Cities 89 (2019) 268–280

Contents lists available at ScienceDirect

Cities

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

Social media and urban mobility: Using twitter to calculate home-work travel matrices

Joaquín Osorio-Arjona*, Juan Carlos García-Palomares

Departamento de Geografía Humana, Universidad Complutense de Madrid, C/Profesor Aranguren, s/n, 28040 Madrid, Spain

ARTICLE INFO

Keywords:
Social networks
Mobility
Geographic information systems
Twitter
Home-work matrices

ABSTRACT

The proliferation of Big Data is beneficial to the study of mobility patterns in cities. This work investigates the use of social media as an efficient tool for urban mobility studies. In this case, the social network Twitter has been used, due to its wealth of spatial and temporal data and the possibility of accessing data free of charge. Using a database of geotagged tweets in the Madrid Metropolitan Area over a two-year period, this article describes the steps followed in the preparation and cleansing of the initial data and the visualisation of the results in Geographic Information Systems in the form of home-work matrices. The Origin-Destination matrices obtained were then compared with the official data provided by the Madrid Transport Consortium from the 2014 Synthetic Mobility Survey. The results of this comparison demonstrate that the level of precision offered by Twitter as a source of geographic information is adequate and efficient, thereby permitting a more in-depth analysis of flows between different zones of interest in the study area.

1. Introduction

The monitoring company *StatCounter* recently revealed that, for first time since 1980, *Android* has replaced the *Windows* operating system as the main internet access mode (Simpson, 2017). The virtual world is entering a new era, in which mobile phones are displacing computers as the main means of interacting with society online. Enormous growth in the use of smartphones affects the quantity of information being generated. In an increasingly technological society, with an ever-greater use of internet, nearly all large cities' inhabitants generate a digital footprint of their activities and movements, a digital trail that can be followed (Blanford, Huang, Savelyev, & MacEachren, 2015). One of the consequences of this digitalization of society is the prominent role of so-called Spatial Big Data: large quantities of spatial information that can be captured, communicated, aggregated, stored, and analysed (Manyika et al., 2011). Shadows of this digitalization are intimately intermingled with offline, material geographies of everyday life (Jin et al., 2017). This new geographic data is produced constantly, in real time, can be acquired easily at a low cost, and can be incorporated and analysed in Geographic Information Systems (Osorio & García-Palomares, 2017).

Internet users are no longer mere passive recipients of information, but have become producers of vast amounts of data, particularly through social networks (García-Palomares, Salas-Olmedo, Moya-Gómez, Condeço-Melhorado, & Gutiérrez, 2018). These social networks enable information to be created, stored, shared, and exchanged with other users, and generate a huge volume of data every day (Cao et al., 2014). Social media data have been used to study the shape of urban agglomeration based on people's activity (Zhen, Cao, Qin, & Wang, 2017), and are useful to analyse urban structure and related socio-economic performance (Martí, Serrano-Estrada, & Nolasco-Cirugeda, 2017; Shen & Karimi, 2016; Zhang, Zhou, & Zhang, 2017). Twitter, an application based on sending short texts or other multimedia content, stands out among the most-used social networks. Geo-located Twitter is a free and easily available global data source that stores millions of digital records on human activity in space and time (Hawelka et al., 2014). According to the company's data, in 2017, approximately 500 million tweets were sent every day all over the planet. According to Twitter's own data,[1] the network has a base of 328 million active users per month, of which 82% generated information from their smartphones.

The new data sources provide information that is extremely useful in mobility studies. If we know the changing location of each person based on their digital footprint, we can analyse their general mobility patterns. Gathering geospatial data provides a unique opportunity to gain valuable insight into information flow and social networking

* Corresponding author.
E-mail addresses: joaquoso@ucm.es (J. Osorio-Arjona), jcgarcia@ghis.ucm.es (J.C. García-Palomares).

[1] Twitter (2016). Twitter usage/company facts. https://about.twitter.com/es/company (Accessed 11/05/).

https://doi.org/10.1016/j.cities.2019.03.006
Received 3 September 2018; Received in revised form 23 January 2019; Accepted 11 March 2019

within a society (Stefanidis, Crooks, & Radzikowski, 2013). Evolution of transport-demand model techniques have developed the need for a high-resolution database, with aggregated economic and socio-demographic attributes for daily travel behaviours modelling (Rashidi, Abbasi, Maghrebi, Hasan, & Waller, 2017). In this respect, Twitter is a particularly valuable source of data for the study of mobility, due to its ease of use on any mobile device, and because it offers the possibility of geotagging messages. Thus, the tweets of users who activate the geo-location service on their accounts, contain, in addition to semantic and temporal information, information about the geographic location from whence the tweet has been sent. It is possible to carry out studies on population distribution at any time (see, for example, Ciuccarelli, Lupi, & Simeone, 2014; Longley & Adnan, 2016, or García-Palomares et al., 2018).

Data from social networks, and in general, most of the data from sources associated with Big Data, are not data created to analyse urban mobility. This is common in the use of social networks in disciplines such as urban planning and mobility. Consequently, the use of social network data to study mobility bears great weaknesses in comparison with data from sources specifically created for this purpose, which are normally household mobility surveys. One of the conditioning factors is the importance of cleaning and pre-processing processes for data, so that the database that is finally used solely contains data from whence reliable mobility information is obtained. Another conditioning factor to bear in mind is data limitation. Thus, social networks provide a low temporal resolution, which is the average time between each tweet sent by a user. This makes it difficult, for example, to obtain data from activity sites related for different reasons to the residence and the workplace. The low temporal resolution also compels data to be collected from much greater time periods. With telephony, we can normally work with 2 or 3 months. However, with social networks, we need greater period samples (for example, 1–2 years). On the other hand, while mobile telephone data refer to very large user samples, samples are smaller in social network data, and are also more biased, since users of these social networks are concentrated into certain sociodemographic groups. Finally, using social network data for a purpose such as studying mobility requires a results validation process. Here, we have made an attempt to insist on this phase, and to validate results at different spatial aggregation scales, in order to see how far we can go with these data in obtaining travel matrices.

In studying urban mobility, the ability to generate Origin-Destination (OD) travel matrices is a fundamental tool for carrying out diagnostics, predicting demand for travel and contributing to modelling of transport and optimisation of network usage (Gao et al., 2014). OD flow matrices provide useful information for ridership forecasting, service planning, and control strategies: for modelling purposes, origins and destinations of movements and modal preferences are needed (Rashidi et al., 2017). Traditionally, household surveys or traffic counts were used to obtain travel matrices. However, these types of sources are expensive, time-consuming, static, and use small population samples (Iqbal, Choudhury, Wang, & González, 2014). Social networks in general, and Twitter in particular, provide new sources for estimating OD travel matrices. They enable us to work with high spatial-temporal resolution data, to access larger population samples, and to obtain this data free of charge (when downloaded via streaming). The aim of this paper is to validate the use of Twitter data in assessing urban mobility patterns by using the Madrid Metropolitan Area as case study. Trips from home to workplaces were chosen as a study motive. The economic value of the workplace is not of academic interest alone, but also of practical interest, as they are key spaces for urban design and planning. Understanding how workers move throughout the day is important to design and manage transport systems and public spaces (Pajević & Shearmur, 2017).

Previous studies focused on obtaining OD travel matrices through new data sources mainly used phone data, so-called *call detail records* (CDRs) (Chen, Ma, Susilo, Liu, & Wang, 2016). However, mobile phone companies are extremely reluctant to hand over their data, and if they do, charges are substantial. Furthermore, information from CDRs is usually linked to the coverage range of the antennae, which means that one frequently works with Voronoi. Using CDRs, spatial resolution is occasionally lower, and they frequently do not match the distribution of land uses or transport zones which are necessary for an accurate modelling of future situations (Järv, Tenkanen, & Toivonen, 2017). Bearing these two main limitations of previous studies based on phone data in mind, the main contribution of this paper is to validate Twitter as an alternative source to phone data in estimating OD travel matrices and detecting home-work trips, benefiting from: a) free-of-charge data; b) the availability of geolocated tweets and the possibility of spatially joining them to transport zones (which are of great interest to transport managers).

When processing Twitter data to obtain the matrices, some methodological improvements were carried out, as opposed to previous works. Thus, information on land uses from Land Registry (Cadastre) maps was used. These maps benefit from high spatial detail, offering greater precision when it comes to detecting home and work places. Furthermore, various factors of expansion were evaluated for the sample. On one hand, the authors worked with expansions based on the origins of the trip, using population data according to place of residence from official sources (Census). On the other hand, the matrices were expanded using data relating to the destination of the trip, using information about the workplace (from National Institute of Social Security records). The estimations were carried out at two different scales of analysis: a) municipality level (with finer spatial detail); b) metropolitan-zone level. The results were validated via comparison with the data from the Synthetic Mobility Survey carried out by the Madrid Transport Consortium in 2014, thereby enabling us to evaluate the quality of the matrices obtained by Twitter data and the sensitivity thereof, both to the type of expansion factor used and to the degree of spatial desegregation.

This article is divided into six sections. After this introduction, Section 2 reviews the research carried out to date using Twitter and OD matrices. Section 3 defines the study area, and Section 4 introduces the data and sources we used. Meanwhile, Section 5 reviews the methodology used in the research. The obtained results are analysed in Section 6 to establish a series of conclusions in Section 7.

2. Literature review

Urban mobility is increasingly more complex and less sustainable, characterised by an increasing number of trips, a greater diversity of reasons for travelling, more intensive use of motorized transport and longer trips (Gutiérrez & García-Palomares, 2007). Faced with this complexity, traditional sources need to be complemented with new data with higher spatial and temporal resolutions. In this context, due to their possibilities for tracking citizen's digital footprints, Big Data emerge as an extremely interesting source in urban mobility studies. These opportunities are shown in the increasing number of studies on urban mobility based on Big Data for estimating OD travel matrices. Chen et al. (2016) carried out a magnificent review of the characteristics of Big Data, the methodological needs required for its use and the problems and opportunities it presents for the study of mobility. This review pays special attention to the use of mobile phone CDRs, the most-used source until now.

The first works based on mobile phone data were used to estimate travel speeds and trip times (Bar-Gera, 2007), to locate home-work anchor points (Ahas, Silm, Järv, Saluveer, & Tiru, 2010), or to calculate the level of traffic on roads that did not have conventional gauging stations (Caceres, Romero, Benitez, & del Castillo, 2012). However, thanks to the temporal resolution offered by mobile phone activity, it is possible to identify the periods when individuals are at a given place and the trips they make between places: their spatial-temporal trajectories. Once these trajectories are obtained for each user, it is possible

to analyse daily mobility patterns and formulate predictive models (De Domenico, Lima, & Musolesi, 2013) or to generate Origin-Destination matrices that quantify the volume of trips (Alexander, Jiang, Murga, & González, 2015; Bonnel, Hombourger, Olteanu-Raimond, & Smoreda, 2015; Caceres, Wideberg, & Benitez, 2007; Kung, Greco, Sobolevsky, & Ratti, 2014; Louail et al., 2015; Toole et al., 2015).

While these works took advantage of the significant sample size and high frequency of mobile phone data (Picornell et al., 2015), the use of phone data presents many drawbacks, which can be partly overcome by using other sources such as social networks' data (Wu, Zhi, Sui, & Liu, 2014). Although mobile phone data offer a greater volume and level of temporal precision due to their high frequency, these sources are extremely difficult to obtain and are dependent upon companies, which means they frequently have an economic cost (occasionally very high), while information from social networks is easily accessible and free.

Twitter, one of the most-used social networks, is free when data are accessed via streaming. Furthermore, phone data have extremely low semantic value, whereas social network data contain information that provides insights into various socio-demographic aspects of the user (for example, the language) or, based on the content of the texts, into his or her opinions and perceptions.

Twitter data were used to analyse population density distribution patterns over the course of the day (Ciuccarelli et al., 2014). Longley and Adnan (2016) did similar work in London with ethnicities and social groups. Recently, García-Palomares et al. (2018) analysed these same hourly distributions of Twitter users, combining the information with types of land use. This information provides an initial overview of the origin and destination zones of trips. However, there were a few cases where Twitter was directly applied to mobility. Salas-Olmedo and Rojas Quezada (2017) mapped mobility patterns to the public spaces of the city of Concepción, in Chile, and, based on those flows, were able to identify potential areas of social exclusion. Other works have used Twitter data to study activity areas of different social groups (Luo, Cao, Mulligan, & Li, 2016; Netto, Pinheiro, Meirelles, & Leite, 2015; Shelton, Poorthuis, & Zook, 2015). Yin, Soliman, Yin, and Wang (2017) proposed alternative boundaries to the administrative boundaries, based on spatial interactions of Twitter users.

Focusing on Twitter as a tool for creating OD matrices, Perez, Dominguez, Rubiales, and Lotito (2015) developed a method for updating matrices based on vehicle traffic in the Buenos Aires metropolitan area using tweets that had been previously filtered and classified into different time periods. In Los Angeles, Gao et al. (2014) researched the efficacy of Twitter as a tool for creating OD matrices by comparing the results with data from the American Community Survey. To do this, they focused on detecting individual trajectories and aggregations of trips to places (using traffic zones as origin and destination areas). Then, through the Pearson correlation coefficient, they validated the results using American Community Survey data. Also, in Los Angeles, and taking the work of Gao et al. (2014) as a starting point, Lee, Gao, and Goulias (2015) sought to validate the quality of Twitter as a source, using an approach whereby they extracted OD relationships on weekdays (in a two-day period), aggregated by spatial zone, and validated them against other sources. For comparison in this work, they used the Tobit auto-regression model with data from the census and from traditional travel demand models. The reliability of Twitter for generating information on mobility was also validated in the work of Lenormand et al. (2014), who compared Twitter data with information from phone networks and official data (census) in the cities of Madrid and Barcelona, concluding that the three sources of information provide comparable results.

Table 1 shows previous works that made travel matrices with data from mobile telephones and Twitter. Telephony data are the big-data source most often-used in mobility studies, and Twitter is the main alternative. Our paper has made an attempt to improve the method used up until now, with the aim of improving and validating Twitter as an alternative to telephony data. Pioneer works using telephony data obtained the number of trips between zones in certain transport corridors (see, for example, Caceres et al., 2007). In a second phase, works were focused on calculated travel matrices with residence-work mobility (see Ahas et al., 2010 or Alexander et al., 2015). Some authors went a bit further, obtaining travel matrices that include other types of trips. This is the case with Picornell et al. (2015), who worked with travel categories such as “other frequent trips” and “infrequent trips”. To obtain travel matrices, these works had to determine, at minimum, the residences and workplaces of the individuals. To this end, they used the most frequent location during working hours (morning or daytime), and the most frequent location during the evening-night hours. Once the matrices were obtained, some works conducted processes to expand the data to the population as a whole, and to verify the results. Methods similar to those used with telephony were used in certain previous works that use Twitter data (for example, Lenormand et al. (2014)). However, barring Lee et al. (2015) with the use of specific land, none of these works have crossed telephony or Twitter data with land use to improve the definition of the residence and workplace. In this article, we did cross this data, enriching the Twitter data with a highly-detailed spatial layer of land use from the Land Registry. Afterward, we completed all stages to obtain travel matrices, expanding the data (based on two different sources: Population Census and Social Security records) and validating the results with two levels of spatial aggregation.

3. Study area

In evaluating the opportunities of Twitter data for estimating OD travel matrices, this study focuses on the Madrid Metropolitan Area. It has an estimated population of 6 million inhabitants (2017), of which 3.2 million live in the city of Madrid. Madrid is the metropolitan area with highest level of population, activity, and services in Spain.

We worked with two levels of spatial desegregation: an initial analysis with 71 spatial units, using the fifty municipalities of the metropolitan area and the twenty-one districts of the municipality of Madrid; and a second level of spatial aggregation in which these units were grouped into eight metropolitan zones, with similar levels of population and their own mobility behaviour patterns (Fig. 1).

4. Data

4.1. Twitter

The initial database for this work contains a total of 2,229,753 tweets, all geotagged and produced by 171,631 users located inside the Madrid Metropolitan Area. These tweets were compiled over a two-year period (from 1st June 2016 to 31st May 2018). Each tweet has information related to the user identification number (ID), username, latitude and longitude, date and time, language, and the hashtags it includes.

4.2. Data on resident and employment population

The data on the resident population used to expand the matrices were obtained from the 2017 census by the National Institute of Statistics (INE). The data related to the location of employment used to expand the travel destinations come from the records of the National Institute of Social Security in 2016 held on the Open Data Portal of the Madrid Council (for the districts in Madrid) and from Institute of Statistics of the Regional Government of Madrid (for the municipalities in the Metropolitan Area).

4.3. Land use data

The locations of the tweets were crossed with land use, adding data on the primary activity of the parcels where the messages are located. To do this, the authors worked with information from the 2017 Land

Table 1
Big Data mobility approaches using OD matrices.

Data source	Authors (year)	Home/work detection	Land data use	Data expansion	Data verification
Mobile phone data	Caceres et al. (2007)	No	No	Yes	No
	Ahas et al. (2010)	Yes	No	No	Yes
	Louail et al. (2015)	Yes	No	No	No
	Alexander et al. (2015)	Yes	No	Yes	Yes
	Bonnel et al. (2015)	No (already given)	No	Yes	Yes
	Toole et al. (2015)	Yes	No	Yes	Yes
	Picornell et al. (2015)	Yes	No	Yes	Yes
Twitter data	Gao et al. (2014)	No	No	No	Yes
	Lee et al. (2015)	No	Yes (business establishments)	No	Yes
	Perez et al. (2015)	Yes	No	Yes	No
	Salas-Olmedo and Rojas Quezada (2017)	No	No	Si	No
	Yin et al. (2017)	No	No	No	No
Mobile phone data and Twitter data	Lenormand et al. (2014)	Yes	No	Yes	Yes

Registry (cadastral data), which contains information with a high spatial level.

4.4. 2014 synthetic mobility survey

To verify and validate the accuracy of Twitter data in estimating OD travel matrices, the authors used the mobility data obtained by the Madrid Transport Consortium from the 2014 Synthetic Mobility Survey. The information in its data file includes codes for the origin and destination zones in the study area (at municipal and district level), reasons for the trip, time of departure, duration of the trip, means of transport used, type of ticket, and number of trips. In total, the Synthetic Mobility Survey recorded the trips of 8689 respondents. This survey uses transport zonification (homogeneous sociomorphological units) as a spatial reference for trip expansion.

5. Methodology

This section shows the methodology used, from obtaining the initial data to verification of results. Fig. 3 summarises the methodological flow.

5.1. Compilation, cleansing and enrichment of Twitter database

The initial tweet database used was collated by streaming Twitter API using a *Python* code and compiled into a *MongoDB* database. The initial tweet database was then incorporated into a Geographic Information System (GIS), and tagged according to their geographical coordinates. Two types of software were used: *QGIS* and *ArcGIS Pro*.

After storing the geotagged data, there was a filtering process and the data were downloaded again (Table 2). The steps below were

Fig. 1. Map of Madrid Metropolitan Area
References: 1) Central Districts, 2) Northern Districts, 3) Southwestern Districts, 4) Southeastern Districts, 5) Metropolitan South Municipalities, 6) Metropolitan East Municipalities, 7) Metropolitan North Municipalities, 8) Metropolitan West Municipalities.

Table 2
Filtering process and expanded download.

Filter	Tweets	Users	Tweets/users
Initial database	2,229,253	171,631	13.0
Detected from bot accounts	−441,716	−296	1492.3
Tweets sent at weekends and holiday	−396,830	−25,713	1.5
From users without spatial mobility	−62,496	−36,856	1.7
From users with low temporal extension	−35,244	−10,521	3.3
From users with less than 5 tweets	−46,213	−24,853	1.9
Database after filtering process	1,246,754	73,392	17.0
Final database with expanded download	1,536,261	73,392	20.9

followed:

- Bot accounts were identified with a database summarising by ID user. ID accounts with more than 1000 tweets and the same coordinates were removed.
- The next step consisted of filtering out tweets posted on Fridays after 2 p.m., at weekends and on public holidays. In this way, we only worked with tweets posted on weekdays.
- The next step in processing Twitter data consisted of eliminating users whose tweets are always generated from very similar locations (user accounts without spatial mobility). To do this, the authors measured the distances of the dispersion deviation of each user tweet and eliminated users with an average distance of less than 50 m in the location of all their tweets from the database (users who do not change location during the compiled data period).
- After this, users with a low temporal extension were eliminated, i.e., users whose messages were concentrated in a period of one week (possible tourists visiting the city).
- Lastly, of the remaining users, all of those who posted less than 5 tweets throughout the whole period covered by the database (low-activity users) were eliminated.

The result of all these cleansing processes was a database with 1,246,754 tweets generated by 73,392 users (Table 2).

Once the data had been cleansed and the users identified by their Twitter ID, the sample was expanded by downloading the most recent messages posted by each user, with the aim of increasing the spatial and temporal precision of the individual movements. This expansion is also free and provides us with the last 3200 tweets from each user for obtaining geotagged messages that were not captured via streaming (Huang & Wong, 2016). After this second downloading process, tweets that did not contain spatial information or that were not located in the study area or posted during the studied period were filtered out. The final database contains 1,536,261 tweets from 73,392 users.

The database obtained with the second download was enriched with information from the Land Registry. To this end, each tweet was associated with the code for the parcel from whence it was sent, and the predominant land use of said parcel. To do so, table-joining operations via a Geographic Information System tracking were used. Information on the parcel from whence the tweet is sent, and its predominant use was used in the next phase of residence and workplace detection. With overlay operations, the district code (for the city of Madrid) and the municipality code where each tweet was sent was assigned to said tweet.

5.2. Home and work places detection

Once the database was cleansed and enriched, the next step consisted of identifying the place of residence and work for each user. To do this, a differentiation was made between daytime (posted between 8:30 am and 8:30 pm) and night-time (the remainder) tweets. Tweets sent at night are associated with the user's residence, and tweets sent during the daytime are associated with the workplace. This methodology has traditionally been used in mobility works by using mobile phone data, using CDR (see, for example, Ahas et al., 2010; Alexander et al., 2015; Picornell et al., 2015).

Fig. 2. Example of home detection by mode parcel.

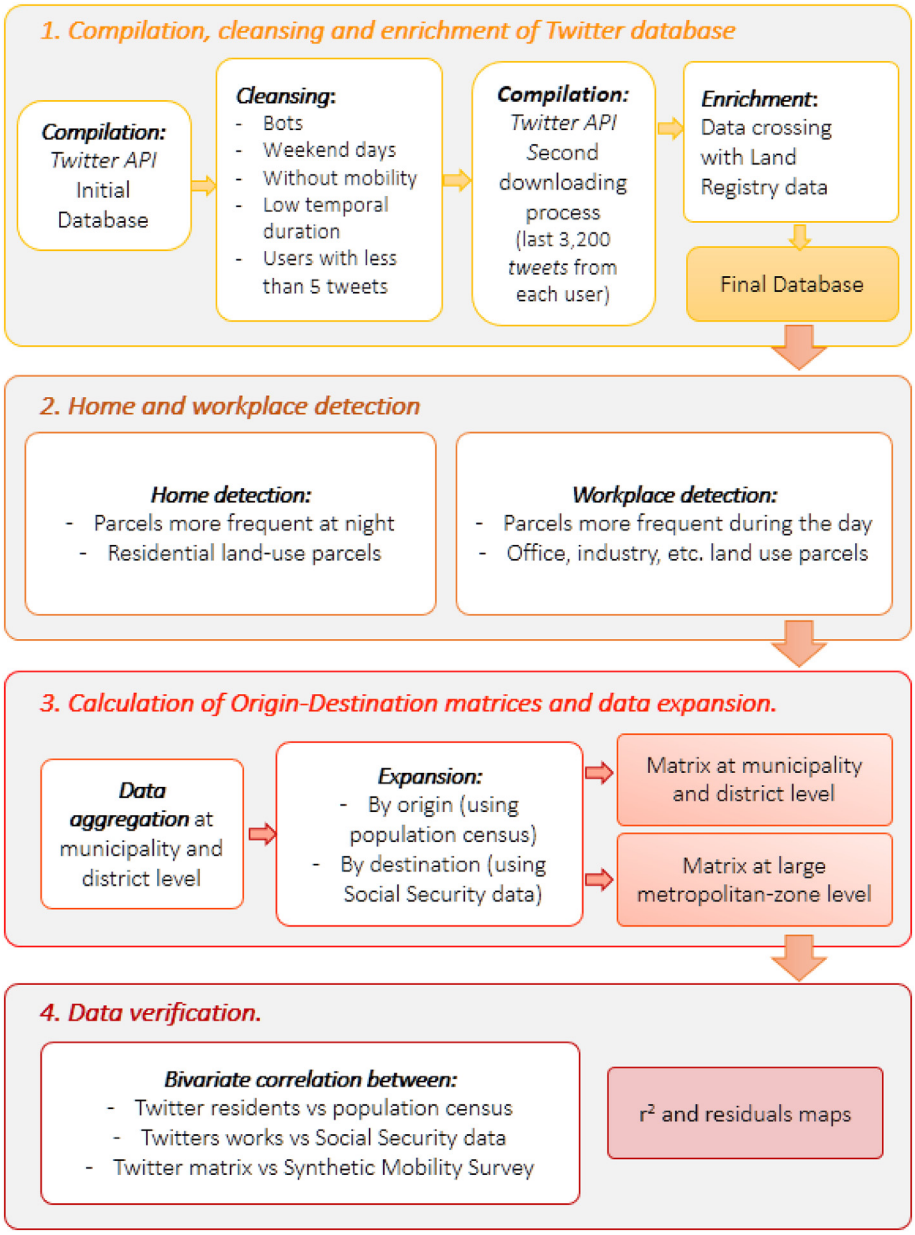

Fig. 3. Methodology diagram flux.

Therefore, to determine the place of residence, we worked with parcels from whence each user sends tweets with the highest frequency at night (residence). Crossing with land use improves this initial classification, solely considering parcels whose use is residential as a residence (Fig. 2). In this way, we eliminated possible residences that were erroneously associated with nighttime workers or leisure spaces.

When we found users with more than three residential parcels with tweets at night, and where the most frequent have the same number of tweets, we used the median centre tool included in the most customary GIS software (in our case, ArcGIS), and from those parcels, we selected the one nearest said median centre. When users with only two parcels and the same number of tweets at night appeared, we discarded them, since we could not clearly decide which of the parcels would be their residence.

The same procedure was followed to determine the location of the places of work of the users, this time working with the tweets posted during the day. Parcels with tweets which land use data predominant use is non-residential, work related (offices, industries, etc.) were selected, removing that way potential leisure time tweets. Finally, parcels where the user most frequently sent tweets were defined as the workplace. Median centre was once again used to establish one single parcel, in the event that one user had more than two parcels with the same number of tweets. Once again, when users with only two parcels and the same number of tweets during the day appeared, we discarded them, since we could not clearly decide which of the parcels would be their workplace.

This method follows, in part, the procedure of detecting home and work anchor points on CDR data by Ahas et al. (2010). This means that the method used here is based on the method customarily used with mobile telephony data, but uses land-use information to improve definition of residence and workplace detection. While Ahas et al. (2010) used network cells to pinpoint the location of mobile phone users, Land Registry data is an excellent alternative for accurate Twitter data enrichment, showing if a user tweeted from home location, work, or leisure centers like shops.

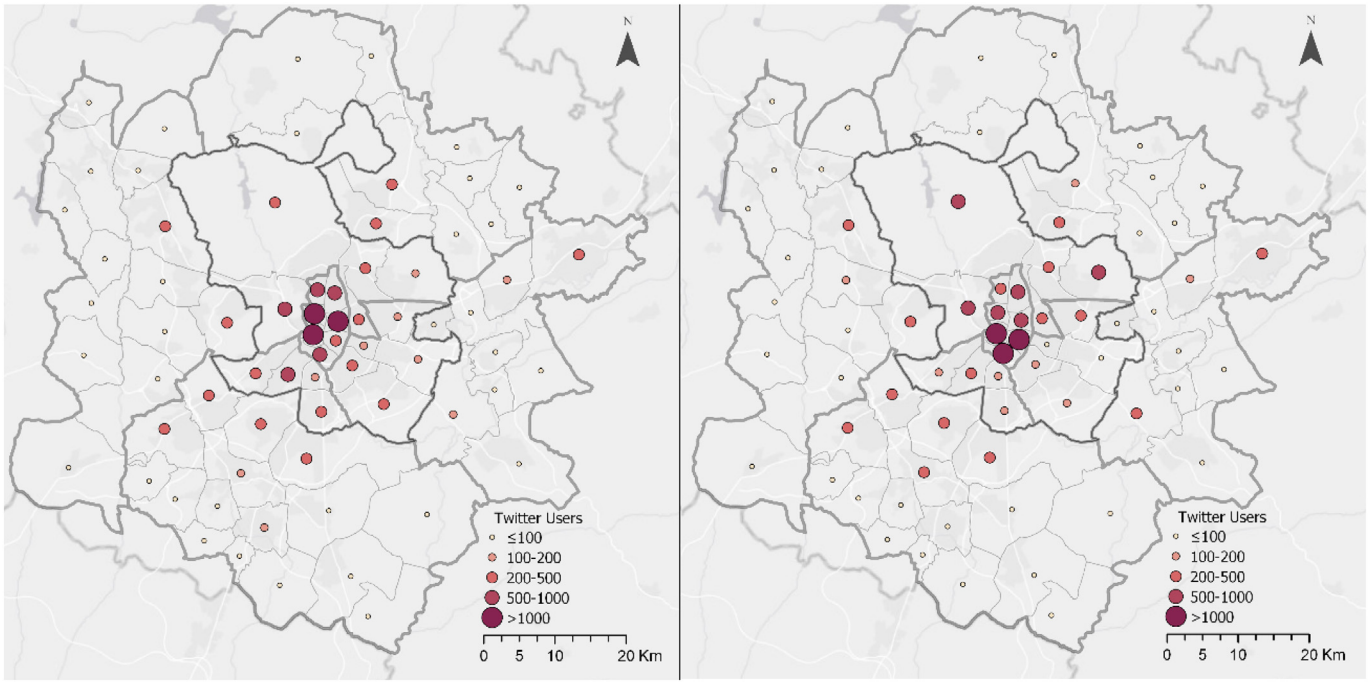

Fig. 4. Population based on points found in Twitter (left: homes; right: places of work).

Table 3
Distribution of the residential population by metropolitan zones.

Zones	Residents (Twitter)	Residents (Census)	Percentage
Central Districts	12,242	503,050	2.43
Northern Districts	1678	382,191	0.44
Southwestern Districts	1188	381,012	0.31
Southeastern Districts	909	331,670	0.27
Metropolitan South	2137	670,412	0.32
Metropolitan East	853	330,085	0.26
Metropolitan North	700	196,703	0.36
Metropolitan West	1037	272,134	0.38
Total	20,744	3,067,257	0.68

Table 4
Distribution of work places by metropolitan zones.

Zones	Work (Twitter)	Place of work (National Institute of Social Security)	Percentage
Central Districts	11,415	995,418	1.15
Northern Districts	3020	486,648	0.62
Southwestern Districts	769	184,728	0.42
Southeastern Districts	760	247,142	0.31
Metropolitan South	2077	348,601	0.60
Metropolitan East	1043	202,823	0.51
Metropolitan North	624	225,288	0.28
Metropolitan West	1036	263,771	0.39
Total	20,744	2,954,419	0.70

5.3. Calculation of origin-destination matrices and data expansion

The relationship matrix was calculated from the relationships between the home (origin) and the place of work (destination) using the user ID. Users with only origin or destination point were filtered out. The result is a matrix with a total of 20,744 users. This individual matrix was aggregated by 21 districts inside the municipality of Madrid and 49 municipalities in the metropolitan area. The total number of relationships in the matrix is 4900 (70 origins by 70 destinations).

The next phase consisted of expanding the travel matrix, coming from Twitter user data (sample), to a matrix representing the entire working-age population (19–55 years) in the Madrid metropolitan area. Two expansion processes were used to this end:

Based on travel origin data, the first was based on population data from the official Census, taking into account the resident population in a 19–55 year age range according to official census data, using the following formula:

$$T_{ij}^{e} = T_{ij} \cdot \frac{p_i}{\widetilde{p}_i}, \forall\, i \in N,$$

where $\frac{p_i}{\widetilde{p}_i}$ are the weights calculated for each district and municipality N, based on the ratio between the total number of p_i inhabitants from Census data, and the $\widetilde{p}_i$ Twitter user samples obtained in the matrix. These weights are multiplied by the value of the Twitter trips obtained in each T_{ij} flow, the result being expanded T_{ij}^{e} flows.

The second expansion factor was calculated by using data related to the workplace (destinations). This figure was divided by the number of workers registered with the National Institute of Social Security in each district or municipality. Using the following formula:

$$T_{ij}^{e} = T_{ij} \cdot \frac{w_j}{\widetilde{w}_j}, \forall\, j \in N,$$

where $\frac{w_j}{\widetilde{w}_j}$ are the weights calculated for each municipality and district N, based on the ratio between the total number of w_j workers from an official source (Institute of Social Security) in the municipality o district j, and the $\widetilde{w}_j$ Twitter-user workers obtained in the municipality o district j.

In short, two expanded matrices were obtained, one based on origin, using data on residents according to Census, and another based on destination, with the number of workers identified by the National Institute of Social Security as an expansion factor.

The expanded matrices from districts and municipalities were aggregated to obtain travel matrices at large metropolitan-zone level (Fig. 1).

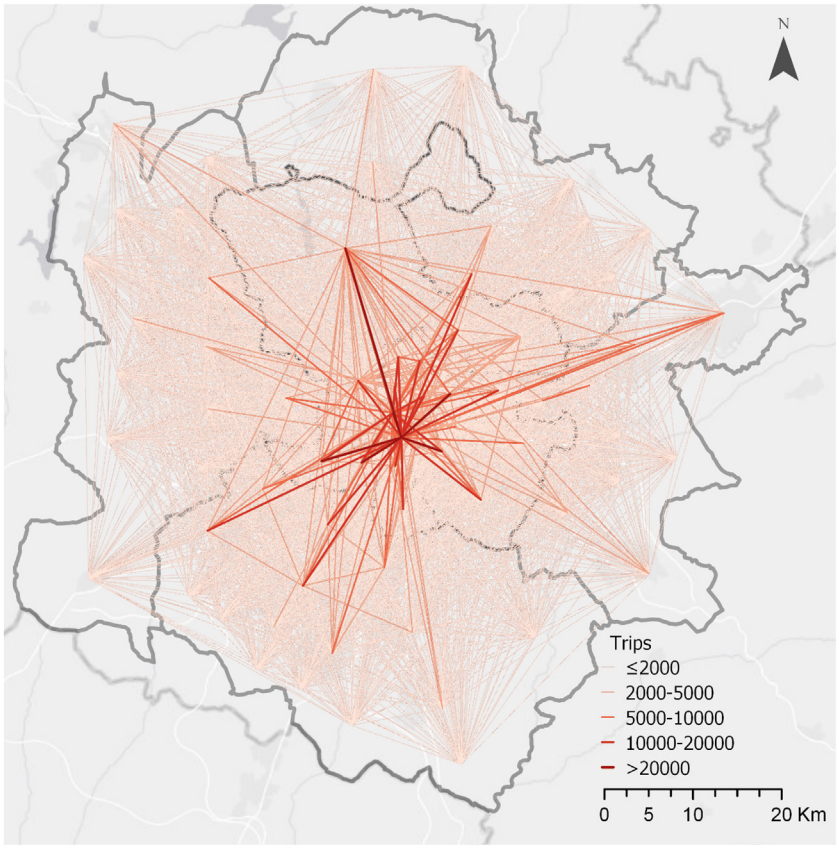

Fig. 5. Travel-flow matrix based on Twitter data in Madrid Metropolitan Area municipalities and districts, expanded by Census population data.

5.4. Data verification

To verify the obtained results, the coefficients of determination (squared Pearson correlation coefficient or r^2) between Twitter home data and census population data, and between Twitter workplaces and Social Security data were calculated. Then, the obtained matrices were evaluated both at district and municipal level and at the aggregated level with the distributions of trips from the Synthetic Mobility Survey of Madrid Transport Consortium. These r^2 coefficients were calculated using the trips from the Synthetic Mobility Survey where the given motive was work, using the following formula:

$$T_{ij}^{EMD} = \alpha + \beta . T_{ij}^{e} + \varepsilon$$

where T_{ij}^{EMD} are the Synthetic Mobility Survey flows and T_{ij}^{e} are our expanded flows calculated using Twitter. Lastly, residuals ε were calculated throught Ordinary Least Squares (OLS) linear regression and mapped to visualise the relationships that present the greatest deviations from the Synthetic Mobility Survey data.

Since the correlation between flows obtained using Twitter data and expanded by the destinations and Synthetic Mobility Survey of Madrid was weak, only the expanded matrix for origin data is shown in the results on its own.

Fig. 3 summarises the methodology used in this work.

6. Results

6.1. The obtained sample: Twitter users according to home and work places

The first results are related to the registered population in each municipality and district in the study area based on the residence points found in the Twitter database. These data comprise the sample of residents that will be used to estimate the OD matrices. As already indicated, the authors based their calculations on a total of just over 20,744 users, for whom they were able to identify the municipality or district of both home and workplaces. These users represent almost 0.7% of the population of the metropolitan area. The largest number of residents are in the municipality of Madrid, especially in the central districts (dynamic zones, inhabited mainly by young people), while the percentages are lower in metropolitan area municipalities. Outside Madrid City, the northern and southern most-populated municipalities stand out as bedroom cities. Municipalities farther from Madrid City tend to have a small, aged population. These demographic characteristics are well-reflected in Twitter quantity data (Fig. 4 and Table 3).

Regarding the number of workplaces identified with Twitter, the sample of 20,744 users represents a percentage of total employment of 0.7%. Once again, the central districts of the municipality of Madrid, as they are important commercial, touristic and offices areas, are better-represented. Another prominent zone is the Madrid airport. Outside Madrid City, main work municipalities are the industrial-oriented municipalities from eastern municipalities, and the southern belt. Again, remote municipalities tend to have few users (Fig. 4 and Table 4).

6.2. OD matrices based on Twitter

Fig. 5 represents the tangle of travel flows from the OD matrix between municipalities and districts in the metropolitan area, based on origin data aggregation by Census population data. There is a clear predominance of centripetal flows, with their origin in peripheral districts and metropolitan municipalities and their destination in the Central Districts. Among the periphery-periphery trips, we can see a prevalence of connexions between some municipalities in the metropolitan area, for example, between the large industrial and logistical municipalities in the Metropolitan South and those in the East. Clearly, the intensity of the relationships diminishes in the more outlying

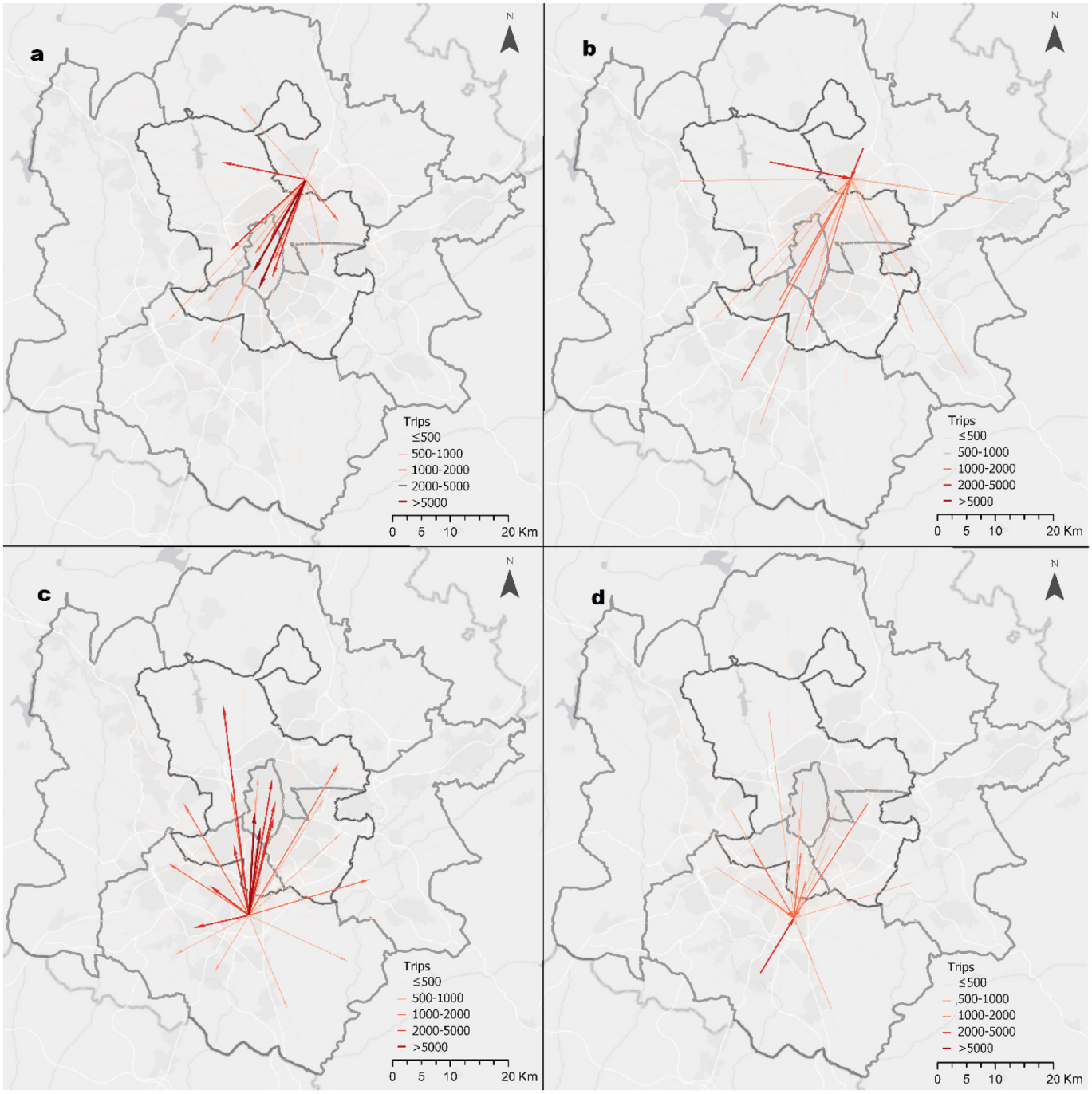

Fig. 6. Generations and attractions of Alcobendas and Getafe (a: generations of Alcobendas, b: attractions of Alcobendas, c: generations of Getafe, d: attractions of Getafe).

municipalities of the metropolitan area.

Another matrix was expanded through Security Social worker data in destination areas. As explained later, the correlation with transport base matrix is low, so it was ruled out because of its poor precision.

From this tangle of relationships, we can obtain information about generations and attractions for each district or municipality. Fig. 6 shows, for example, origin and destination trips in two spaces in the south and the north of the metropolitan area. These are two spaces of pronounced residential nature, but in recent years they have received activities and companies because of decentralising processes. Both the generations and, above all, the attractions of the municipality of Alcobendas in the Metropolitan North, a tertiary space with high incomes, are linked with the spaces around it and with the municipality of Madrid and the Metropolitan East and South zones, while the municipality of Getafe in the south of Madrid is shown as a bedroom area, generating trips mainly to other southern municipalities and Madrid City.

The aggregated matrix for large metropolitan zones helps to view these relationships between the different zones of study area more easily. If we simplify the flows by metropolitan zones we can clearly see the predominance, on one hand, of intrazonal trips, evidence of the growing relationships between municipalities and nearby districts, and, on the other, the attractions of the Central Districts (Fig. 7 and Table 5).

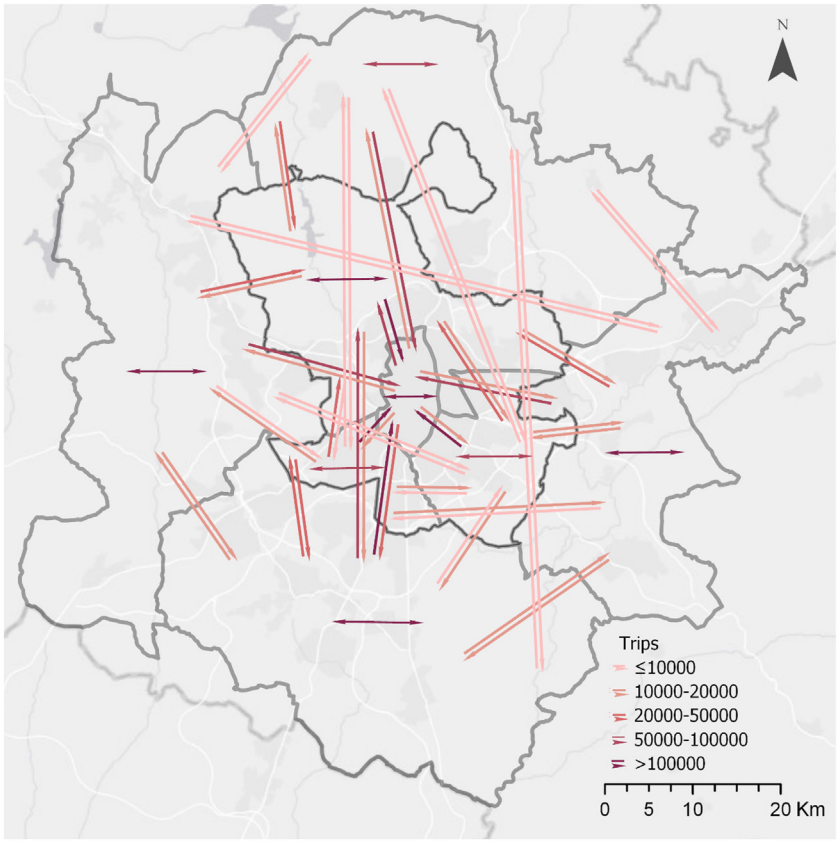

Fig. 7. Travel-flow matrix based on Madrid metropolitan zones Twitter data.

Table 5
Number of travel flows based on Twitter data in Madrid metropolitan zones.

	Central Districts	Northern Districts	Southwestern Districts	Southeastern Districts	Southern Municipalities	Eastern Municipalities	Northern Municipalities	Western Municipalities	Total
Totals									
Central D.	339,149	72,842	15,094	13,354	22,596	15,624	10,097	15,678	504,434
Northern D.	184,937	126,909	5513	13,083	12,649	11,981	14,465	12,492	382,029
Southwestern D.	172,575	39,938	85,021	12,286	36,571	11,778	8372	14,519	381,060
Southeastern D.	138,181	43,237	7702	99,248	13,048	15,177	6520	8455	331,568
Southern M.	169,942	62,502	20,908	9724	364,165	14,363	7436	19,685	668,725
Eastern M.	74,446	35,679	4850	10,823	14,405	174,004	8460	7422	330,089
Northern M.	71,320	27,988	4677	3920	8883	8130	66,454	5315	196,687
Western M.	76,491	43,342	4936	5923	12,654	5544	4382	118,785	272,057
Totals	1,227,041	452,437	148,701	168,361	484,971	256,601	126,186	202,351	3,066,649
Percentages									
Central D.	67.23	14.44	2.99	2.65	4.48	3.10	2.00	3.11	100.00
Northern D.	48.41	33.22	1.44	3.42	3.31	3.14	3.79	3.27	100.00
Southwestern D.	45.29	10.48	22.31	3.22	9.60	3.09	2.20	3.81	100.00
Southeastern D.	41.68	13.04	2.32	29.93	3.94	4.58	1.97	2.55	100.00
Southern M.	25.41	9.35	3.13	1.45	54.46	2.15	1.11	2.94	100.00
Eastern M.	22.55	10.81	1.47	3.28	4.36	52.71	2.56	2.25	100.00
Northern M.	36.26	14.23	2.38	1.99	4.52	4.13	33.79	2.70	100.00
Western M.	28.12	15.93	1.81	2.18	4.65	2.04	1.61	43.66	100.00
Total	40.01	14.75	4.85	5.49	15.81	8.37	4.11	6.60	100.00

6.3. Quality of the obtained data

The coefficient of determination was obtained between the number of resident users according to Twitter and official population data, so the sample quality can be assessed. The r^2 value obtained is relatively high, at 0.72.[2] In the same way, the r^2 value, the result of correlating Twitter workplaces with Security Social workers' data, scores at 0.70. In this sense, the sample values are relatively well-matched to the residents and workers' distribution in the study area, except in the case of the central districts, where the sample is much greater that would be expected, given their resident population. This discrepancy in central

[2] The Central District, which presents a volume of identified residents and workers using *Twitter* which is much higher than the total number of residents (*footnote continued*) and workers, has been eliminated. Thus, the correlation including this district falls to an r2 of 0.08 for residents, and 0.32 for workers.

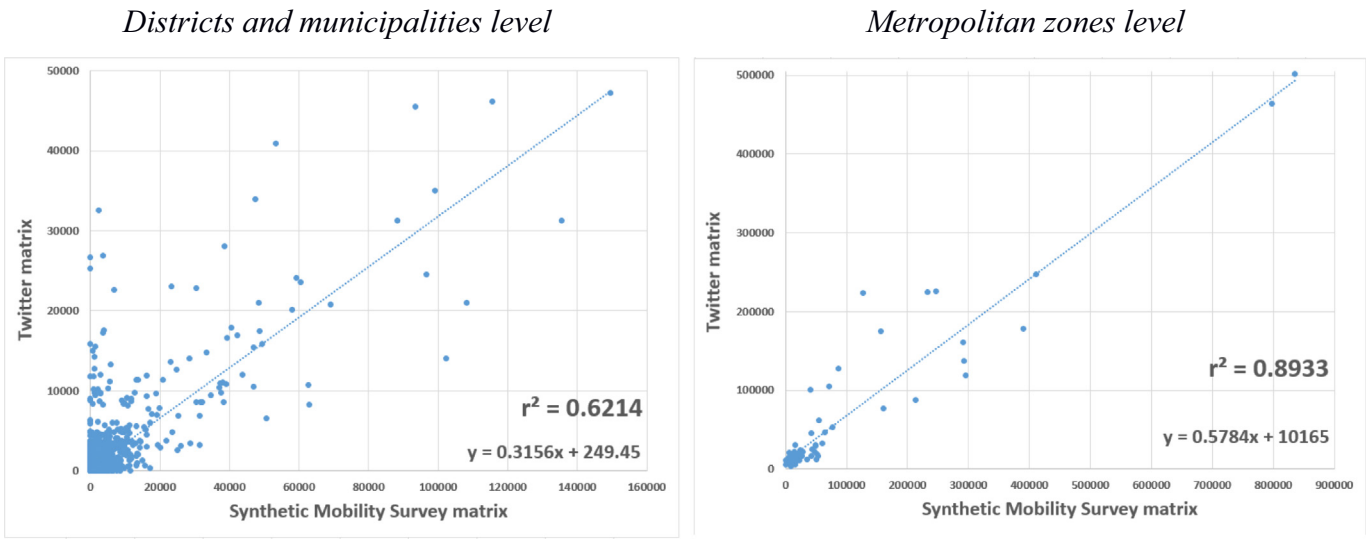

Fig. 8. Bivariate correlation between trip values from Twitter and trip values from the Synthetic Mobility Survey.

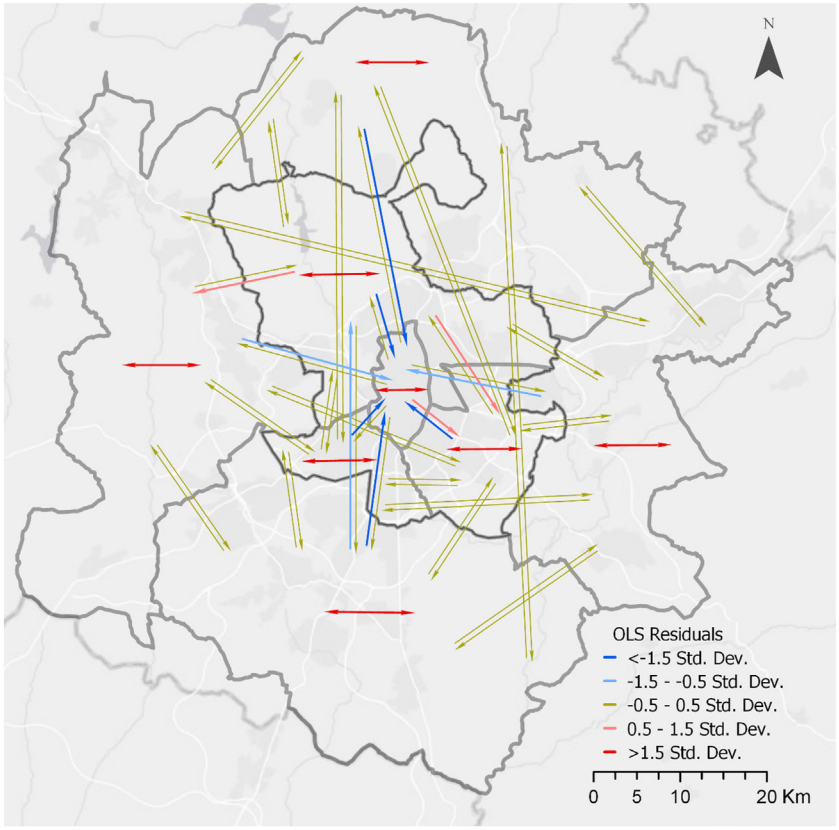

Fig. 9. Residuals map distribution from bivariate correlation between Twitter trip values and Synthetic Mobility Survey trip values (metropolitan-zone level).

spaces is because Twitter includes groups that are not considered in official statistics, such as unregistered foreigners and visitors not included in official data. Furthermore, Madrid citizens are very active at night, especially in the central districts. We should bear in mind that tourism is a major factor in these zones and it has a high level of nocturnal mobility, which makes it difficult to extract residential zones, as these are easily confused with nocturnal activity. However, the correlation values obtained indicate that the number of residents detected by Twitter is a good sample for studying the distribution of the population (see also García-Palomares et al., 2018).

If we study the relationship between the value of the flows detected in the matrix obtained using Twitter data and expanded with the Census using data from the Synthetic Mobility Survey, at municipal and district level, the r^2 coefficient is 0.62. However, when the matrix obtained using Twitter data is expanded by the destinations using the number of employees registered with the National Institute of Social Security, the match drops to 0.27. This difference may be due to a greater bias in access to social networks according to the type of job, and to the fact that the National Institute of Social Security does not register students. In any event, the results demonstrate the convenience of expanding the data based on origins and the distribution of the resident population.

When the matrices are aggregated according to large metropolitan

Table 6
Residuals matrix from bivariate correlation between Twitter trip values and Synthetic Mobility Survey trip values (metropolitan-zone level).

	Central Districts	N Districts	SW Districts	SE Districts	Southern Municipalities	Eastern Municipalities	Northern Municipalities	Western Municipalities
Central Districts	96,639 (1.70)	17,401 (0.31)	24,961 (0.44)	34,691 (0.61)	−22,745 (−0.40)	−11,514 (−0.20)	23,268 (0.41)	−3872 (−0.07)
Northern Districts	−145,650 (−2.56)	133,888 (2.35)	8504 (0.15)	33,230 (0.58)	−3027 (−0.05)	2714 (0.05)	21,425 (0.38)	32,881 (0.58)
Southwestern Districts	−105,099 (−1.85)	−10,426 (−0.18)	125,886 (2.21)	2884 (0.05)	−25,804 (−0.45)	−2690 (−0.05)	1833 (0.03)	−599 (−0.01)
Southeastern Districts	−124,342 (−2.19)	−5022 (−0.09)	15,356 (0.27)	93,558 (1.64)	5585 (0.10)	−3417 (−0.06)	11,512 (0.20)	−4742 (−0.08)
Southern Municipalities	−219,625 (−3.86)	−67,130 (−1.18)	8300 (0.15)	12,386 (0.22)	82,668 (1.45)	−6450 (−0.11)	9868 (0.17)	−3707 (−0.07)
Eastern Municipalities	−75,006 (−1.32)	−17,127 (−0.30)	5658 (0.10)	13,574 (0.24)	−15,349 (−0.27)	55,154 (0.97)	2602 (0.05)	8141 (0.14)
Northern Municipalities	−98,992 (−1.74)	−50 (0.00)	41 (0.00)	9667 (0.17)	−8515 (−0.15)	166 (0.00)	82,760 (1.45)	4023 (0.07)
Western Municipalities	−64,499 (−1.13)	−20,805 (−0.37)	9663 (0.17)	5389 (0.09)	4264 (0.07)	6248 (0.11)	8137 (0.14)	51,278 (0.90)

*Standard deviations in parenthesis.

zones and the results are compared once again with those obtained in the Synthetic Mobility Survey transport matrix, the quality of the results increases significantly, to give an r^2 match of 0.89 for resident users (Fig. 8). It is then is possible to map the residuals (Fig. 9) in such a way that we can see in which relationships the greatest deviations occur, whether due to overestimation (more trips by Twitter data in contrast with Synthetic Mobility Survey) or underestimation. In general, the results show how the Twitter matrix tends to overestimate the Central Districts destination flows and, by contrast, to underestimate the attractions of the internal trips (Table 6).

7. Conclusions

The emergence of geotagged Big Data and the enormous possibilities it offers are revolutionising studies and professional practice in transport planning. One of the spheres in which the use of the new data is having the greatest impact is mobility, especially for obtaining OD travel matrices. Most works have been based on the use of mobile phone data (the so-called CDRs) (see Chen et al., 2016), which enable the collection of enormous population samples with high temporal resolution. However, accessing phone data is costly and is always dependent on telephone companies.

In this article, we worked with Twitter data to obtain travel matrices as an alternative to mobile phone data. Obviously, the samples used are much less representative than those available using phone data, thereby increasing errors. However, when Twitter data are downloaded via streaming, access is free, and this enables many institutions to gain an initial overview of mobility in their area.

As with CDRs, Twitter data need to be pre-processed and processed before mobility matrices are obtained. In this article we have tried to show the steps that must be followed when working with Twitter data and to adopt the necessary measures to enable us to obtain more reliable information regarding mobility. One of these improvements is enriching Twitter data by using information on land uses, by working with high spatial detail layers such as Land Registry. In this way, it is possible to enhance the definition of the origins from residential spaces and the destinations in activity spaces.

Based on Twitter data, the results obtained provide an approximation to the dynamic of work-related mobility inside the Madrid Metropolitan Area and enable us to identify the main zones that generate and attract trips, and the intensity of the flows between them. In addition to its low cost, Twitter has the advantage of being constantly updated, in such a way that once a sufficient volume of tweets has been downloaded and the first matrices have been obtained, the data continue to update themselves permanently, thereby enabling us to identify tendencies and changes over time, the effect of transport and mobility policies, measures adopted with regard to transport networks and the impact of certain events or incidents in the city.

To verify the results, the matrices obtained were compared with the official data of the Madrid Transport Consortium, generated by a Synthetic Mobility Survey carried out in 2014. This validation process demonstrated that the results are acceptable when working at municipal and district level in the metropolitan area, and very good when aggregated into large zones. This validation also served to contrast different sources when it comes to expanding the data, with the validation based on the expansion of the origins of trips using resident population data, offering a better match.

Nevertheless, some problems were detected in the results and it is important to consider them. The greatest problems appear to be related to the central spaces, such as the case of Central Districts in Madrid and, to a lesser extent, some adjacent districts. Finetuning the land use data from Land Registry and enhanced data enrichment could help to improve results in future research. In turn, different time slots can be applied to employees in this study when it comes to determining home and work activities. Data bias (predominant Twitter use by 20–39 years young people, less available data in low-income areas) and a lack of sample representativeness are limitations to be overcome (Rashidi et al., 2017).

In any case, some of the problems detected can be rectified over time by increasing the size of the samples used. Thus, the volume of data compiled increases if the data continue to be downloaded via streaming and the matrices are constantly updated with a larger number of tweets. The possibility of being able to frequently update the matrices suggests that this type of information is extremely applicable when it comes to efficient planning of urban mobility and the transport systems that serve it. At the same time, the availability of much larger data samples will enable new research in the future, such as the construction of travel matrices according to time slots or monitoring of different social groups' mobility.

Acknowledgments

The authors gratefully acknowledge funding from the Spanish Ministerio de Educación, Cultura y Deporte (Program FPUAP2015-0147), the Spanish Ministerio de Economía, Industria y Competitividad (MINECO) and European Regional Development Fund (ERDF) (Project TRA2015-65283-R) and the Madrid Regional Government (SOCIALBIGDATA-CM, S2015/HUM-3427).

References

Ahas, R., Silm, S., Järv, O., Saluveer, E., & Tiru, M. (2010). Using mobile positioning data to model locations meaningful to users of mobile phones. *Journal of Urban Technology, 17*(1), 3–27. https://doi.org/10.1080/10630731003597306.

Alexander, L., Jiang, S., Murga, M., & González, M. C. (2015). Origin-destination trips by purpose and time of day inferred from mobile phone data. *Transportation Research Part C: Emerging Technologies, 58*, 240–250. https://doi.org/10.1016/j.trc.2015.02.018.

Bar-Gera, H. (2007). Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: A case study from Israel. *Transportation Research Part C: Emerging Technologies, 15*(6), 380–391. https://doi.org/10.1016/j.trc.2007.06.003.

Blanford, J. I., Huang, Z., Savelyev, A., & MacEachren, A. M. (2015). Geo-located tweets. Enhancing mobility maps and capturing cross-border movement. *PLoS One, 10*(6), 1–16. https://doi.org/10.1371/journal.pone.0129202.

Bonnel, P., Hombourger, E., Olteanu-Raimond, A.-M., & Smoreda, Z. (2015). Passive mobile phone dataset to construct origin-destination matrix: Potentials and limitations. *Transportation Research Procedia, 11*, 381–398. https://doi.org/10.1016/j.trpro.2015.12.032.

Caceres, N., Romero, L. M., Benitez, F. G., & del Castillo, J. M. (2012). Traffic flow estimation models using cellular phone data. *IEEE Transactions on Intelligent Transportation Systems, 13*(3), 1430–1441. https://doi.org/10.1109/TITS.2012.2189006.

Caceres, N., Wideberg, J. P., & Benitez, F. G. (2007). Deriving origin–destination data from a mobile phone network. *IET Intelligent Transport Systems, 1*(1), 15–26. https://doi.org/10.1049/iet-its:20060020.

Cao, G., Wang, S., Hwang, M., Padmanabhan, A., Zhang, Z., & Soltani, K. (2014). A scalable framework for spatiotemporal analysis of location-based social media data. *Computers, Environment and Urban Systems, 51*, 70–82. Retrieved from http://arxiv.org/abs/1409.2826.

Chen, C., Ma, J., Susilo, Y., Liu, Y., & Wang, M. (2016). The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transportation Research Part C: Emerging Technologies, 68*, 285–299. https://doi.org/10.1016/j.trc.2016.04.005.

Ciuccarelli, P., Lupi, G., & Simeone, L. (2014). *Visualizing the data city: Social media as a source of knowledge for urban planning and management.* Springerhttps://doi.org/10.1007/978-3-319-02195-9.

De Domenico, M., Lima, A., & Musolesi, M. (2013). Interdependence and predictability of human mobility and social interactions. *Pervasive and Mobile Computing, 9*(6), 798–807. https://doi.org/10.1016/j.pmcj.2013.07.008.

Gao, S., Yang, J., Yan, B., Hu, Y., Janowicz, K., & McKenzie, G. (2014). Detecting origin-destination mobility flows from geotagged tweets in greater Los Angeles area. *Eight international conference on geographic information science (GIScience'14)*. Retrieved from http://www.geog.ucsb.edu/~sgao/papers/2014_GIScience_EA_DetectingODTripsUsingGeoTweets.pdf.

García-Palomares, J. C., Salas-Olmedo, M. H., Moya-Gómez, B., Condeço-Melhorado, A., & Gutiérrez, J. (2018). City dynamics through twitter: Relationships between land use and spatiotemporal demographics. *Cities, 72*, 310–319. https://doi.org/10.1016/j.cities.2017.09.007.

Gutiérrez, J., & García-Palomares, J. C. (2007). New spatial patterns of mobility within the metropolitan area of Madrid: Towards more complex and dispersed flow networks. *Journal of Transport Geography, 15*(1), 18–30. https://doi.org/10.1016/j.jtrangeo.2006.01.002.

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., & Ratti, C. (2014). Geo-located twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science, 41*(3), 260–271. https://doi.org/10.1080/15230406.2014.890072.

Huang, Q., & Wong, D. W. S. (2016). Activity patterns, socioeconomic status and urban spatial structure: What can social media data tell us? *International Journal of Geographical Information Science, 30*(9), 1873–1898. https://doi.org/10.1080/13658816.2016.1145225.

Iqbal, M. S., Choudhury, C. F., Wang, P., & González, M. C. (2014). Development of origin–destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies, 40*, 63–74. https://doi.org/10.1016/j.trc.2014.01.002.

Järv, O., Tenkanen, H., & Toivonen, T. (2017). Enhancing spatial accuracy of mobile phone data using multi-temporal dasymetric interpolation. *International Journal of Geographical Information Science, 31*(8), 1630–1651. https://doi.org/10.1080/13658816.2017.1287369.

Jin, X., Long, Y., Sun, W., Lu, Y., Yang, X., & Tang, J. (2017). Evaluating cities' vitality and identifying ghost cities in China with emerging geographical data. *Cities, 63*, 98–109. https://doi.org/10.1016/j.cities.2017.01.002.

Kung, K. S., Greco, K., Sobolevsky, S., & Ratti, C. (2014). Exploring universal patterns in human home-work commuting from mobile phone data. *PLoS One, 9*(6), 1–15. https://doi.org/10.1371/journal.pone.0096180.

Lee, J. H., Gao, S., & Goulias, K. G. (2015). *Can twitter data be used to validate travel demand models? GEOTRANS report 2015-5-03, 1–27*.

Lenormand, M., Picornell, M., Cantú-Ros, O. G., Tugores, A., Louail, T., Herranz, R., & Ramasco, J. J. (2014). Cross-checking different sources of mobility information. *PLoS One, 9*(8), 1–10. https://doi.org/10.1371/journal.pone.0105184.

Longley, P. A., & Adnan, M. (2016). Geo-temporal twitter demographics. *International Journal of Geographical Information Science, 30*(2), 369–389. https://doi.org/10.1080/13658816.2015.1089441.

Louail, T., Lenormand, M., Picornell, M., García Cantú, O., Herranz, R., Frias-Martinez, E., & Barthelemy, M. (2015). Uncovering the spatial structure of mobility networks. *Nature Communications, 6*(1), 1–8. https://doi.org/10.1038/ncomms7007.

Luo, F., Cao, G., Mulligan, K., & Li, X. (2016). Explore spatiotemporal and demographic characteristics of human mobility via twitter: A case study of Chicago. *Applied Geography, 70*, 11–25. https://doi.org/10.1016/j.apgeog.2016.03.001.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity. *McKinsey Global Institute*, (June). Retrieved from http://www.mckinsey.com/Insights%0A/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_%0Ainnovation

Martí, P., Serrano-Estrada, L., & Nolasco-Cirugeda, A. (2017). Using locative social media and urban cartographies to identify and locate successful urban plazas. *Cities, 64*, 66–78. https://doi.org/10.1016/j.cities.2017.02.007.

Netto, V. M., Pinheiro, M., Meirelles, J. V., & Leite, H. (2015). *Digital footprints in the cityscape: Finding networks of segregation through big data. International conference on location-based social media data, (March)*. 1–15.

Osorio, J., & García-Palomares, J. C. (2017). Nuevas fuentes y retos para el estudio de la movilidad urbana. *Cuadernos Geográficos, 56*(3), 247–267. Retrieved from http://revistaseug.ugr.es/index.php/cuadgeo/article/view/5352/5858.

Pajević, F., & Shearmur, R. G. (2017). Catch me if you can: Workplace mobility and big data. *Journal of Urban Technology, 24*(3), 99–115. https://doi.org/10.1080/10630732.2017.1334855.

Perez, A. J., Dominguez, L. D., Rubiales, A. J., & Lotito, P. A. (2015). *Optimización de matrices origen-destino estimadas a partir de datos georeferenciados en redes sociales. 13º Simposio Argentino de Investigación Operativa.* 47–56.

Picornell, M., Ruiz, T., Lenormand, M., Ramasco, J. J., Dubernet, T., & Frías-Martínez, E. (2015). Exploring the potential of phone call data to characterize the relationship between social network and travel behavior. *Transportation, 42*(4), 647–668.

Rashidi, T. H., Abbasi, A., Maghrebi, M., Hasan, S., & Waller, T. S. (2017). Exploring the capacity of social media data for modelling travel behaviour: Opportunities and challenges. *Transportation Research Part C: Emerging Technologies, 75*, 197–211. https://doi.org/10.1016/j.trc.2016.12.008.

Salas-Olmedo, M. H., & Rojas Quezada, C. (2017). The use of public spaces in a medium-sized city: From twitter data to mobility patterns. *Journal of Maps, 13*(1), 40–45. https://doi.org/10.1080/17445647.2017.1305302.

Shelton, T., Poorthuis, A., & Zook, M. (2015). Social media and the city: Rethinking urban socio-spatial inequality using user-generated geographic information. *Landscape and Urban Planning, 142*, 198–211. https://doi.org/10.1016/j.landurbplan.2015.02.020.

Shen, Y., & Karimi, K. (2016). Urban function connectivity: Characterisation of functional urban streets with social media check-in data. *Cities, 55*, 9–21. https://doi.org/10.1016/j.cities.2016.03.013.

Simpson, R. (2017). Android overtakes windows for first time. *StatCounter Global Stats.*. Retrieved April 20, 2018, from http://gs.statcounter.com/press/android-overtakes-windows-for-first-time.

Stefanidis, A., Crooks, A., & Radzikowski, J. (2013). Harvesting ambient geospatial information from social media feeds. *GeoJournal, 78*(2), 319–338. https://doi.org/10.1007/s10708-011-9438-2.

Toole, J. L., Colak, S., Sturt, B., Alexander, L. P., Evsukoff, A., & González, M. C. (2015). The path most traveled: Travel demand estimation using big data resources. *Transportation Research Part C: Emerging Technologies, 58*, 162–177. https://doi.org/10.1016/j.trc.2015.04.022.

Wu, L., Zhi, Y., Sui, Z., & Liu, Y. (2014). Intra-urban human mobility and activity transition: Evidence from social media check-in data. *PLoS One, 9*(5), 1–13. https://doi.org/10.1371/journal.pone.0097010.

Yin, J., Soliman, A., Yin, D., & Wang, S. (2017). Depicting urban boundaries from a mobility network of spatial interactions: A case study of Great Britain with geo-located twitter data. *International Journal of Geographical Information Science, 31*(7), 1293–1313. https://doi.org/10.1080/13658816.2017.1282615.

Zhang, P., Zhou, J., & Zhang, T. (2017). Quantifying and visualizing jobs-housing balance with big data: A case study of Shanghai. *Cities, 66*, 10–22. https://doi.org/10.1016/j.cities.2017.03.004.

Zhen, F., Cao, Y., Qin, X., & Wang, B. (2017). Delineation of an urban agglomeration boundary based on Sina Weibo microblog 'check-in' data: A case study of the Yangtze River Delta. *Cities, 60*, 180–191. https://doi.org/10.1016/j.cities.2016.08.014.