

# Using Open Data to Analyze Urban Mobility from Social Networks

Caio Libânio Melo Jerônimo, Claudio E. C. Campelo, Cláudio de Souza Baptista

Federal University of Campina Grande, Brazil

caiolibanio@copin.ufcg.edu.br, {campelo, baptista}@computacao.ufcg.edu.br

**Abstract.** The need to use online technologies that favor the understanding of city dynamics has grown, mainly due to the ease in obtaining the necessary data, which, in most cases, are gathered with no cost from social networks services. With such facility, the acquisition of georeferenced data has become easier, favoring the interest and feasibility in studying human mobility patterns, bringing new challenges for knowledge discovery in GIScience. This favorable scenario also encourages governments to make their data available for public access, increasing the possibilities for data scientist to analyze such data. This article presents an approach to extracting mobility metrics from Twitter messages and to analyzing their correlation with social, economic and demographic open data. The proposed model was evaluated using a dataset of georeferenced Twitter messages and a set of social indicators, both related to Greater London. The results revealed that social indicators related to employment conditions present higher correlation with the mobility metrics than any other social indicators investigated, suggesting that these social variables may be more relevant for studying mobility behaviors.

Categories and Subject Descriptors: H.2.8 [Database Applications]: Data mining; H.2.8 [Database Applications]: Spatial databases and GIS; H.3.3 [Information Search and Retrieval]: Clustering

Keywords: mobility patterns, open Government data, statistical correlations analysis

## 1. INTRODUCTION

With the growth of large cities and the increasing demand for better public services, the need for a clear understanding of city dynamics becomes mandatory. Understanding the citizens behavior is necessary for the development of inclusive policies and improvements in mobility strategies made by governments.

Urban Mobility Patterns represent models of human behavior in an urban environment [Luo et al. 2016] and are especially relevant as the analysis of these patterns may influence public transportation systems, public safety, traffic engineering, health systems, and many other fields related to the planning of urban centers [Noulas et al. 2012; Wilson and Bell 2004]. Mobility patterns are also present in studies related to recommendation systems [Hao et al. 2010; Zheng et al. 2010] as well as in research on trajectories [Bagrow and Lin 2012; Hsieh et al. 2012]. There are many studies addressing this subject, however, most of them are concerned with data from cell phone networks [Gonzalez et al. 2008; Jiang et al. 2013; Palchykov et al. 2014], Wifi networks [Chaintreau et al. 2007; Zhang et al. 2012] or GPS signals [Rhee et al. 2011; Zhao et al. 2014]. Although these studies help understand mobility patterns, they have restrictions on privacy, as well as on precision, specially when using cell phone networks data.

The massive usage of social networks by different classes of people have lead to an increase in the amount of data that are generated in the internet, enabling the rise of new studies related to knowledge

---

Copyright©2017 Permission to copy without fee all or part of the material printed in JIDM is granted provided that the copies are not made or distributed for commercial advantage, and that notice is given that copying is by permission of the Sociedade Brasileira de Computação.

discovery. This phenomenon, allied to the usage of smartphones in daily life, and the capability of these devices to generate georeferenced data, have also favored studies related to mobility patterns, specially in urban centers. It is known that large cities might have considerable economic, social and demographic discrepancies among their regions, which can influence the way locals move within such urban centers. Analyzing how these factors influence urban mobility is a major challenge to be considered, for example, in Points of Interest (POI) recommendation, in route prediction, or in urban planning systems.

In this article we propose a model to extract mobility metrics from georeferenced data collected from social networks and to find statistical correlations between these metrics and social, economic and demographic open data from an urban center. Georeferenced information obtained from social media is typically imprecise and fragmented. For example, many people only post messages from certain locations (e.g from home, from work), even though they visit many other places; many users remain inactive for long periods. Thus, this research aims at verifying the feasibility of extracting mobility information from social media data that is relevant enough to be correlated with other open social data.

To evaluate our model, we collected Twitter messages from Greater London for one year, in a total of 19,456,798 messages, aiming the extraction of mobility metrics from these messages. For the social data, we used the public platform London Datastore<sup>1</sup>. These two dataset were supplied to our model, allowing the discovery of some correlations between these data, indicating the feasibility of using a model of this kind. Through the experiments we conducted, we found that the social variables related to employment conditions tend to correlate with the mobility metrics analyzed in this study, specially the following variables: economically inactive people, employment rate, unemployment rate and persons with no qualifications.

The obtained results suggest that it is possible to extract relevant knowledge from data by combining techniques already applied in mobility studies with statistical correlation tests, favoring studies aiming to analyze data produced by humans or any kind of sensors in urban environments. The output provided by the model can be used, for example, by government decision makers to understand the dynamics of populations and to better plan the locations of public services. Such an output can also serve as an input to recommendation systems, improving their recommendations of places and services according to people's social behaviors.

This article is an extension of Jerônimo et al. [2016], published in *Geoinfo* 2016. The main contributions of this article, in comparison to the previous one, are: (1) a more detailed discussion of related work, including different researches and a table contrasting their main characteristics; (2) more detailed explanations of the methods of home and activity center detection, with the description of the algorithms used; (3) improvements in the results visualization by adding heat maps, showing the density of posts along the city of London; (4) validation of the home detection method, with the collaboration of volunteers; (5) an improved explanation of the mobility properties; (6) a deeper analysis of the data, showing them by temporal ranges, improving the number of correlations found.

The remainder of this article is organized as follows. The next section presents related work. Section 3 describes the mobility metrics used and some relevant related concepts. Then the experiments conducted to validate our model are addressed in Section 4. Section 5 discusses obtained results. Section 6 concludes the article and points to future directions of this research work.

## 2. RELATED WORK

Most studies related to mobility behaviors use data derived from cell phone networks, RFID devices, GPS based data, or Wifi networks. Recently, studies have addressed the task of extracting and

<sup>1</sup>London DataStore: <http://data.london.gov.uk/>

identifying mobility behaviors from social media data. This tendency is a consequence of the way that these networks offer their data, since this information is mostly available for public access, reducing financial costs applied to research projects.

Yuan et al. [2013] proposed a probabilistic model called W4 (Who + Where + When + What) to extract from Twitter messages aspects of mobility related to the users of this social network. The authors considered the spatial and temporal dimensions, and also the activities performed by the users.

Considering social networks as new data sources for current and future research in many different fields, some researchers have analyzed the suitability of this kind of data source. Jurdak et al. [2015] analyze mobility patterns considering spatial and temporal aspects related to the major cities of Australia, in order to demonstrate that Twitter data can be as efficient as the data provided by mobile phone networks. Similarly, the research presented by Hawelka et al. [2014] only considers spatial and temporal aspects of Twitter data, however, they deal with global mobility scales (between countries). The study aims at revealing global mobility patterns related to these messages, demonstrating that these data have similar properties to other kinds of data sources used in different studies.

Hasan et al. [2013] categorize mobility patterns through user's activities around three major cities: New York, Chicago and Los Angeles. The authors consider, in addition to spatial and temporal aspects, the semantics of displacements. To do this, they analyze georeferenced messages from Twitter, using links to the Foursquare platform, which allows them to identify and categorize the check-ins as: (1) at home; (2) work; (3) meal; (4) entertainment activity; (5) recreation and (6) shopping. Yin et al. [2015] propose a probabilistic model called Topic-Region-Model (TRM) to identify semantic, temporal and spatial patterns from check-in data of both Foursquare and Twitter networks, allowing the authors to study the decision of these users by certain Points of Interest (POIs) in situations where a user is not in his/her city of origin, making it difficult to perform recommendations.

Blanford et al. [2015] investigate spatiotemporal characteristics in mobility patterns between Kenya's political boundaries using georeferenced data from Twitter messages. For the temporal analysis, days and months of user messages are considered. For the spatial analysis, the authors consider the displacement between the main cities of the country, using a proprietary GIS tool to view these data.

Nguyen and Szymanski [2012] used data collected from Gowalla<sup>2</sup> to create and validate human mobility models taking into account the user's friends circles and how it would affect these patterns. One of the main findings is that the incidence of check-ins in the same location is not a good indicator of friendship between users, at least in this social network.

Most research relating mobility patterns and social network data only focus on the spatial and temporal aspects of these patterns. However, other aspects might be considered when studying mobility patterns, specially economic, social and demographic factors. In this context, the works that consider these dimensions are restricted and limited to a few variables in a social context.

Steiger et al. [2015] explore the semantics of messages posted in Twitter from the region of London, inferring the message location (home or work), allowing the identification of residential and commercial regions in the city. In order to validate the results, the authors correlate the findings with census data, showing that their model inferred work regions better than home regions.

Cheng et al. [2011] investigate georeferenced Twitter messages, considering, in addition to spatial and temporal aspects, variables related to income, popularity on the social network, and the content of the messages. The authors try to find out some relations between these variables and the mobility patterns encountered in the Twitter messages. They conclude that people who live in cities with a higher average income, tend to get around for longer distances. Luo et al. [2016] investigate, in addition to spatial and temporal aspects, the ethnicity, age and gender of the Twitter users. These

---

<sup>2</sup>This network was acquired by Facebook and ended its activities in 2012.

Table I. Comparative table of related works

Work	Considers social aspects	Use of Government open data	Considers spatial and temporal aspects	Social circle	Use of message content	Use of multiple social networks	Use of POI	Use of home places
Luo et al. [2016]	x	x	x					x
Steiger et al. [2015]	x	x	x		x			x
Jurdak et al. [2015]			x					
Yin et al. [2015]			x		x	x	x	
Blanford et al. [2015]			x					
Hawelka et al. [2014]			x					
Yuan et al. [2013]			x		x			
Hasan et al. [2013]			x		x	x		
Nguyen and Szymanski [2012]			x	x				
Cheng et al. [2011]	x	x	x		x			x
<b>Our proposal</b>	<b>x</b>	<b>x</b>	<b>x</b>				<b>x</b>	<b>x</b>

social variables have been inferred from the users' profiles on Twitter and from public information provided by the government. The authors analyze how these three variables influence the mobility patterns extracted from the messages of Twitter related to the city of Chicago. They conclude that ethnicity was the most determining factor with regard to mobility patterns, possibly because this variable may express some socioeconomic characteristics of these users, demonstrating some level of segregation imposed to foreign people.

Table I shows a comparison of the related works classified using relevant characteristics present in the literature. The last line of Table I shows the characteristics of the proposed model.

Considering the related works present in Table I, it is possible to observe that there are few studies addressing social aspects and how these aspects may impact in mobility behavior of urban populations. This lack of studies considering social variables and mobility behavior represents a relevant gap in the

literature, increasing the need for models that are able to handle these two dimensions of data, which could help in understanding the cities dynamics, and how people behave in these cities.

### 3. MOBILITY METRICS AND SOCIAL ANALYSIS

Figure 1 shows an overview of our approach to detecting correlations between mobility metrics and social, economic and demographic data.

Our model accepts as input a set of Twitter messages in JSON format. The first step consists in filtering the Twitter messages. Initially, the model filters out messages without geographic coordinates and those whose latitude/longitude coordinates do not point to a location inside the area of interest. Filtering out this kind of data is important because most messages posted in social networks only have an approximation of location provided by the IP address, thus, missing the real coordinates of the user. In addition, most social networks API allow queries for a specific region (e.g., a city), but they do not respect official borders of the regions, resulting in messages out of these real limits. Then, still in this filtering process, messages posted by stationary users are also filtered out. We consider as stationary those users whose messages are located within the same area, based on a predefined radius. Filtering this kind of users is important due to the fact that these users generally represent companies that report community information such as weather or traffic conditions, which are not relevant for the proposed model. Finally, similarly to [Birkin et al. 2014], we remove all users with less than 20 messages, avoiding noise in the data provided by users with few posts.

In the second and third steps, the model detects the home and the activity centers (most visited regions) for each user. These concepts are used by the model to identify social characteristics and recurrent behaviors of the users. The fourth step extracts mobility metrics that are frequently used in mobility patterns analysis, that are: radius of gyration and user displacement distance [Luo et al. 2016; Cheng et al. 2011; Gonzalez et al. 2008; Hasan et al. 2013]. In this process, the collected messages are analyzed and the two properties are extracted for each user, allowing the identification of mobility characteristics present in data, favoring the calculation of the correlations between these properties and the social data. These statistical properties will be explained in further sessions.

The fifth step consists in the correlation analysis, where the model receives the mobility metrics extracted by the previous process, and accepts tabular data containing the social variables and the polygon related to each region of the city/region to be analyzed, allowing the model to calculate the correlations between mobility and social data. After all correlations are calculated, the model generates a correlation matrix. This matrix contains the correlations found between the mobility metrics and the social variables. The model returns this matrix as an output file, allowing an easy visualization of the results. Samples of this matrix can be found in Table II, Table III, Table IV and Table V.

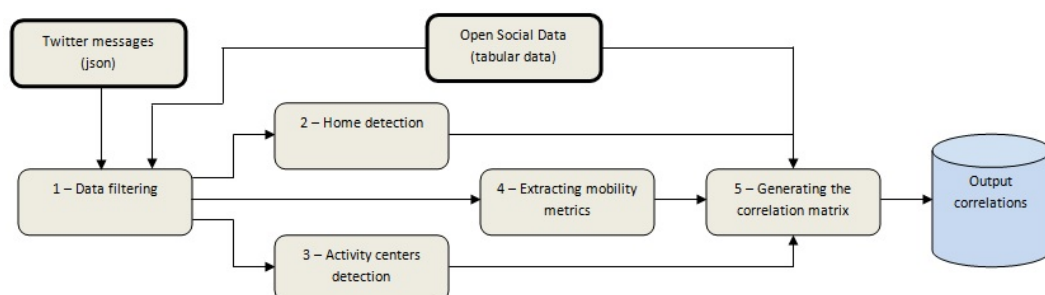


Fig. 1. Basic flow process of the proposed model

### 3.1 Radius of gyration

As a popular metric of mobility, the radius of gyration represents the standard deviation of distances between points of a trajectory and the center of mass of these points. This metric was chosen due to its large adoption in the literature, and the ability that this value has to express relevant patterns of general mobility. For example, this metric can measure how far and how frequently a user moves. A low radius of gyration indicates that a specific user tends to travel mostly locally, with few long-distance check-ins, while a high value of this metric generally indicates that the user moves predominantly for long distances in relation to its center of mass [Cheng et al. 2011]. This metric can be formalized in Equation 1 as:

$$r = \sqrt{\frac{1}{m} \cdot \sum_{i=1}^m (p_i - p_c)^2} \quad (1)$$

- $r$  represents the value for the radius of gyration for a user;
- $m$  is the number of messages for a user;
- $p_i$  represents a particular point where the message was posted;
- $p_c$  represents the user center of mass (centroid);
- $(p_i - p_c)$  is the distance between a particular point from the user's centroid;

For the extraction of this metric, the model calculates the geographical centroid of the messages posted by the user and then, the distances between this centroid and the user's messages. After these calculations, the obtained distances are used in the above equation to extract the radius of gyration of each user in the dataset.

### 3.2 User displacement distance

The second metric adopted in this research, and also widely used in the literature as an indicator of mobility behavior, is the user displacement distance. This metric of mobility represents the sum of the distances between all consecutive messages or check-ins, reflecting the total distance traveled by the user around the analyzed geographic region.

Cheng et al. [2011] show that this metric approximates to a power law, considering the frequency of displacements and the displacements in miles, where  $\delta_r$  represents the displacement and  $-\beta = 1.8845$ , suggesting that the behavior of user's displacement for Twitter messages follows a Lévy Flight distribution [Rhee et al. 2011]. This distribution is characterized as a mixture of short and random displacements, with occasional long jumps.

$$P(\delta_r) \propto \delta_r^{-\beta} \quad (2)$$

Shin et al. [2008] find similar results for the displacement distribution by analyzing GPS data in different scenarios, such as metropolitan area and college campuses. In opposition to prevailing Lévy flight random walk models, Gonzalez et al. [2008], by analyzing data from cell phone calls, highlight that human displacements have a significant level of temporal and spatial regularity, mainly because they tend to return to a few highly frequented locations.

For this metric, the model calculates the distance between all consecutive messages of each user and then, these distances are summed in order to obtain the user displacement distance.

### 3.3 Activity centers

An activity center can be defined as a location that a user frequently visits. The locations for an activity center could be a restaurant, home, place of work or any location where the user post his/her messages with some frequency. This concept represents an important parameter to express the life patterns of a user, indicating the user's preferences for certain places or regions in a city.

To identify such activity centers, we used a popular clustering algorithm called DBSCAN (Density Based Spatial Clustering of Applications with Noise) [Ester et al. 1996]. This is a density based clustering algorithm that clusters points that are closely located, and it can deal with points that are located in a low density regions, treating these points as noise. The algorithm has two parameters: the maximum radius of the neighborhood to be considered to form a cluster ( $\epsilon$ ); and the minimum number of points that a cluster must have (*minPts*). This algorithm presents advantages over other clustering algorithms, such as: it does not require the user to specify the number of clusters for the execution; it can deal with noisy data efficiently; it can find a cluster surrounded by another cluster; and it relies on two variables only.

### 3.4 Home detection

The region of a user's home represents an important characteristic that needs to be considered in a study that analyzes social, economic and demographic data. This is mainly because the user's home location may express social conditions, and these conditions might, in part, influence how the citizens move across an urban center.

Detecting the home location based on the center of mass of all user check-ins can lead to problems of splitting-the-difference, where a user that travels to distant regions over the city will have her home located at the middle of these regions [Cheng et al. 2011].

In this study, we consider home locations as the most intense activity centers during the night time, following other existing works [Luo et al. 2016; Huang et al. 2014]. For this, we select the messages posted between 8pm and 6am (on weekdays only); then we apply the DBSCAN algorithm to cluster the points of these messages; and finally we select the cluster with the greatest number of points as the user home.

## 4. EXPERIMENTAL EVALUATION

This section presents the experimental evaluation conducted to validate our proposed model for identifying possible correlations between mobility metrics extracted from Twitter messages and social open data provided by governmental organizations.

### 4.1 Social data and maps

To perform the experiments, the city of London was chosen as a case study, specially due to the large volume of social, economic and demographic data publicly available for this city. In this study, data were collected from the governmental platform London Datastore<sup>3</sup>.

The social data collected for these experiments is related to the year of 2011, since this was the most recent data for the majority of the collected variables. Observed variables are related to the following categories: population age, family structure, ethnic groups, country of birth, house prices, economic activity, qualifications, health, car or van availability and religion. Additionally, we used the data level of LSOA (Lower Super Output Area), which is the smallest area used to divide the city of

<sup>3</sup>London DataStore: <http://data.london.gov.uk/>



Fig. 2. City of London divided by LSOA regions

London for the available data. Each LSOA has an average population of 1,722 inhabitants. Figure 2<sup>4</sup> shows the map of the city of London used in this research. The map is divided by LSOA regions, and is originally made by the ONS (Office for National Statistics). This is under the terms of the Open Government License (OGL)<sup>5</sup> and UK Government Licensing Framework and is therefore free for this kind of use. This map was also acquired from the London Datastore platform.

#### 4.2 Twitter dataset

For this set of experiments, we collected georeferenced Twitter messages from the dataset of [de Oliveira 2017]. The messages were collected from November 26th 2014 to November 22nd 2015, totalizing 19,456,798 messages from the region of London. From the initial set of messages, our model filtered out the messages without geographic coordinates information, resulting into 7,680,200 georeferenced messages and a total of 351,656 who have posted these messages. Then, after removing messages whose locations are outside the region of interest, messages posted from stationary users, and messages posted by light users, the dataset was further reduced to 6,203,474 georeferenced messages and 52,974 users, with an average of 117.1 messages per user.

#### 4.3 Design of Experiments

The experiments executed in this research had the prior objective of answering the following questions:

- Research question (Q1): Do the characteristics of a user's home region have an influence on his/her mobility behaviors?
- Research question (Q2): Do the characteristics of the regions of a user's activity centers have an influence on his/her mobility behaviors?

To answer these questions, the proposed model performs Kendall's correlation test to calculate possible correlations between user's mobility metrics and open social data. The Kendall's correlation

<sup>4</sup>England map adapted from OpenStreetMap: [http://wiki.openstreetmap.org/wiki/Greater\\_London](http://wiki.openstreetmap.org/wiki/Greater_London)

<sup>5</sup>Contains National Statistics data ©Crown copyright and database right [2012] and Contains Ordnance Survey data ©Crown copyright and database right [2012].



coefficient range from -1 to 1, giving the strength/direction of the relationship as the value approximate to these extremes. A correlation coefficient of 0 indicates no correlation. We chose this test due to the fact that it neither requires a specific distribution of data nor a linear relation among the variables within the dataset [Croux and Dehon 2010; Chok 2010]. The Kendall's correlation is based on the ranking of concordant pairs of variables, measuring the strength of the relation between them. The Equation 2 shows the formalization of the Kendall tau-b coefficient<sup>7</sup>, which is better adjusted for the occurrence of ties in data [Knight 1966] (e.g. pairs with repeated values), and thus, is adopted in this research.

$$\tau_b = \frac{n_c - n_d}{\sqrt{((n_0 - n_1) * (n_0 - n_2))}} \quad (3)$$

— $n_0 = n * (n - 1) / 2$ ;

— $n_c$  is the number of concordant pairs;

— $n_d$  is the number of discordant pairs;

— $n_1 = \sum_i t_i(t_i - 1) / 2$ ;

— $n_2 = \sum_j u_j(u_j - 1) / 2$ ;

— $t_i$  is the number of tied values in the  $i^{th}$  group of ties in  $x$ ;

— $u_j$  is the number of tied values in the  $j^{th}$  group of ties in  $y$ ;

We divided the experimental evaluation into two experiments. Experiment 1 consisted in analyzing the first two metrics (radius of gyration and displacement distance). For this, we performed the correlation test between these metrics and all social data variables related to users' home locations (previously detected), selecting the correlations with  $\tau_b \geq 0.25$  and with  $p\text{-value} < 0.05$ . The  $p\text{-value}$  is used to determine statistical significance in a hypothesis test [Hung et al. 1997]. The formalization of this value would be quite long for this article, but it is largely described in the statistical literature. For the calculation of this statistic, we use the R<sup>6</sup> API. This approach enables us to identify situations where the social, economic and demographic status of the region where a user's home is located may have an influence on his/her mobility behaviors, allowing us to answer Question Q1. For the calculation of the correlations, we use the Apache Commons Mathematics Library<sup>7</sup>.

The Code 1 shows the method of home detection. The algorithm receives a list with all users as a parameter and calculates the estimated location of their residences. In line 2, the algorithm iterates over all users. In line 3, it retrieves the messages posted during weekdays and between 8pm and 6am only. In line 4, we cluster the dataset of points for each user using the DBSCAN algorithm with  $\epsilon = 45$  meters and  $minPts = 4$ . The result of this execution is a list of lists, where each line of this list represents a cluster detected by the DBSCAN. In line 5, the algorithm returns the biggest cluster detected by the DBSCAN. After this process, in line 6, the algorithm calculates the centroid of the biggest cluster found in previous line. This centroid is considered the home location of the user, and this point is assigned to the user in line 7.

In order to validate the home detection method, we selected five volunteers to analyze satellite images of randomly selected points considered as home places by the model. Each volunteer was assigned to analyze ten images and classify them as residential places or non-residential places. It was considered as residential places the centroids that were located near buildings that could be easily identified as a house or residential apartment. Places that did not match the characteristics of a home place, in a residential neighborhood, were classified as non-residential places. With the proposed evaluation, the volunteers found a total of 31 residential places, totalizing 62% of the sample. Figure 3 shows an example of a place considered as a home location.

<sup>6</sup><https://www.r-project.org/>

<sup>7</sup><https://commons.apache.org/proper/commons-math>

**Code 1 - Detecting home locations for each user**

```

1: function detectHomeLocatoin(listUsers)
2:   for user in listUsers
3:     listPoints = user.getAllMessagesAsPointsInHomeTime();
4:     listOfClusters = executeDBSCAN( $\epsilon$ , minPts, listPoints);
5:     biggestCluster = listOfClusters.getBiggestCluster();
6:     centroidPoint = biggestCluster.calculateCentroid();
7:     user.setHomePoint(centroidPoint);

```

**Code 2 - Detecting activity centers**

```

1: function calculateMedianFromAC(listUsers)
2:   for user in listUsers
3:     listPoints = user.getAllMessagesAsPoints();
4:     listOfClusters = executeDBSCAN( $\epsilon$ , minPts, listPoints);
5:     user.addActivityCenters(listOfClusters);

```

Experiment 2 has been conducted aiming at answering Question Q2. For this, we used the concept of activity centers described in Section 3.3. First, we clustered the dataset of points for each user with the DBSCAN algorithm in order to find the user's activity centers. After this clustering process, we calculate the medians for all social data variables related to the regions in which these clusters were formed, for each user. Then we performed the correlation tests over these medians and the users' mobility metrics, allowing us to answer Question Q2. The Code 2 shows the method of activity centers detection in detail.

In line 2 of Code 2, we iterate over all users present in dataset. In line 3, the coordinates of all messages owned by the user are retrieved. In line 4, we cluster the dataset of points for each user with the DBSCAN algorithm using  $\epsilon = 45$  meters and  $minPts = 4$  in order to find the user's activity centers. The result of this execution is a list of lists, where each line of this list represents a cluster detected by the DBSCAN. In line 5, the clusters are assigned to each user. For Experiments 1 and 2, we adopted the value of 45 meters for filtering stationary users.

Code 3 shows the process of generating the correlation matrix. In line 1, the method receives a matrix of the mobility metrics extracted from users in previous step. In this matrix, each column represents one mobility metric extracted, and the lines represent each user. In lines 4 and 6, the

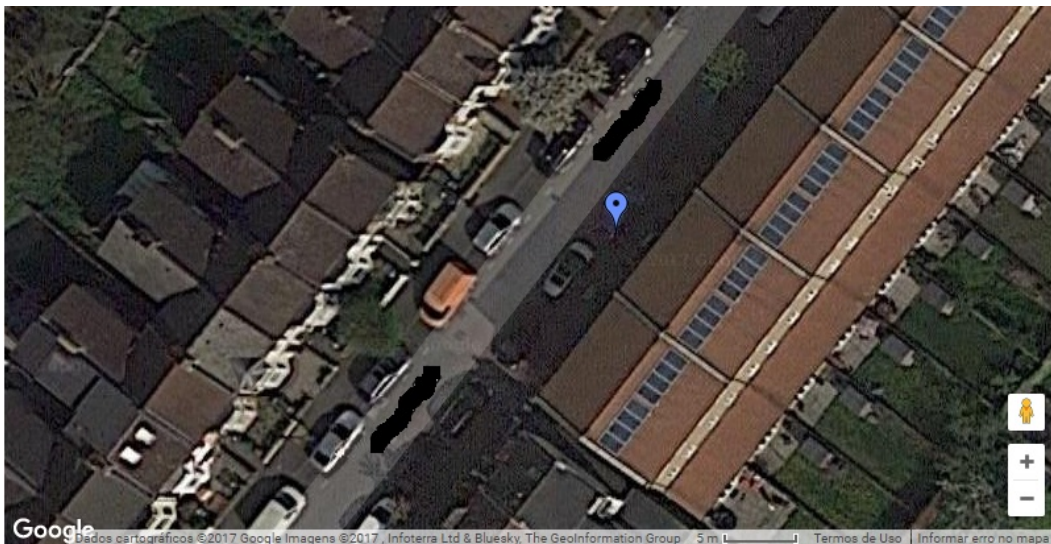


Fig. 3. Example of a centroid located in a residential zone, thus considered as a home location by the volunteer

**Code 3 - Generating the correlation matrix**


---

```

1: function generatingCorrelationMatrix(mobilityMatrix)
2:   for user in listUsers
3:     if experiment_1
4:       userSocialArray = findSocialIndicatorsByRegion(user.homePoint())
5:     else #experiment 2
6:       userSocialArray = findMedianOfSocialIndicatorsByRegion(user.clusteringPoints())
7:     socialDataMatrix.add(userSocialArray)
8:   mobilityAndSocialMatrix = concatenateMatrix(mobilityMatrix , socialDataMatrix)
9:   correlationMatrix = kendallCorrelationCalculator(mobilityAndSocialMatrix)

```

---

method *findSocialIndicatorsByRegion(point)* recovers from a database the social indicators for each region (home or AC) of each user. For the home location (Experiment 1), an array containing the social indicators for the home region of the user is returned. For the AC (Experiment 2), an array containing the medians of each social indicator is returned. In opposite to the home location, where a user has only one home location, a user can have multiple AC (multiple points), thus the median is used to represent the values of each social indicator. In both arrays, each position of the array contains the value of a social indicator. In line 7, the method adds the array returned in line 4 or 6 into a matrix. After this process, in line 8, the matrix of mobility metrics (as parameter) and the matrix of social indicators are concatenated, forming a single matrix. In line 9, this matrix is passed to the Apache API, where the correlation matrix will be constructed by correlating each column of the original matrix. The result is a matrix containing the correlation coefficients between the variables.

To perform these experiments, we divided the users into three categories: Category 1, 2 and 3. Respectively, they group the users who have posted at least 1,000 messages (679 users); 2,500 messages (168 users); and 5,500 messages (33 users). This division was made with the objective of identifying correlations that can only be found for heavy users, possibly due to the imprecision and fragmentation of messages posted in the Twitter network. Figures 4, 5 and 6 show a heat map for each of these categories, indicating the density of posts for the entire region of London. The darkest regions indicate a high density of posts.

In the Figures 4, 5 and 6, it is possible to notice the variations on density when we reduce the number of users, considering users with the highest number of posts (Figure 6) where, in this case, the users have at least 5,500 messages. From Figure 4 (users with at least 1,000 messages), it can be observed that there is a tendency of users to post their messages in the central regions of the city. This tendency, and a possible bias, clearly reduces when we consider the users of Categories 2 and 3.



Fig. 4. Density of posts for users with more than 1,000 messages in Twitter (Category 1)

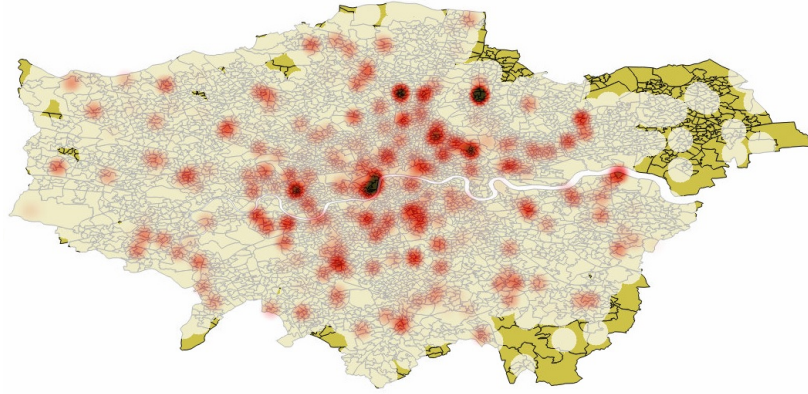


Fig. 5. Density of posts for users with more than 2,500 messages in Twitter (Category 2)

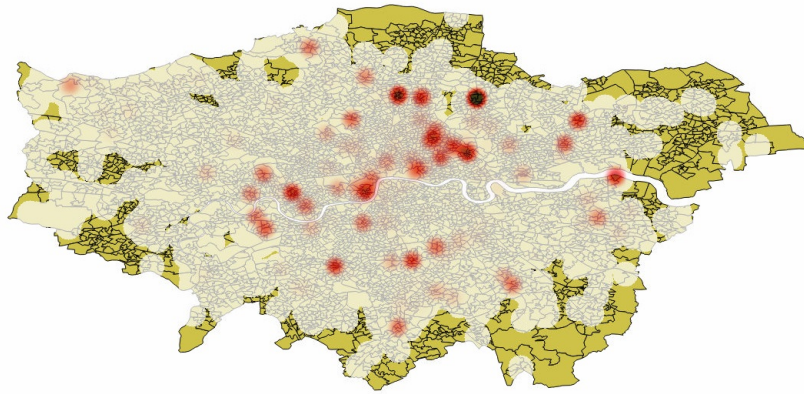


Fig. 6. Density of posts for users with more than 5,500 messages in Twitter (Category 3)

## 5. RESULTS AND DISCUSSION

The histograms of the radius of gyration and displacement distance [Cheng et al. 2011] are shown in Figure 7. In these histograms, it is possible to visualize the frequency at which the values of both variables occur in the extracted mobility dataset. For a clearly visualization of the values, the displacement distance histogram is shown in a log10 scale while the radius of gyration histogram is represented by its original values. Both variables were calculated in meters. In the first histogram, it can be seen that the majority of users have their radius of gyration between 3,000 meters and 4,000 meters, totalizing 7,554 users. The second histogram shows that most users have their displacement distance between 100,000 meters and 316,228 meters, representing 19,514 users.

In order to capture correlations that could only be found in specific periods of a daily life, the messages were analyzed considering temporal ranges. For this, we considered messages posted in workdays and messages posted during the weekends. Since that these two temporal ranges can express distinct behaviors, for example, working routine during the workdays and recreation during the weekends, they may reveal correlations that could not be found without considering these ranges.

Aiming at answering Question Q1, after the generation of the correlation matrix, we selected the most relevant correlations found by the model. No significant correlation has been found for users from Categories 1 and 2 (users with at least 1,000 and 2,500 messages); however, some significant



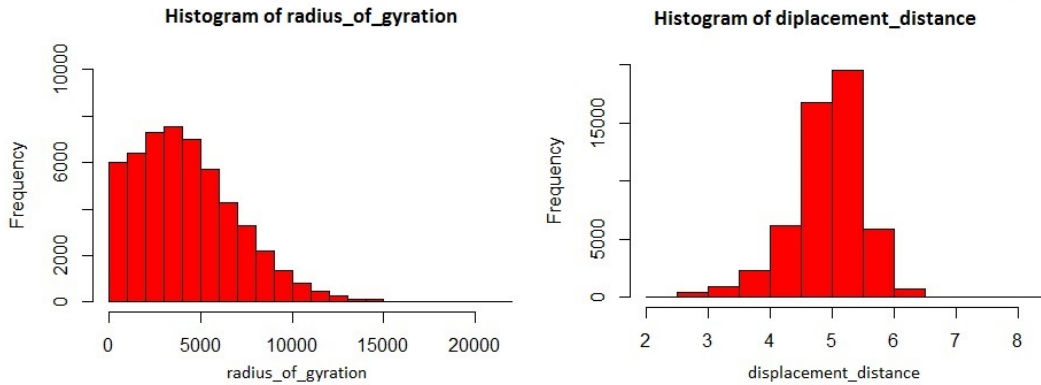


Fig. 7. Radius of gyration and displacement distance [Cheng et al. 2011] histograms. The displacement distance histogram is in a log10 scale.

correlations were found for users from Category 3. The most significant correlations found for this group are shown in Table II and Table III. In these tables, we used the notation of a tuple  $(\tau_b, p\text{-value})$ , where the first element represents the Kendall's tau-b (the correlation coefficient for this test), and the second represents the significance of the executed test, where a  $p\text{-value}$  less than 0.05 allows the rejection of the null hypothesis for the correlation test, denoting that there may be a real correlation between the metrics of mobility and social data.

Table II shows the correlations between the radius of gyration and the social indicators of the home regions of the users. In this table, it can be observed that a significant number of correlations were found between this mobility indicator and the social variables related to employment conditions. As an example, we found correlations between the radius of gyration and the social indicator "Employment rate", with  $\tau_b = 0.26$ . This positive correlation suggests that when the value of the radius of gyration increases for a user, the social indicator "Employment rate" of his home region tends to increase also. Now, considering the social indicator "Persons with no qualifications", we can observe a negative correlation of  $\tau_b = -0.27$  without considering any temporal range, and  $\tau_b = -0.30$  in weekends. This negative correlation suggests that users that have a higher radius of gyration tend to live in regions that the social indicator "Persons with no qualification" is lower.

For the Table III, that shows the correlations between the mobility indicator User Displacement Distance and the social indicators, it can also be seen a predominance of correlations related to employment conditions. For instance, the correlation between this mobility indicator and the social variable "Unemployment rate" is  $\tau_b = -0.38$ , indicating that as the mobility indicator increases, the social indicator "Unemployment rate" of home regions tends to decrease. This result indicates that users with longer traveled distances tend to live in regions with the low unemployment rate, even considering messages posted in workdays and weekends. Similar results were found for the social

Table II. Results for Experiment 1 (correlating radius of gyration with social variables). Values for users from Category 3.

Social Data	Without temporal range	Workdays	Weekends
Economically inactive people	(-0.35, 0.003)	(-0.40, 0.001)	-
Employment rate	(0.26, 0.03)	(0.30, 0.01)	(0.25, 0.04)
Persons with no qualifications	(-0.27, 0.02)	-	(-0.30, 0.01)
Economically active unemployed	-	-	(-0.28, 0.02)
Unemployment rate	-	-	(-0.29, 0.01)
Day to day activities limited a little	-	-	(-0.25, 0.04)
Muslim	-	-	(-0.26, 0.03)

Table III. Results for Experiment 1 (correlating user displacement distance with social variables). Values for users from Category 3.

Social Data	Without temporal range	Workdays	Weekends
Age 0-15	(-0.27, 0.02)	-	(-0.33, 0.007)
Age 65 plus	-	-	(-0.26, 0.03)
Couple with children	(-0.31, 0.01)	(-0.29, 0.01)	(-0.35, 0.003)
Black /african/caribbean/black british	-	-	(-0.25, 0.03)
Economically inactive people	(-0.28, 0.02)	(-0.32, 0.007)	-
Employment rate	(0.29, 0.01)	(0.30, 0.01)	(0.26, 0.03)
Unemployment rate	(-0.38, 0.001)	(-0.36, 0.002)	(-0.37, 0.002)
Persons with no qualifications	(-0.29, 0.01)	(-0.29, 0.01)	(-0.29, 0.01)
Sales	-	(0.25, 0.03)	(0.25, 0.03)
Economically active unemployed	-	(-0.30, 0.01)	(-0.31, 0.01)
Day to day activities limited a little	-	(-0.36, 0.003)	(-0.37, 0.002)

Table IV. Results for Experiment 2 (correlating the radius of gyration and social variables related to users' activity centers). Values for users from Category 3.

Social Data	Without temporal range	Workdays	Weekends
Employment rate	(0.25, 0.03)	(0.27, 0.0401)	-
Persons with no qualifications	(-0.31, 0.01)	(-0.36, 0.005)	-
No cars or vans in household	-	(-0.28, 0.03)	-
Economically active unemployed	-	-	(-0.27, 0.04)

indicators "Economically inactive people", "Persons with no qualifications" and "Economically active unemployed".

The results obtained for Experiment 1 allowed us to answer Question Q1, since we could find correlations with some statistical significance, denoting that there are correlations between users' mobility metrics and social aspects of their home location, specially for variables related to employment conditions.

For Experiment 2, where we analyze possible correlations between the two mobility metrics (radius of gyration and displacement distance) and the social variables related to users' activity centers, we found no significant correlations for users from Categories 1 and 2. Again, the most significant correlations were found for user from Category 3.

Table IV shows the results of the correlations between the radius of gyration of the users and the social indicator related to the region of their Activity Centers. These results show some conformance with the first experiment. Here, again, the occurrence of social indicators related to employment conditions is greater than any other social indicator. For instance, there is a positive correlation between the radius of gyration and the social indicator "Employment rate", with  $\tau_b = 0.25$  for the messages without a temporal range, and messages posted in workdays. This result indicates that users with a high radius of gyration tend to visit regions that the social indicator "Employment rate" is also high. An opposite result was found with the radius of gyration and the social indicator "Persons with no qualifications", with  $\tau_b = -0.36$  considering workdays, denoting that, as the radius of gyrations of user increases, he will tend to visit regions with a low number of persons without professional qualifications. For the Table V, that shows the correlations for the mobility indicator "User Displacement Distance", the highest correlation was found with the variable "Employment Rate", with a  $\tau_b = 0.38$  for workdays. For this table, it is also possible to observe the prevalence of social indicators related to employment features.

It is important to note that even finding possible correlations between the mobility metrics and social data, these correlations were not classified as strong correlations, as the highest value was of  $\tau_b = -0.40$ . We believe that the fragmented nature of Twitter messages can add some inaccuracies to the results. For example, poor users might post significantly more than users from rich locations, bringing

Table V. Results for Experiment 2 (correlating the user displacement distance and social variables related to users' activity centers). Values for users from Category 3.

<b>Social Data</b>	<b>Without temporal range</b>	<b>Workdays</b>	<b>Weekends</b>
Employment rate	(0.36, 0.002)	(0.38, 0.002)	(0.26, 0.04)
Unemployment rate	(-0.32, 0.007)	-	(-0.29, 0.03)
Persons with no qualifications	(-0.28, 0.02)	(-0.28, 0.03)	-
Couple household without dependent children	-	(0.28, 0.02)	-
Economically inactive people	-	(-0.27, 0.03)	-
Economically active unemployed	-	-	(-0.30, 0.02)
Day to day activities limited a lot	-	(-0.33, 0.01)	-
3 cars or vans in household	-	(0.28, 0.03)	-
Muslim	-	(-0.31, 0.01)	-
Other religion	-	(0.26, 0.04)	-

imprecisions to the results. This kind of problem was partially mitigated by the segmentation of users based on the number of messages they have posted (Categories 1, 2 and 3). For users in Category 3, which provided the best results, we could visually observe that they were homogeneously distributed over the city of London. Furthermore, it is not possible to extrapolate the results obtained from the correlations to the whole population of London, as these results were based on certain Twitter profiles.

## 6. CONCLUSIONS

Social networks are playing an important role in enabling the sharing of information across the Internet, increasing the amount of data that are public available for scientific studies. In accordance with this scenario, governments are tending to make their data available for public access, what favor studies considering these two kinds of data.

This research presented a model to allow the identification of correlations between mobility patterns and social, economic and demographic variables. This model identifies mobility metrics from georeferenced Twitter messages, detect users' home locations and activity centers, then looks for correlations with the social data supplied to the model. An experimental evaluation was conducted using data from the city of London. This city was chosen due to the high availability of Twitter messages in the time interval where these messages were collected and also for the availability of many social indicators for this city.

This study confirms that it is possible to identify some correlations between mobility data extracted from social media and social indicators. In the results obtained from our experiments, relevant correlations were found for variables associated with employment conditions (economically inactive people, employment rate, unemployment rate and persons with no qualifications). In practice, the presented model can provide knowledge for governments, facilitating the decisions for investments in specific areas of a city, increasing the integration of segregated populations. The results can also help recommendation systems in performing suggestions of products related to the mobility behavior of a population of specific regions.

Additionally, the fragmented nature of Twitter messages makes the task of finding correlations even challenging, forcing us to reduce the number of users to be considered in the experiments (only those with a large number of posts). This indicates the need of performing additional experiments involving more heavy users, to make the correlations more significant. Further work also includes applying this model to the analysis of other regions in the world and the development of a graphical user interface using the proposed model. Moreover, we intend to formulate additional mobility metrics, enhancing the analysis of mobility and the discovery of relevant correlations.

## REFERENCES

- BAGROW, J. P. AND LIN, Y.-R. Mesoscopic structure and social aspects of human mobility. *PloS one* 7 (5): e37676, 2012.
- BIRKIN, M., HARLAND, K., MALLESON, N., CROSS, P., AND CLARKE, M. An examination of personal mobility patterns in space and time using twitter. *International Journal of Agricultural and Environmental Information Systems (IJAEIS)* 5 (3): 55–72, 2014.
- BLANFORD, J. I., HUANG, Z., SAVELYEV, A., AND MACEachREN, A. M. Geo-located tweets. enhancing mobility maps and capturing cross-border movement. *PloS one* 10 (6): e0129202, 2015.
- CHARENTREAU, A., HUI, P., CROWCROFT, J., DIOT, C., GASS, R., AND SCOTT, J. Impact of human mobility on opportunistic forwarding algorithms. *IEEE Transactions on Mobile Computing* 6 (6): 606–620, 2007.
- CHENG, Z., CAVERLEE, J., LEE, K., AND SUI, D. Z. Exploring millions of footprints in location sharing services. *ICWSM* vol. 2011, pp. 81–88, 2011.
- CHOK, N. S. *Pearson's versus Spearman's and Kendall's correlation coefficients for continuous data*. M.S. thesis, University of Pittsburgh, 2010.
- CROUX, C. AND DEHON, C. Influence functions of the spearman and kendall correlation measures. *Statistical Methods & Applications* 19 (4): 497–515, Nov, 2010.
- DE OLIVEIRA, M. G. *Ontology-driven urban issues identification from social media*. Ph.D. thesis, Federal University of Campina Grande, Brazil, 2017.
- ESTER, M., KRIEGEL, H.-P., SANDER, J., XU, X., ET AL. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. Vol. 96. AAAI Press, Menlo Park, California, pp. 226–231, 1996.
- GONZALEZ, M. C., HIDALGO, C. A., AND BARABASI, A.-L. Understanding individual human mobility patterns. *Nature* 453 (7196): 779–782, 2008.
- HAO, Q., CAI, R., WANG, C., XIAO, R., YANG, J.-M., PANG, Y., AND ZHANG, L. Equip tourists with knowledge mined from travelogues. In *Proceedings of the 19th international conference on World wide web*. ACM, Proceeding WWW '10 Proceedings of the 19th international conference on World wide web, Raleigh, North Carolina, USA, pp. 401–410, 2010.
- HASAN, S., ZHAN, X., AND 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*. UrbComp '13. ACM, New York, NY, USA, pp. 6:1–6:8, 2013.
- HAWELKA, B., SITKO, I., BEINAT, E., SOBOLEVSKY, S., KAZAKOPOULOS, P., AND RATTI, C. Geo-located twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science* 41 (3): 260–271, 2014.
- HSIEH, H.-P., LI, C.-T., AND LIN, S.-D. Exploiting large-scale check-in data to recommend time-sensitive routes. In *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*. UrbComp '12. ACM, New York, NY, USA, pp. 55–62, 2012.
- HUANG, Q., CAO, G., AND WANG, C. From where do tweets originate?: A gis approach for user location inference. In *Proceedings of the 7th ACM SIGSPATIAL International Workshop on Location-Based Social Networks*. LBSN '14. ACM, New York, NY, USA, pp. 1–8, 2014.
- HUNG, H. M. J., O'NEILL, R. T., BAUER, P., AND KOHNE, K. The behavior of the p-value when the alternative hypothesis is true. *Biometrics* 53 (1): 11–22, 1997.
- JERÔNIMO, C. L. M., CAMPELO, C. E., AND DE SOUZA BAPTISTA, C. Analyzing mobility patterns from social networks and social, economic and demographic open data. In *Proceedings of the XVII Brazilian Symposium on Geoinformatics (GeoInfo 2016)*. pp. 32–43, 2016.
- JIANG, S., FIORE, G. A., YANG, Y., FERREIRA, JR., J., FRAZZOLI, E., AND GONZÁLEZ, M. C. A review of urban computing for mobile phone traces: Current methods, challenges and opportunities. In *Proceedings of the 2Nd ACM SIGKDD International Workshop on Urban Computing*. UrbComp '13. ACM, New York, NY, USA, pp. 2:1–2:9, 2013.
- JURDAK, R., ZHAO, K., LIU, J., ABOUJAOUDE, M., CAMERON, M., AND NEWTH, D. Understanding human mobility from twitter. *PloS one* 10 (7): e0131469, 2015.
- KNIGHT, W. R. A computer method for calculating kendall's tau with ungrouped data. *Journal of the American Statistical Association* 61 (314): 436–439, 1966.
- LUO, F., CAO, G., MULLIGAN, K., AND LI, X. Explore spatiotemporal and demographic characteristics of human mobility via twitter: A case study of chicago. *Applied Geography* vol. 70, pp. 11 – 25, 2016.
- NGUYEN, T. AND SZYMANSKI, B. K. Using location-based social networks to validate human mobility and relationships models. In *Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on*. IEEE, ASONAM, Istanbul, Turkey, pp. 1215–1221, 2012.
- NOULAS, A., SCCELLATO, S., LAMBIOTTE, R., PONTIL, M., AND MASCOLO, C. A tale of many cities: universal patterns in human urban mobility. *PloS one* 7 (5): e37027, 2012.



- PALCHYKOV, V., MITROVIĆ, M., JO, H.-H., SARAMÄKI, J., AND PAN, R. K. Inferring human mobility using communication patterns. *arXiv preprint arXiv:1404.7675* 4 (6174): 6, 2014.
- RHEE, I., SHIN, M., HONG, S., LEE, K., KIM, S. J., AND CHONG, S. On the levy-walk nature of human mobility. *IEEE/ACM transactions on networking (TON)* 19 (3): 630–643, 2011.
- SHIN, R., HONG, S., LEE, K., AND CHONG, S. On the levy-walk nature of human mobility: Do humans walk like monkeys? In *Proceeding IEEE INFOCOM*. pp. 924–932, 2008.
- STEIGER, E., WESTERHOLT, R., RESCH, B., AND ZIPF, A. Twitter as an indicator for whereabouts of people? correlating twitter with uk census data. *Computers, Environment and Urban Systems* vol. 54, pp. 255–265, 2015.
- WILSON, T. AND BELL, M. Comparative empirical evaluations of internal migration models in subnational population projections. *Journal of Population Research* 21 (2): 127–160, 2004.
- YIN, H., CUI, B., HUANG, Z., WANG, W., WU, X., AND ZHOU, X. Joint modeling of users' interests and mobility patterns for point-of-interest recommendation. In *Proceedings of the 23rd ACM International Conference on Multimedia*. MM '15. ACM, New York, NY, USA, pp. 819–822, 2015.
- YUAN, Q., CONG, G., MA, Z., SUN, A., AND THALMANN, N. M. Who, where, when and what: Discover spatio-temporal topics for twitter users. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '13. ACM, New York, NY, USA, pp. 605–613, 2013.
- ZHANG, Y., WANG, L., ZHANG, Y.-Q., AND LI, X. Towards a temporal network analysis of interactive wifi users. *EPL (Europhysics Letters)* 98 (6): 68002, 2012.
- ZHAO, K., MUSOLESI, M., HUI, P., RAO, W., AND TARKOMA, S. Explaining the power-law distribution of human mobility through transportation modality decomposition. *arXiv preprint arXiv:1408.4910* 5 (9136): 21, 2014.
- ZHENG, V. W., ZHENG, Y., XIE, X., AND YANG, Q. Collaborative location and activity recommendations with gps history data. In *Proceedings of the 19th International Conference on World Wide Web*. WWW '10. ACM, New York, NY, USA, pp. 1029–1038, 2010.