has been developed mainly between London and Barcelona. He has obtained a MA in Urban and Regional Planning form the School of Public Administration of Catalunya in 2010, a MA in Urban Policy from the Universitat Politècnica de Catalunya in 2010 and a MA in Urban Design from the University of Westminster in 2006. Oriol has recently participated in different studies to evaluate the role of cities from the point of view of pedestrians and has developed different education projects to bring urban studies closer to schools.

Andreas Kaltenbrunner is the head of the Digital Humanities Department at Eurecat. His research is centered on social media and social network analysis. Andreas uses methods from computer science and the study of complex systems to resolve sociological research questions. He obtained his PhD in Computer Science and Digital Communication in 2008 from the Universitat Pompeu Fabra with a thesis about stochastic effects in human and neural communication patterns. Afterwards he joined the Information Retrieval Group of the technology center Barcelona Media. There he analyzed structural, temporal and behavioral patterns in human interactions in social media websites. He later co-founded in October 2010 as senior researcher the Social Media research line within Barcelona Media and led it from May 2013 onwards. Since June 2015 he leads the Digital Humanities Research Unit at the newly formed technology center Eurecat.