Cities 72 (2018) 310–319

Contents lists available at ScienceDirect

Cities

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

City dynamics through Twitter: Relationships between land use and spatiotemporal demographics

Juan Carlos García-Palomares*, María Henar Salas-Olmedo, Borja Moya-Gómez, Ana Condeço-Melhorado, Javier Gutiérrez

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

ARTICLE INFO

Keywords:
Social media
Twitter
Spatiotemporal demographics
Land use
Urban geography

ABSTRACT

Social network data offer interesting opportunities in urban studies. In this study, we used Twitter data to analyse city dynamics over the course of the day. Users of this social network were grouped according to city zone and time slot in order to analyse the daily dynamics of the city and the relationship between this and land use. First, daytime activity in each zone was compared with activity at night in order to determine which zones showed increased activity in each of the time slots. Then, typical Twitter activity profiles were obtained based on the predominant land use in each zone, indicating how land uses linked to activities were activated during the day, but at different rates depending on the type of land use. Lastly, a multiple regression analysis was performed to determine the influence of the different land uses on each of the major time slots (morning, afternoon, evening and night) through their changing coefficients. Activity tended to decrease throughout the day for most land uses (e.g. offices, education, health and transport), but remained constant in parks and increased in retail and residential zones. Our results show that social network data can be used to improve our understanding of the link between land use and urban dynamics.

1. Introduction

Internet users are no longer mere passive recipients of information, but have become producers of vast amounts of data, particularly through social networks. Many social media apps (Twitter, Foursquare and Facebook are three common examples) have geo-location features that (optionally) includes the ability to attach locational information in the form of coordinates provided by the units GPS or place names provided by the user, thereby enabling individual users to leave a digital 'geographic footprint' of their movement when posting a message (Blanford, Huang, Savelyev, & MacEachren, 2015). These digital data shadows are intimately intermingled with offline, material geographies of everyday life (Jin et al., 2017; Shelton, Poorthuis, & Zook, 2015). High spatial and temporal social media data reveals activity patterns information moving beyond the night-time residential geographies of conventional geodemographic data sources (Longley, Adnan, & Lansley, 2015). Social media data have been used to delineate the boundary of urban agglomeration based on people's activity (Zhen, Cao, Qin, & Wang, 2017), and can help to portray urban structure and related socioeconomic performance (Martí, Serrano-Estrada, & Nolasco-Cirugeda, 2017; Shen & Karimi, 2016; Zhang, Zhou, & Zhang, 2017).

Twitter has become an important sensor of the interactions between individuals and their environment (Frias-Martinez & Frias-Martinez, 2014). Twitter data contain precise spatial and temporal information, which can be used to perform spatiotemporal demographic analyses. It has been observed that residential zones on the periphery of cities generate more tweets in the evening, when people have returned to their homes, whereas areas of activity in the city centre are especially active during the day, when people visit them to undertake activities such as work or shopping (Ciuccarelli, Lupi, & Simeone, 2014). The spatiotemporal pattern of tweets posted from different zones in the city indicates the existence of the inherent linkage between the human urban activity pattern and the underlying land use structure (Zhan, Ukkusuri, & Zhu, 2014).

The aim of this study was to use information from the social network Twitter in order to analyse city dynamics throughout the day and the relationship between this and land use. We estimated the number of unique active users in each zone of the city of Madrid throughout the day, as a proxy for the changing location of the population. Then, we analysed Twitter activity (active users) throughout the day according to land use, yielding typical activity profiles for each land use. Lastly, we calculated multiple regression models (OLS) for each time slot to

* Corresponding author.
E-mail addresses: jcgarcia@ghis.ucm.es (J.C. García-Palomares), msalas01@ucm.es (M.H. Salas-Olmedo), bmoyagomez@ucm.es (B. Moya-Gómez), acondeco@ghis.ucm.es (A. Condeço-Melhorado), javiergutierrez@ghis.ucm.es (J. Gutiérrez).

http://dx.doi.org/10.1016/j.cities.2017.09.007
Received 13 June 2017; Received in revised form 5 September 2017; Accepted 16 September 2017
Available online 28 September 2017

explain Twitter activity based on square metres of each land use in each zone of the city. An analysis of the coefficients of the independent variables should indicate the changing influence of the different land uses according to time slots. These coefficients were used to predict activity patterns in a future urban development in Madrid according to a range of scenarios of land use mix and density. The underlying hypothesis was that if each land use has a specific influence on Twitter activity, then the built-up area of each land use should satisfactorily explain the distribution of tweeters in the city in each time slot. This approach provides useful information for urban planning, not only because it sheds light on urban dynamics in relation to land use, but also because it can be used as a tool in the planning process when it comes to evaluating future urban developments.

This paper contributes to the existing literature in several ways. Firstly, works related to Twitter use and land use attempt to infer land use on the basis of temporal profiles in Twitter use. This, in reality, is of limited use given that there are techniques (remote sensing) which are more reliable and easier to apply in order to obtain land use maps. In our case, the objective is not to infer land use, but to learn about the daily dynamics of the city with regard to land use. Secondly, we used explanatory OLS models which linked land use and Twitter activity. To the best of our knowledge, previous works have been limited to performing descriptive analyses. These explanatory models are of great interest for urban planning. On the basis of predicted land use in new urban developments it is possible to obtain a proxy of the activity (demand) at different times of day in those future new developments. Thirdly, in contrast to previous studies, the land use data in this paper do not come from land use maps but from the Cadastre. Cadastral data provide precise information on the mix of uses of built-up land in each spatial unit and not simply the predominant land use. This is of great importance for adjusting the explanatory models. In contrast with other previous works, in this paper we have worked with a detailed classification of land use, which is useful for urban planning. Fourthly, another distinguishing feature of our study is that the unit of analysis is the Twitter user rather than the geotagged tweet. We focused not on the spatiotemporal distribution of tweets but on that of Twitter users, as a proxy for the changing location of the population throughout the day. This approach mitigates the bias arising from varying intensities of Twitter use, which gives more weight to users who tweet more often, and above all, it provides useful information for the provision of public sector services and private sector business activities.

The remainder of this paper is structured as follows. Section 2 presents a literature review on the use of Twitter data in urban studies, especially in relation to land uses. Section 3 reports the methodology and data. Section 4 contains an analysis of the results and Section 5 presents the main conclusions.

2. Human activities and land uses: spatiotemporal patterns of digital footprints in the city

Social network users generate a vast amount of data. Worldwide, 500 million Twitter messages are posted every day, there are 7000 million check-ins on Foursquare, and > 80 million photos are uploaded on Instagram. Most studies based on social network data have used Twitter (Murthy, 2013), due in large part to the fact that the data (tweets) can be freely downloaded by connecting to the Twitter Streaming API. This real-time, public, and cost-free data stream covers only around 1–2% of the whole stream of tweets (Andrienko et al., 2013). Some of the tweets are geotagged. These facilitate profiling of usage across space as well as time and have been found to be a useful tool for urban research (Lansley & Longley, 2016).

Within a city from the centre to the periphery, the densities decrease in general, so tweet densities would be a good surrogate of population densities (Jiang, Ma, Yin, & Sandberg, 2016). Maps of the spatial distribution of tweets can be enriched with user IDs, tweet content and language. Thus, Longley et al. (2015) inferred the gender, age and ethnicity of tweeters from the user ID (username field) and analysed the spatial distribution of the tweets posted by each of the groups identified. Lansley and Longley (2016) and Andrienko et al. (2013) focused on tweet content in order to examine frequently tweeted words and their spatiotemporal patterns. Meanwhile, Mocanu et al. (2013) used the language of tweets to identify linguistically specific urban communities. It is also possible to analyse the degree of social mixing in the use of space, tracking the movement of demographic groups within cities (Netto, Pinheiro, Meirelles, & Leite, 2015; Shelton et al., 2015). In contrast to the information provided by official sources, which offer data on place of residence, these studies on multiculturalism and social mixing analyse the use of space throughout the day.

Through spatiotemporal monitoring of the tweets posted by each user, it is possible to deduce population mobility patterns (Blanford et al., 2015; Longley & Adnan, 2016; Wu, Zhi, Sui, & Liu, 2014), identify demographic groups based on user names (Luo, Cao, Mulligan, & Li, 2016) and obtain origin-destination matrices (Gao et al., 2014). The reliability of Twitter data in mobility studies has been validated by Lenormand et al. (2014), who compared Twitter data with mobile phone records and official data (census) and concluded that the three sources offer comparable results.

The spatial distribution of tweets in the urban core changes over the course of the day, reflecting changes in the location of the population. Areas of the city with similar time profiles can be grouped using clustering algorithms in order to identify clusters of common tweeting activity. Thus, Frias-Martinez, Soto, Hohwald, and Frias-Martinez (2012) and Frias-Martinez and Frias-Martinez (2014) identified four clusters with specific tweeting activity signatures that basically corresponded to the following types of land use: business, leisure/weekend, nightlife and residential. Using a similar methodology, Zhan et al. (2014) inferred four broad categories of land use based on a spatiotemporal analysis of Twitter data: residential, retail, open space/recreation and transportation/utility. A similar approach has been adopted in other studies (e.g., Pei et al., 2014; Ríos & Muñoz, 2017), in which clustering algorithms have been used to group city zones with similar profiles according to mobile phone activity, or Chen et al. (2017) in China using the social media "*Tencent*". These methodologies based on clustering the tweeting activity profiles of different areas in the city could be used as an alternative to satellite imagery, to infer urban land uses based on social network data. However, the land use categories obtained are very generic, and their usefulness for urban planning is therefore limited.

3. Data and methodology

3.1. Twitter data pre-processing

One of the most widespread social networks is undoubtedly Twitter, a platform for posting messages with a maximum of 140 characters, known as tweets. The service has > 270 million active users around the world. Roughly 80% of active Twitter users access the service via a mobile telephone (Lansley & Longley, 2016). Since 2010, Twitter provided users with the ability to include their location either by attaching coordinates or a place name while tweeting, therefore making it possible to locate tweets geographically both in space and over time (Blanford et al., 2015). Tweets are automatically geotagged by the GPS of mobile devices, provided that the user has enabled this function. Geotagged Tweets account for approximately just 1% of all the messages that are sent using the Twitter service (Blanford et al., 2015).

Twitter data were downloaded and pre-processed as follows:

a) Data download

The data used for this study was downloaded via the Twitter Streaming API over two consecutive years (from January 2012 to December 2013). Only geotagged tweets were downloaded, selecting those that covered the municipality of Madrid. Besides the coordinates,

each tweet also carries information about the user ID, the date and time the tweet was posted, the device language setting, device type and the text of the message. We downloaded over 6.8 billion tweets in the city of Madrid. The data were loaded into a GIS (ArcGIS 10.4), creating a layer of points with the coordinates *x y* of the position from which each of them was posted. Subsequently, we selected tweets that had been posted on typical workdays (Tuesday, Wednesday and Thursday), obtaining a total of 3.07 million tweets.

b) Spatial and temporal aggregation of data to obtain the number of unique active users

The same user often posts several tweets from the same location at the same time. The number of such tweets can be extremely high with some users, leading to an overestimation of the presence of this type of user at these locations and times. It is therefore necessary to analyse unique users rather than tweets. To this end, the tweets were aggregated spatially and temporally (every quarter of an hour) depending on the user ID, to obtain the presence of unique active users in each spatial unit rather than the number of tweets posted.

Spatial aggregation of Twitter data was based on the transport zones established by the Madrid transport authority. These zones were designed to measure the daily dynamics of the population (mobility of the population from a spatiotemporal perspective) by differentiating between zones which generate journeys (those where the population is concentrated during the night) and zones which attract journeys (those where the population is concentrated during the day), which makes them suitable for the objectives of the study. Although most transport zones present some degree of mixed land use, in general they form relatively homogeneous spatial units from the point of view of land use, making it possible to recognize important elements of the city (for example, big parks or shopping centres). In addition, these zones are large enough to register a significant number of tweets. The possibility of using cells or hexagons (Shelton et al., 2015) was ruled out. Cells undoubtedly have the advantage of mitigating the problem of modified spatial units (Openshaw, 1984) since they form spatial units of the same size and shape; however, such units are more heterogeneous from the point of view of land use, and the aim of our research was to relate the number of Twitter users to land use.

The temporal sequence obtained of the presence of unique active users in each transport zone every quarter of an hour can be viewed in the attached Video 1, which shows how hot spots shifted from the centre, in the morning and the afternoon, towards the periphery, at night.

c) Normalisation of data and calculation of the number of users per transport zone and time of day

One of biases presented by Twitter when using temporal information is the different use made of this social network over the course of the day. Twitter is used more in the evening than in the morning or at mid-day (Fig. 1). The highest peak in users occurred at 10 pm (over 7%), whereas this figure fell to < 4% at 8–9 am.

To eliminate this bias, daily distribution data were grouped by time slot and the distributions were normalised, so that the total number of unique active users per time slot was equalised to 100,000, using the following formula:

$$T_{zhn} = \frac{T_{zh}^{*}}{T_h} 100000,$$

where T_{zhn} is the normalised number of unique active users in zone z at time slot h; T_{zh} is the number of unique active users in zone z at time slot h and T_h is the total number of active tweeters at time slot h.

Subsequently, the data were aggregated into four time slots to facilitate analysis: a) morning (07:00 to 12:59), b) afternoon (13:00 to 17:59), c) evening (18:00 to 22:00) and d) night (22:00 to 23:59).[1]

3.2. Analysis of spatial distribution patterns

Initially, the spatial distribution patterns were analysed by mapping unique active users according to large time slots. Descriptive statistics were obtained for these distributions. In this case, since the data were normalised, the mean distributions were similar in all cases, but their standard deviations showed the degree of concentration or dispersion of Twitter users. High standard deviation values are associated with more concentrated spatial patterns (users are concentrated more in some zones and less in others). In contrast, low standard deviation values indicate user dispersion (more uniform distribution throughout the zones).

To compare the distributions according to time slot, we conducted a bivariate Ordinary Least Squares (OLS) analysis. ArcGIS10.4 software was used to obtain both the coefficient of determination and the spatial distribution of the residuals. The coefficient of determination indicates the degree of relationship between two variables (the overlap between data distributions), while an analysis of normalised residual plots identifies differences between the distributions. Although we analysed all the relationships between the four major time slots, we focused primarily on comparing night-time distribution (as a baseline reference) with each of the other time slots.

3.3. Temporal profiles according to land use

Temporal profiles according to land use were calculated using the normalised distributions of unique active users according to zone every quarter of an hour and the predominant land use in each zone. The total number of unique active users every quarter of an hour in a zone was assigned to the predominant land use, and then the total number of users according to land use was summed for each quarter of an hour. This approach yielded the temporal distribution of unique active users for each land use.

To obtain temporal profiles of unique active Twitter users by land use, each transport zone was characterised in terms of the percentage of built-up area pertaining to each land use, based on cadastral data.[2] Cadastral data provides information on the amount of square metres of building floor space for each type of land use in each building in the city. Initially, in order to obtain the temporal profiles, three main types of zone were distinguished: residential (when > 66.6% of built surface in the zone was residential), activity (when more than the 66.6% was non-residential, e.g. offices, industry, retail or education) and mixed residential (all other cases). In addition, the areas of activity were classified in 9 types: offices, industry, retail, health, education, culture, large transport terminals, parks and others. Fig. 2 shows the predominant land use in each of the zones.

[1] The assumption that the normalised number of unique active users reflected the location of the population should be contrasted with the ground truth. Official data record the location of the population according to place of residence (night-time residential geographies), but do not provide information on the location of the population over the course of the day. Therefore, a comparison with the ground truth can only be performed with night-time data (Lin & Cromley, 2015). As in previous studies (e.g. Luo et al., 2016; Salas-Olmedo and Rojas, 2017), we defined “home” as the place most frequently visited by a user at night-time. Thus, we included users who tweeted between 22:00 and 24:00 from residential buildings. In order to verify if the number of the detected residents out of the Twitter users in each transport zone adequately reflected the distribution of the population according to official data (Register of Inhabitants 2013, Spanish National Statistics Institute), we calculated the correlation coefficient between the two variables, obtaining an r2 value of 0.46, indicating that the number of resident tweeters is a good proxy for the distribution of the population. The difference between both sources can be partly explained by the fact that Twitter includes groups that are not accounted in official statistics, such as tourists and no registered foreigners. Furthermore, Madrid citizens' enjoy doing leisure activities (i.e. restaurants, bars) during the nigh time, especially in the city centre, In addition, night workers who tweet from their workplace between 20:00 and 24:00 may be mistaken for residents.

[2] http://www.catastro.meh.es/esp/acceso_infocat.asp.

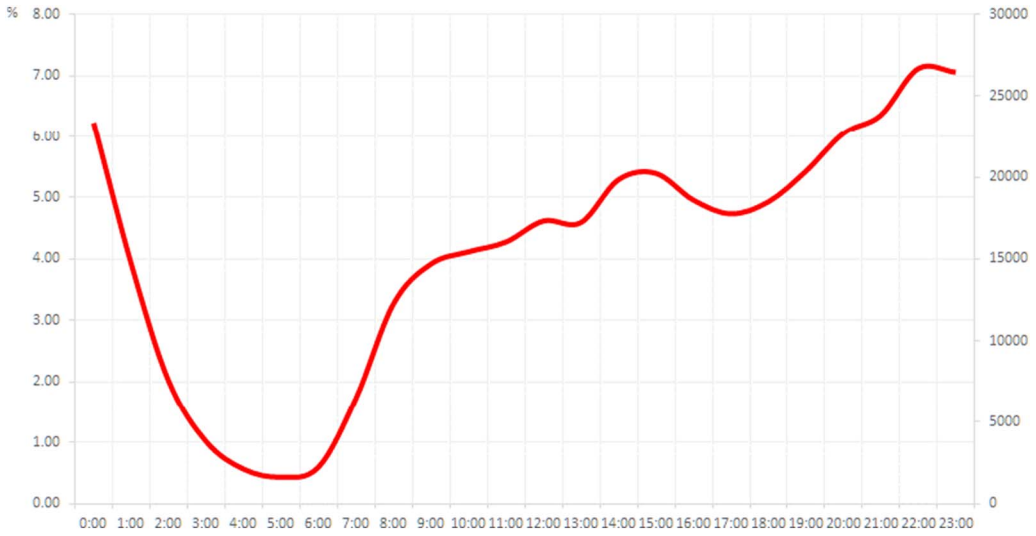

Fig. 1. Distribution of the number of Twitter users in Madrid according to time slot.

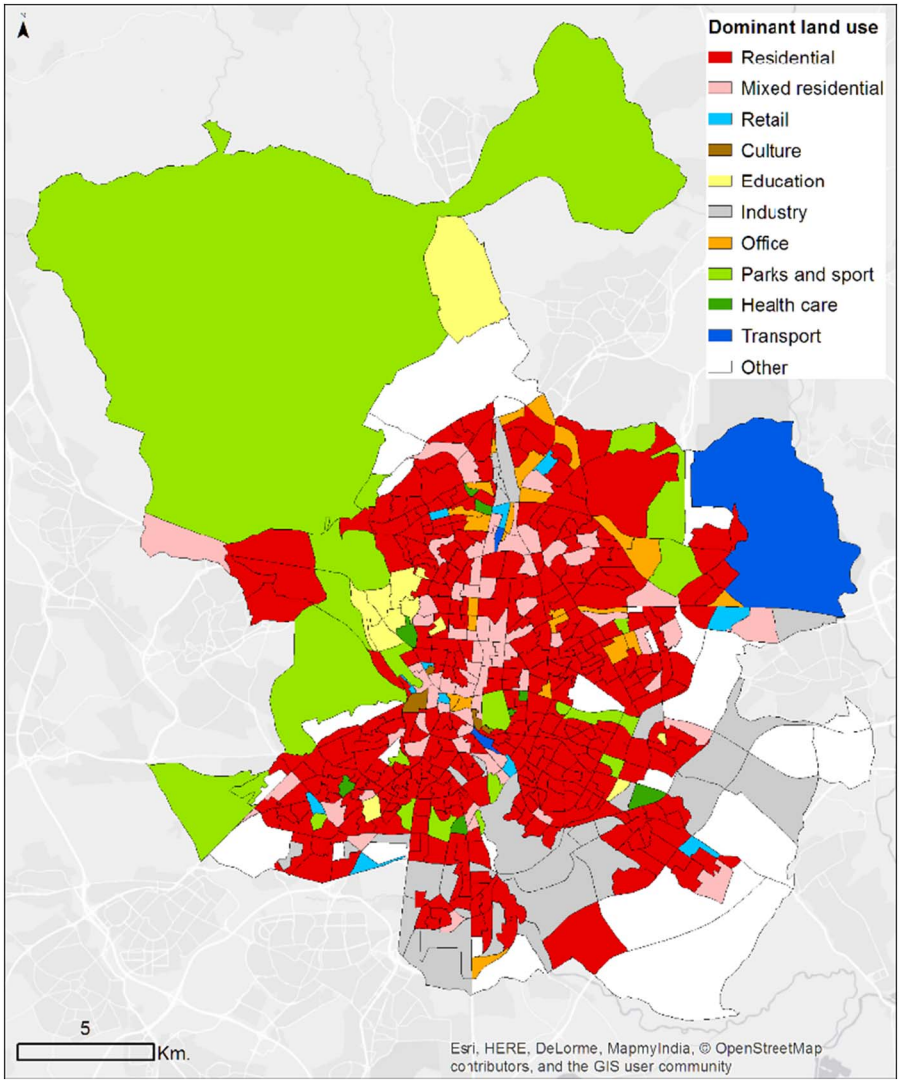

Fig. 2. Predominant land uses according to transport zone.

3.4. Ordinary Least Squares (OLS) models

Lastly, the relationship between Twitter user distribution and land use was analysed using four multiple regression (OLS) models. The dependent variables were the normalised distributions of unique active users according to zone in each of the major time slots (Morning, Afternoon, Evening and Night). The explanatory variables were the amount of building floor area square metres of each type of land use in

each of the zones in the city, based on cadastral data.

Given the marked centre-periphery drop in the intensity of Twitter use (see Section 4.1), distance to the city centre was included in the models as a control variable.

The OLS models were calculated in two steps. In the first step, we included all the independent variables considered initially (land uses and distance to the centre). In the second step, we eliminated non-significant variables and calculated the models again. The results shown in Table 4 correspond to the second step, and the variables that proved non-significant in the first step and were discarded have been left blank.

3.5. Application of the OLS model to future urban development

As an example of the use of the models obtained in urban planning tasks, it is proposed that they be applied to estimating daily activity in a future urban development by comparing four scenarios. "*Madrid Nuevo Norte*",[3] a new business district situated 10.8 km from the city centre, is used as an example. The reference scenario is the Madrid City Council's proposal for this future urban development. The reference scenario is compared to three fictitious scenarios, in which the density and mix of land use varies. In order to generate the fictitious scenarios, the characteristics of the urban developments of *La Défense* in Paris and *Potsdamer Platz* in Berlin have been applied. Specifically, the characteristics of the four scenarios proposed are (Table 1):

- The *base scenario* is the Madrid City Council's proposal. The proposal has a gross development potential (m^2 which can be built on/m^2 of land) of 1.05 m^2. The objective is a space which is relatively balanced in terms of land use, with 40% residential land, 34% offices and the rest devoted to other uses (commerce, education, health, parks, etc.).
- The second scenario ("*Madrid Nuevo Norte-Office*") has the same development potential proposed in the Council's plan, but with a greater specialisation in offices, as the residential land in the initial proposal is considered as offices.
- In the third scenario ("*Potsdamer Platz*" *Scenario*) the distribution of land use in the base scenario remains the same but the development potential is reduced to 0.55 m^2. The "*Potsdamer Platz*" development in Berlin was taken as a reference, having a mix of land use similar to the Madrid City Council proposal but a much lower density.
- Finally, the "*La Défense*" scenario proposes high density and a low mix of land use. La Défense, in Paris, is taken as a reference. The proposal involves increasing the development potential to 2.25 m^2, and changing the residential, educational and health care land to offices.

4. Results

4.1. Distribution of tweeters according to time slot

Twitter activity in the different city zones varied substantially according to time slot. For the purpose of conducting comparisons, nighttime was taken as the reference scenario (Fig. 3). The differences in distribution according to time slot can be analysed using bivariate correlations (Table 2). As is to be expected in a place with a wide variety of land uses such as Madrid, the coefficients of determination between the time distributions were relatively high in all cases, especially between successive time slots in the day. Taking the night as the reference (in bold in Table 2), the greatest differences (lowest correlations) were observed between night and morning (homes vs activities). In contrast, the correlations were highest between night and afternoon, and especially between night and evening (when many people have returned home after a day's work).

The residual plots of the correlations between night and the rest of the time slots show the different behaviour of residential spaces and areas of activity throughout the day (Fig. 4). Comparing night and morning, areas of activity were more active in the latter slot (positive residuals in red), especially in office and mixed zones in the centre (with a high presence of offices) and in large complexes (e.g. university campuses and hospitals) and transport terminals (airport and railway stations). In contrast, residential areas presented negative residuals (in blue). Comparing night and evening, many areas of activity ceased activity while leisure centres and shopping areas became more active.

The descriptive statistics show that the lowest standard deviation occurred at night, indicating that active users were distributed throughout all zones in the city (Table 3). The standard deviation increased in the morning, when the population congregated in areas of activity, especially for work, but little use was made of leisure and shopping zones. The highest standard deviation value was recorded in the evening, when the population was more concentrated in the centre, engaged in shopping and leisure activities.

4.2. Temporal Twitter activity profiles according to predominant land use

Table 4 shows the number of unique users according to predominant land uses in each spatial unit and major time slots. Residential land use accounted for the highest number of users. For areas of activity, offices accounted for the highest number of unique users, followed by education and retail. An analysis of data density (normalised unique users/Ha) revealed that the highest values corresponded to land uses that attract a large number of people throughout the day (transportation, culture and retail), whereas parks and industry attracted very low densities.

Table 4 also shows the total number of users by time slot and land use. To facilitate comparisons, the graph in Fig. 5 uses relative values to depict very different temporal profiles according to the main types of land use. The curve for residential areas indicates greater activity at night than during the day, while the areas of activity have a much more uneven profile, with a very sharp peak in the morning and a marked drop at night, and as expected, mixed areas show an intermediate situation between the two previous categories.

The temporal distribution of active users in areas of activity was obtained by combining specific profiles associated with different types of land use (Fig. 6). The main transportation infrastructures (airport and railway stations) presented a very early and marked peak. Industry also began very early, although the fluctuations were less pronounced over the course of the day. Education, health and offices showed activity some time later and can be differentiated by the marked peak in the morning in education, which was much less pronounced for offices. Parks and sports areas showed a very stable distribution throughout the day, whereas retail areas were particularly active in the afternoon, after the working day had ended.

4.3. Influence of land use on Twitter activity: OLS models

The relationship between Twitter user distribution and land use (built-up area according to the cadastral data) was analysed using four multiple regression (OLS) models. The results of the OLS models obtained for the different time slots (Table 5) confirmed a close relationship between the spatial distribution of land uses and Twitter activity, since the adjusted coefficients of determination ranged between 0.61 (at night) and 0.76 (in the morning), all with F-statistic values significant at the 0.000 level. Of the explanatory variables considered initially, only two (health and industry) were not significant in all models.[4] At night (from 22 to 24 h) the variables education, culture and parks were not significant either. Most of the population is at home

[3] http://www.madrid.es/portales/munimadrid/es/Inicio/Actualidad/Noticias/Presentado-el-plan-Madrid-Nuevo-Norte?vgnextfmt=default&vgnextoid=c60a71d9fc38d510VgnVCM1000001d4a900aRCRD&vgnextchannel=a12149fa40ec9410VgnVCM100000171f5a0aRCRD.

Table 1
Proposed scenarios.

Types of land use	Scenario base "Madrid nuevo Norte" (development potential 1.05 m^2)	Scenario "Madrid nuevo Norte" (Office) (development potential 1.05 m^2)	Scenario "Potsdamer Platz" (Berlin) (development potential 0.55 m^2)	Scenario "La Défense" (Paris) (development potential 2.25 m^2)
Residence [m^2]	1,100,000	0	575,000	0
Office [m^2]	900,000	2,000,000	500,000	4,285,000
Retail [m^2]	120,000	120,000	63,000	120,000
Transport [m^2]	50,000	50,000	50,000	50,000
Park [m^2]	300,000	300,000	300,000	300,000
Education [m^2]	90,000	90,000	48,000	0
Culture [m^2]	60,000	60,000	30,000	60,000
Health care [m^2]	30,000	30,000	16,000	0
Total [m^2]	2,650,000	2,650,000	1,582,000	4,815,000
Distance to city centre [m]	10,800	10,800	10,800	10,800

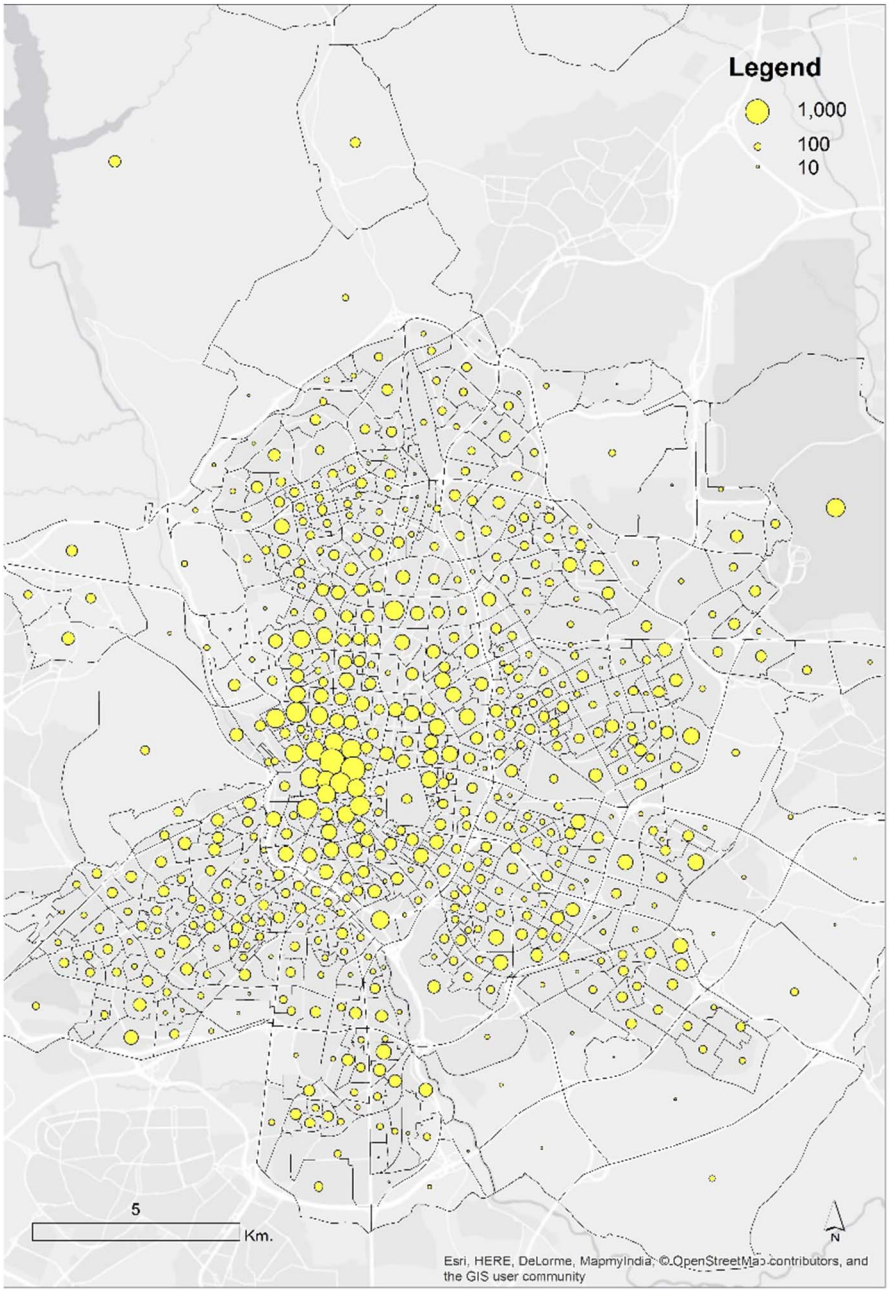

Fig. 3. Active Twitter users at night.

Table 2
Relationships in the distribution of users according to time slot (r2).

	Morning: 08:00 to 14:00	Afternoon: 14:00 to 19:00	Evening: 19:00 to 22:00	Night: 22:00 to 24.00
Morning: 08:00 to 14:00	1			
Afternoon: 14:00 to 19:00	0.89	1		
Evening: 19:00 to 22:00	0.71	0.89	1	
Night: 22:00 to 24:00	**0.49**	**0.72**	**0.81**	**1**

* P value < 0.001.

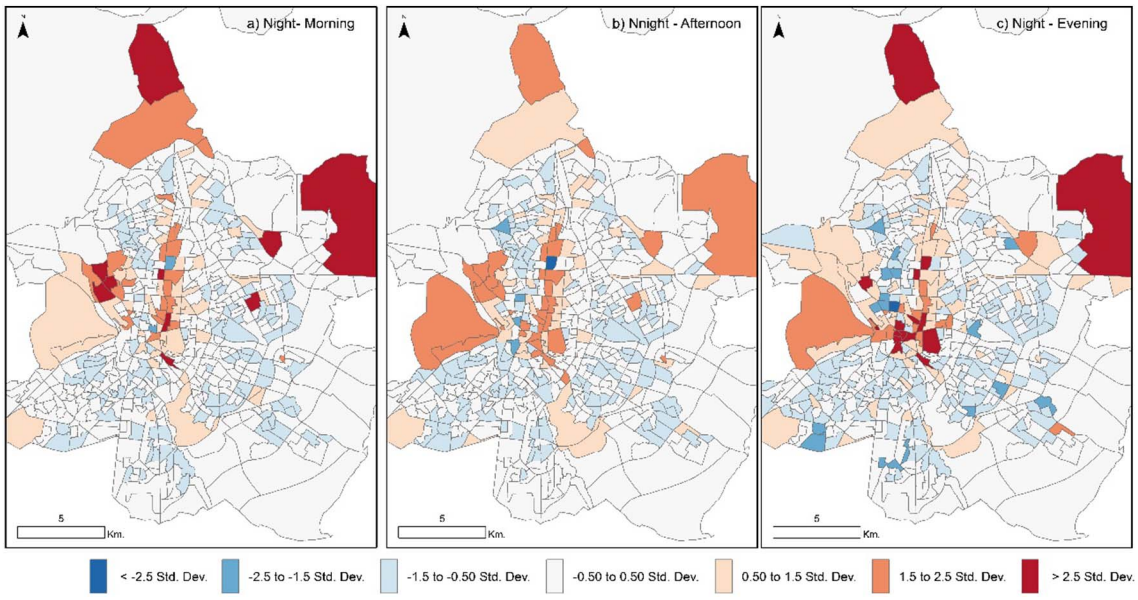

Fig. 4. Residuals in the bivariate correlations of the distribution of active users at night and the rest of the time slots.

Table 3
Descriptive statistics of the distribution according to time slot.

	Morning: 08:00 to 13:59	Afternoon: 14:00 to 18:59	Evening: 19:00 to 21:59	Night: 22:00 to 23:59
No. of zones	584	584	584	584
Minimum	1	1	0	0
Maximum	2015	1299	1536	1099
Total	100,000	100,000	100,000	100,000
Mean	171.2	171.2	171.2	171.2
Standard deviation	166.21	150.7	169.6	141.8

in this time slot (residential). The presence of categories other than residential in the night model might be due primarily to leisure activities, which were encompassed in the category of others, but were also scattered throughout the city centre in areas where residential use was mixed with other land uses such as offices and retail. Irrespective, the coefficients were very low at night for offices, retail and transportation. The VIF parameter was < 2 for all significant variables in the four models, indicating that there was no problem of collinearity between the explanatory variables.

As expected, the coefficients of the four models presented positive signs for the various categories of land use and a negative sign for distance to the centre. Thus, the greater the surface area of each land use category, the higher the number of active tweeters, and the further

Table 4
Temporal distribution of active users according to predominant land use and time slot (normalised data).

Use	Morning: 08:00 to 13:59	Afternoon: 14:00 to 18:59	Evening: 19:00 to 21:59	Night: 22:00 to 23:59	Total day	Normalised unique user/Ha
Residential	57,022	61,330	61,782	69,784	63,237	4.10
Mixed	19,876	19,277	21,061	18,520	18,802	6.11
Activity total	23,102	19,393	17,157	11,696	17,961	1.28
Activity						
Retail	2520	2790	3046	1932	2359	6.30
Culture	698	644	639	372	623	7.16
Education	4466	3187	2015	1039	2653	3.44
Industry	2367	1995	1730	1567	2166	0.48
Office	5974	4856	4289	2871	4358	4.44
Park and sport	2400	2278	2403	1721	2117	0.49
Health care	1378	964	745	473	1006	4.28
Transport	1323	1059	893	407	1068	15.14
Other	1974	1621	1398	1313	1612	0.80
Total	100,000	100,000	100,000	100,000	100,000	3.07

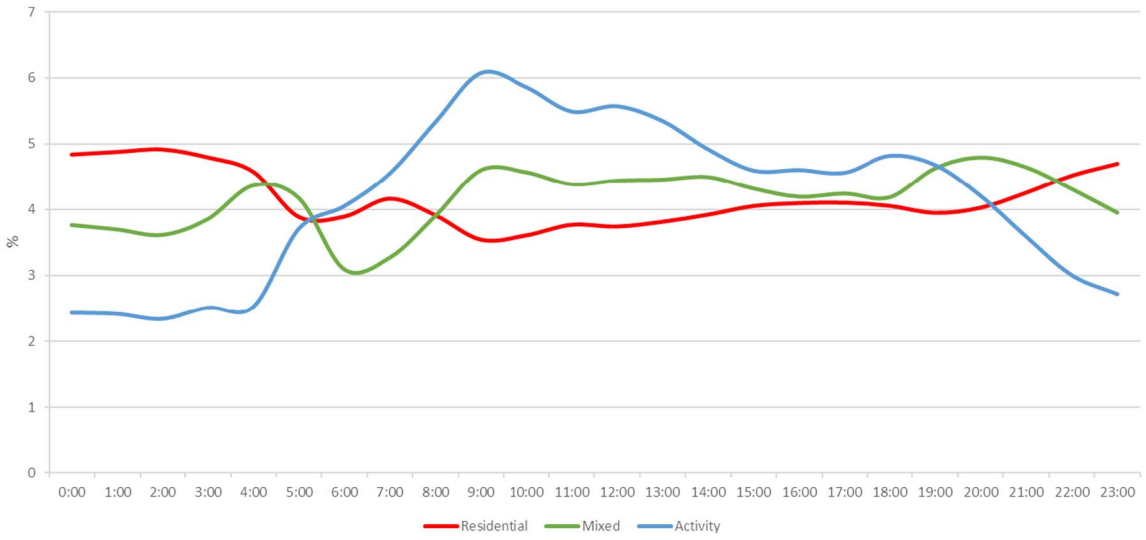

Fig. 5. Temporal distribution of active users according to the main types of land use (percentage).

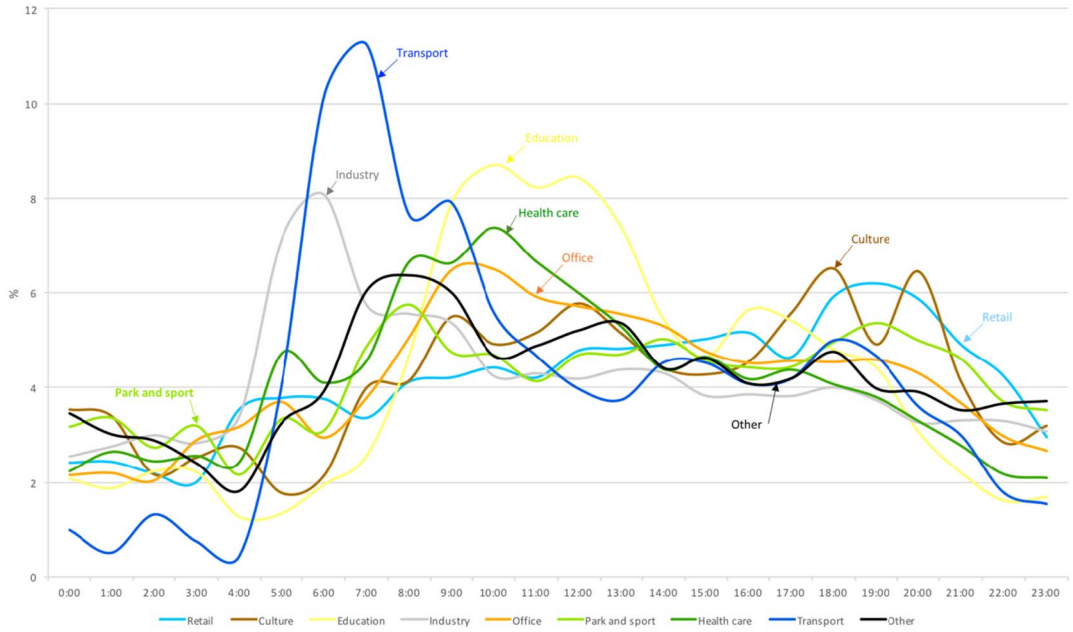

Fig. 6. Temporal distribution of active users according to predominant land use (areas of activity) (percentage).

away from the centre, the lower the amount of Twitter activity. The coefficients offer relevant information on the elasticities between number of tweeters and each of the predictor variables. In general, the highest coefficients corresponded to the variables education, culture and retail, indicating that an increase of one unit (one built-up square metre) in these land use categories resulted in a higher increase in tweeters than that recorded for the other categories.

The variation in the coefficients of the independent variables over the course of the day is consistent with what was observed in the Twitter activity profiles according to predominant land use. The coefficients fell sharply throughout the day in education and transport, whereas this drop was less pronounced in offices and culture. In contrast, the coefficients rose throughout the day in retail, parks and residential zones. The coefficient of the variable distance to the city centre was lower at night, indicating that the centre-periphery gradient of Twitter activity fell at night, given that there is more residential land use on the periphery than in the centre.

4.4. Application of the model: activity in new developments according to each scenario

The application of the OLS models to the different proposals for future urban developments makes it possible to evaluate their future sociodemographic dynamics (activity). Fig. 7 shows Twitter activity throughout the day for the four scenarios proposed for "Madrid Nuevo Norte". The results show clearly that the proposals which promote a mix of land uses (*base scenario* and "Potsdamer Platz" scenario) produce flatter temporal profiles and therefore a demand for services which is

[4] Industry is not an activity conducive to workers making use of social networks (in fact, the density of tweets posted from predominantly industrial areas was very low), but this may vary considerably according to the degree to which these spaces are occupied by companies in the service sector (offices). The second case is more surprising, since patients in waiting rooms have time that they could fill by making use of social networks. One explanation for this finding might be the existence of extremely varied health spaces, ranging from large hospitals to private consultants, with a very different intensity of Twitter use. The fact that health and industrial land uses do not play a significant role is a weakness of the model, since this conditions the application of the model to future urban developments in which these categories of land use are dominant.

Table 5
Results of multiple regression (OLS) models.

Independent variable	Coefficients			
	Morning	Afternoon	Evening	Night
Intercept	123.24*	112.72*	125.48*	110.35*
Residence [m^2]	0.000167*	0.000228*	0.000255*	0.000386*
Office [m^2]	0.000667*	0.000481*	0.000443*	0.000140*
Retail [m^2]	0.000791*	0.001044*	0.001333*	0.000700*
Transport [m^2]	0.000620*	0.000395*	0.000347*	0.000154*
Park [m^2]	0.000012*	0.000016*	0.000016*	
Education [m^2]	0.001909*	0.001169*	0.000626*	
Culture [m^2]	0.001640*	0.001226*	0.001034*	
Health care [m^2]				
Industry [m^2]				
Other [m^2]	0.000055*	0.000049*	0.000056*	0.000044*
Distance to city centre [m]	− 0.013546*	− 0.012342*	− 0.015579*	− 0.011259*
No. Observations	584	584	584	584
R^2	0.763*	0.683*	0.613*	0.609*
Adj. R^2	0.759*	0.678*	0.609*	0.605*
AIC	6811	6867	7121	6912

Variables obtaining non-significant results in the OLS with all variables were discarded and have been left blank.
* Significant at the 0.01 level.

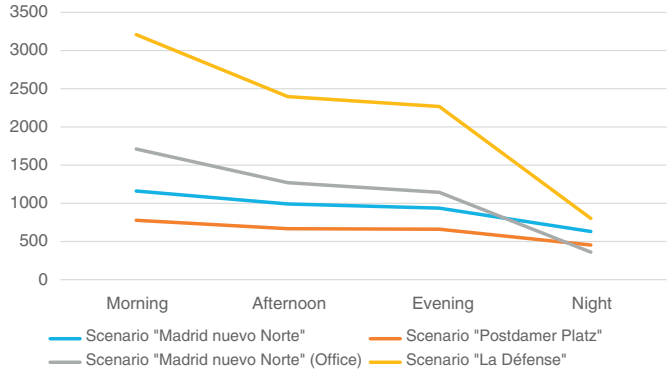

Fig. 7. Twitter activity profiles in the future urban development of Madrid Nuevo Norte according to the 4 proposed scenarios.

more stable in time than the proposals with a lower mix of land uses. As expected, the increase in development potential produces an increase in activity and therefore a greater potential demand for services.

5. Final remarks

Social networks and big data provide previously unavailable information about urban dynamics, opening up new opportunities in the field of urban studies (Graham & Shelton, 2013). In this study, we used data from one of the most widespread social networks (Twitter) to analyse spatiotemporal demographics in a city. Obviously, the data must be filtered and processed before analysis. In this case, we downloaded and geotagged each of the tweets and aggregated the data spatially and temporally in order to obtain the number of active users according to city zone and time slot. In contrast to other studies that have employed classification algorithms to obtain groups of zones and compare them with land use distribution (inferring land use from Twitter activity), our approach pursues a deeper understanding of the influence of land use on activity in the city. Our groups of land uses were predefined according to a more accurate and useful classification for urban planning than the main groups identified in previous research.

Our analysis of the correlation between the distribution of active tweeters in the night time slot and the different day time slots provides an initial insight into city dynamics, revealing the zones which gain or lose activity in each of the time slots over the course of the day. The former are usually located in the centre and the latter predominate in the periphery, with the logical exception of certain areas of activity such as universities, offices or the airport. Our approach has made it possible to visually identify that offices, education and the main transportation terminals gain activity in the morning, whereas retail areas are most active in the evening.

Subsequently, we linked Twitter activity and land use. First, we performed a descriptive analysis based on typical Twitter activity profiles according to the predominant land use in each transport zone. A marked contrast was observed between residential zones, with increased activity at night, and areas of activity, which were much more active during the day. For most activities, the curve fell from the morning to the evening, except for parks (which had a very stable curve) and retail areas (which showed more activity in the evening than in the morning). Transport terminals and industry presented an early peak which was much more pronounced in the former than in the latter. When the predominant use was education, health or offices, this peak appeared later on in the day, being especially marked in education and less so in offices.

An analysis of zone activity profiles is a simplification, since we only considered the predominant land use in each zone, when most of them host several land uses. Furthermore, this is also a purely descriptive analysis. This limitation is overcome by using OLS analysis in order to examine the influence of the different land uses in each zone on Twitter activity according to time slot. Compared with previous studies, more descriptive, our regression outcomes offer an explanation about the relation between land use and active Twitter users' distribution, considering the land use mix in each zone. As expected, the coefficients of the independent variables in all models presented positive signs for the land use variables and a negative sign for distance to the centre. The variation in coefficients over the course of the day is consistent with the results obtained for Twitter activity profiles according to predominant land use: rising for residential and retail zones, and falling for education, offices and transport.

This works confirms something which we already know from our observation of the city without the need to use Big Data: different types of areas within a city have different spatiotemporal profiles. However, the new sources of data make it possible to measure, analyse, model and predict. In this paper, we measure the number of active tweeters in each zone of the city by time slots, as a proxy for the presence of population, and analyse the changes between time slots by means of bivariate correlation analyses and the study of the residuals obtained. We then also analysed the relationship between land uses and active tweeters using temporal profiles and modelled this relationship by means of a multiple regression analysis. The coefficients obtained in this multiple regression analysis serve to predict activity patterns in new urban developments, as has become evident in the case of the new urban development “Madrid Nuevo Norte”. The scenarios analysed for this new urban development demonstrate that the mix of uses generates “flatter” time temporal profiles of activity and therefore a demand for services which is more stable in time, while specialisation of uses (for example, office space) produces profiles which are much more “abrupt” and therefore a greater fluctuation in the demand for services, making these less efficient. Furthermore, it has been demonstrated that density also significantly affects patterns of activity, thus increasing or reducing the potential demand for services.

Our analyses provide useful information for urban planning. The results shed light on urban dynamics in relation to land use. This is of interest as regards the provision of services in the public sector (for example, risk assessment and population evacuation plans) and for private sector business activities (potential demand according to zone and time of day). In addition, knowledge of the link between activities and land use can be used to predict future patterns of activity in new urban developments, by using the OLS model to estimate the

spatiotemporal distribution of the population according to the envisaged land uses in a new development.

It is clear that social networking data present biases. In the case of Twitter in Spain, although generating an enormous amount of data every day, it is only used by a 25% of the population (We Are Social: https://wearesocial.com/sg/blog/2017/01/digital-in-2017-global-overview). Twitter users are not representative of the whole population. For example, among Twitter users in Spain, women (46%), people over 54 (6%) and those who have not completed university studies (59%) are under-represented (). The sample of mobile phone users is much larger than that of Twitter users and is better distributed among the various segments of the population, thus reducing the problems associated with sample bias and allowing greater spatial disaggregation than with Twitter data (less spatial units with 0 tweets), but mobile phone data are extremely expensive. In contrast, Twitter data is free and this is a significant advantage compared with mobile phone data. Our analysis using Twitter data offers consistent results, so that Twitter seems to be a good alternative to mobile phone records when it comes to analysing general trends in the pulse of the city. Future research can determine whether Twitter data and mobile phone records offer comparable results in studying the pulse of the city.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.cities.2017.09.007.

Acknowledgments

The authors gratefully acknowledge funding from the ICT Theme (611307) of the European Union's Seventh Framework Program (INSIGHT project-Innovative Policy Modeling and Governance Tools for Sustainable Post-Crisis Urban Development, GA 611307), the Madrid Regional Government (SOCIALBIGDATA-CM, S2015/HUM-3427) and the Spanish Ministry of Economy and Competitiveness and the European Regional Development Fund (DynAccess, TRA2015-65283).

References

Andrienko, G., Andrienko, N., Bosch, H., Ertl, T., Fuchs, G., Jankowski, P., & Thom, D. (2013). Thematic patterns in georeferenced tweets through space-time visual analytics. *Computing in Science & Engineering, 15*(3), 72–82.

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), e0129202.

Chen, Y., Liu, X., Li, X., Liu, X., Yao, Y., Hu, G., ... Pei, F. (2017). Delineating urban functional areas with building-level social media data: A dynamic time warping (DTW) distance based k-medoids method. *Landscape and Urban Planning, 160*, 48–60.

Ciuccarelli, P., Lupi, G., & Simeone, L. (2014). *Visualizing the data city.* Springer International Publishing17–22.

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

Frias-Martinez, V., Soto, V., Hohwald, H., & Frias-Martinez, E. (2012). Characterizing urban landscapes using geolocated tweets. *Privacy, security, risk and trust (PASSAT), 2012 international conference on and 2012 International Conference on Social Computing (SocialCom)* (pp. 239–248). IEEE.

Gao, S., Yang, J. A., Yan, B., Hu, Y., Janowicz, K., & McKenzie, G. (2014). Detecting Origin-Destination mobility flows from geotagged tweets in Greater Los Angeles Area. *Eighth international conference on geographic information science (GIScience'14)*.

Graham, M., & Shelton, T. (2013). Geography and the future of big data, big data and the future of geography. *Dialogues in Human Geography, 3*(3), 255–261.

Jiang, B., Ma, D., Yin, J., & Sandberg, M. (2016). Spatial distribution of city tweets and their densities. *Geographical Analysis, 48*, 337–351.

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.

Lansley, G., & Longley, P. A. (2016). The geography of Twitter topics in London. *Computers, Environment and Urban Systems, 58*, 85–96.

Lenormand, M., Picornell, M., Cantú-Ros, O., Tugores, A., Louail, T., Herranz, R., ... Ramasco, J. J. (2014). Cross-checking different sources of mobility information. *PloS One, 9*, e105184.

Lin, J., & Cromley, R. G. (2015). Evaluating geo-located Twitter data as a control layer for areal interpolation of population. *Applied Geography, 58*, 41–47.

Longley, P. A., & Adnan, M. (2016). Geo-temporal Twitter demographics. *International Journal of Geographical Information Science, 30*(2), 369–389.

Longley, P. A., Adnan, M., & Lansley, G. (2015). The geotemporal demographics of Twitter usage. *Environment & Planning A, 47*(2), 465–484.

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.

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.

Mocanu, D., Baronchelli, A., Perra, N., Gonçalves, B., Zhang, Q., & Vespignani, A. (2013). The twitter of babel: Mapping world languages through microblogging platforms. *PloS One, 8*(4), e61981.

Murthy, D. (2013). *Twitter: Social communication in the Twitter age.* John Wiley & Sons.

Netto, V. M., Pinheiro, M., Meirelles, J. V., & Leite, H. (2015). Digital footprints in the cityscape. *International conference on social networks*. USA: Athens.

Openshaw, S. (1984). *The modifiable areal unit problem.* Norwich, England: Geobooks.

Pei, T., Sobolevsky, S., Ratti, C., Shaw, S. L., Li, T., & Zhou, C. (2014). A new insight into land use classification based on aggregated mobile phone data. *International Journal of Geographical Information Science, 28*(9), 1988–2007.

Ríos, S. A., & Muñoz, R. (2017). Land use detection with cell phone data using topic models: Case Santiago, Chile. *Computers, Environment and Urban Systems, 61*, 39–48.

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.

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

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), e97010.

Zhan, X., Ukkusuri, S. V., & Zhu, F. (2014). Inferring urban land use using large-scale social media check-in data. *Networks and Spatial Economics, 14*(3–4), 647–667.

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.

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.