Sensing Urban Land Use with Twitter Activity

Vanessa Frias-Martinez, Victor Soto, Heath Hohwald and Enrique Frias-Martinez

Telefonica Research, Madrid – Spain
{vanessa,vsoto,heath,efm}@tid.es

Abstract

Individuals generate vast amounts of geolocalized content through the use of mobile social media applications. In this context, Twitter has become an important sensor of the interactions between individuals and their environment. Building on this idea, the authors propose the use of geolocated tweets as a complementary source of information for urban planning applications, focusing on the characterization of land use. The authors' proposed technique automatically determines land uses in urban areas by clustering geographical regions with similar tweeting activity patterns. Three case studies are presented and validated for Manhattan, London and Madrid using Twitter activity and land use information provided by the city planning departments. Results indicate that geolocated tweets can be used as a powerful data source for urban planning applications.

Keywords: modeling human behavior, land use detection, clustering

1. Introduction

Cell phones have become one of the main sensors of human behavior, thanks, among others, to their growing penetration and wealth of user applications such as Whatsapp, Facebook, Twitter, Foursquare or Flickr. From

messaging to social networking, these tools are used by citizens on the go. In fact, the mobile nature of cell phones promotes the use of such applications anytime, anywhere, thereby generating vast amounts of human behavioral information. Additionally, many mobile social media applications allow users to add geolocation information to their profiles or to the information they share, enhancing the richness of the behavioral datasets. For example, Twitter offers the possibility of recording the user's geographical coordinates each time a tweet is generated. From this perspective, there is a potential use of using geolocated user-generated content as a complementary source of information for urban planning applications.

Urban planning is a process that focuses on the control and on the design of urban environments in order to increase the well being of citizens. An important concern in urban planing is the characterization of urban land use. Such information is usually gathered through direct observation or using questionnaires that attempt to capture how citizens interact with their urban environment. Nevertheless, this approach has some limitations such as the resiliency of citizens to provide such information or the cost of running questionnaires, which highly limits the frequency with which the information is captured. Alternative approaches such as GIS (Geographic Information Systems) provide satellite imagery that might reveal some types of land use information through image processing techniques. However, such techniques fail to provide real time information as images are not captured frequently.

Here we present a novel approach for sensing urban land use that exclusively makes use of spatial (geo-tagged) and temporal (time-stamped) information, without accessing personal details or the content of the user-

generated information. By doing so, our techniques preserve privacy and can potentially be applied and/or complemented with any other mobile social media dataset with geolocation information. In order to validate to which extent tweetting activity can be used to characterize urban land use, we study three urban environments: Manhattan (NYC), London (UK) and Madrid (Spain) using geolocated tweets and land use information provided by city planning departments.

2. Related Work

Our work arises as a combination of two smart cities research areas mainly crowd modeling and urban computing for urban planning. Different authors have used a variety of user-generated content services for implementing such solutions. Wakamiya *et al.* [1] and Fujisaka *et al.* [2] studied how to exploit geotagged tweets and the semantics of its content to interpret individual and crowd behavior *i.e.,* how individuals and groups of people move across geographical areas. They propose models of aggregation and dispersion as a proxy to understand the bursty nature of human mobility. Similarly, Kinsella *et al.* [3] used geolocated tweets, together with their content, to create language models at varying levels of granularity (from zip codes to countries). The authors use these models to predict both the location of the tweet and the user based on location changes. There are interesting results using geotagged information from Foursquare and Flickr to model land use in urban environments. For example, Noulas *et al.* [4] have used the geolocated information provided by Foursquare to model crowd activity patterns in London and New York City using spectral clustering. For that purpose,

the authors characterize the activity patterns identified by the clusters using the predefined Foursquare categories that give an indication of the type of check-in location (restaurants, academic, etc.). As such, this approach gives an approximated understanding of land use. In a related work Cranshaw et al. [5] present a new clustering model designted to study social dynamics on a large scale using two FourSquare datasets. The results are validated with personal interviews that confirm the clusters identified.

3. Sensing Urban Land Uses using Twitter

The identification of urban land uses from geotagged tweets using the spatial (localization) and temporal (timestamp) information has two steps: land segmentation and land use detection.

3.1. Land Segmentation with Geotagged Data

Given that we want to sense land uses in different urban regions, the first step consists on partitioning the land into different segments, which can then be characterized by its tweet usage. The partitioning of the area considered has to preserve the topological properties of the geolocalized tweets, while respecting the actual shape of the geographical area under study.

We approached this problem using Self-Organizing Maps (SOM) [6], which reduce the input data dimensionality to be able to represent its distribution as a map. In our case, the input data are the latitude & longitude pairs that represent the geolocalized tweets over a period of time for a specific urban area. Thus, we use a SOM to build a map that segments the urban land into geographical areas with different concentrations of tweets.

The SOM consists of a collection of N neurons organized in a grid $[p, q]$, with $N = p * q$. Since we can choose any initial size $[p, q]$ for the map, our method explores different map sizes and selects as the best land segmentation map the topology that minimizes the Davies-Bouldin clustering index [7]:

$$DB = \frac{1}{n} \sum_{i=1}^{n} \max_{i \neq j} \left(\frac{\sigma_i + \sigma_j}{d(w_i, w_j)} \right) \quad (1)$$

The DB index is chosen because the partition with minimum DB value will minimize the maximum sum of a pair of standard deviations σ_i and σ_j $i \neq j$ and maximize the distance between cluster representatives, ensuring that even the most disperse clusters concentrate its points (geolocated tweets) inside a compact cluster.

As a result of the process, we obtain a map where each neuron represents a pointer to a region with a high density of tweets. Additionally, areas with larger concentrations of tweets will have larger numbers of neurons geographically located nearby. Finally, Voronoi tessellation is applied over the location of the neurons in order to compute the land segments that each neuron represents. These land segments are used as the elements for the characterization of land use.

3.2. Detecting Urban Land Uses

We characterize each land segment by its average tweet activity, which will then be used to identify common land uses across land segments. For each land segment s, a tweet-activity vector X_s representing the average tweetting behavior is built as:

1. An activity vector $x_{s,n}$ for land segment s is built for each day $n = 1, ..., d$ in the twitter dataset.

2. Each day n in the activity vector contains 72 components $x_{s,d}(t), t = 1, ..., 72$ where each one represents the number of tweets generated in land segment s during a 20-minute interval t in day d.
3. An average activity vector for land segment s is computed for both weekdays $X_{s,wkd}$ and weekends $X_{s,wkn}$ as $X_{s,wkd}(t) = \frac{\sum_{n=1}^{d} x_{s,n}(t)}{n}, t = 1, ..., 72$ where n is a weekday and $X_{s,wkn}(t) = \frac{\sum_{n=1}^{d} x_{s,n}(t)}{n}, t = 1, ..., 72$ where n is a weekend day.
4. The final activity vector is represented as the concatenation of weekday and weekend average activity vectors $X_s = \{X_{s,wkd}, X_{s,wkn}\}$ and is normalized as:

$$\hat{X}_s(t) = \frac{X_s(t)}{\sum_{t=1}^{72} X_{s,wkd}(t) + \sum_{t=1}^{72} X_{s,wkn}(t)}. \quad (2)$$

After this process, each land segment s is represented by a unique activity vector X_s with 144 elements representing the average weekday and weekend tweeting activity computed in 20-minute timeslots. These activity vectors are used to automatically identify and characterize urban land uses using spectral clustering to reveal clusters of common tweeting behaviors across land segments [8, 4]. We posit that the land use can be derived from a careful analysis of the tweeting behaviors in each cluster, based on its activity vector as well as on its physical layout in the city.

Spectral clustering treats the data clustering as a graph partitioning problem without any assumption on the form of the clusters. It starts by constructing a similarity graph (D) and a weight matrix (W) which are then used to construct a Laplacian matrix L. The algorithms performs a dimensionality reduction and then it applies a clustering algorithm, typically k-means.

	Dataset			
	Total	Mean	Area($km2$)	Density (tweets/$km2$)
World	24130423	492457.61	-	-
Manhattan	247381	5048.59	60	84.13
London	312513	6377.81	150	42.51
Madrid	46931	957.77	88	10.88

Table 1: Dataset Characteristics

In order to identify the number of clusters k eigengap detection is specially suited. With this approach the number of clusters k is defined by the point where there is a drop in the magnitude of the eigenvalues of the Laplacian matrix arranged in increasing order. Once the best value of k is selected, the method outputs the clusters of land segments.

In order to analyze the type of land use associated to each cluster, we average the activity vectors of all the land segments in the cluster and compute an average activity vector that represents the tweeting activity for that cluster. Finally, we hypothesize the land use for each cluster based on its tweeting activity and its distribution across the urban environment under study.

4. Evaluation of Land Uses

We present an evaluation of our land use detection method for three cities: Manhattan (NYC), London (UK) and Madrid (Spain). We have selected these three cities because they show different densities of Twitter activity computed as the number of daily tweets per square kilometer in the urban

perimeter considered: Manhattan has the highest Twitter density (84.13 tweets/km^2) followed by London with about half of that (42.51 tweets/km^2) and Madrid with a density of 10.88 tweets per square kilometer. As such, these three cities represent different cultural and behavioral Twitter attitudes useful to evaluate our propossed aproach.

The objective of this evaluation is twofold: (1) to analyze to which extent the land use identification algorithm detects different types of land uses and (2) to understand the impact of the density of tweets on the accuracy of the method proposed. For that purpose, we apply the algorithm on datasets of Twitter activity from each one of the three cities considered and draw hypothesis regarding land uses based on both cluster tweeting activity and their location. To validate our results, we contrast our clusters and land uses hypotheses against real land use information collected by the corresponding city planning department.

4.1. Twitter Datasets

Twitter users are allowed to tag tweets with their current geospatial location. Specifically, users can set their geographical location by specifying a city or region by themselves or by allowing Twitter to track their GPS longitude and latitude coordinates. When a new tweet is produced, Twitter records the geographical information of the user at that moment, along with a variety of other meta data. Given that we want to model land use within an urban environment, we require highly granular geolocations. Thus, we only collect tweets whose location is automatically recorded by Twitter through the GPS and not self-reported by the user.

We used the Twitter Streaming API [9] to gather geolocalized tweets in

near real-time. The streaming API enables a high-throughput stream to be established with Twitter by which a large volume of public statuses of tweets can be gathered. Specifically, the Twitter steaming API provides a sample of all tweet public statuses, currently about one percent of the full Firehose set of tweets. Our final Twitter dataset consists of 49 days (seven weeks) of geolocated tweets worldwide from October 25th to December 12th, 2010. Although our study focuses on Manhattan, London and Madrid, we collected tweets worldwide mostly for sanity purposes.

Table 1 shows the general statistics for the dataset collected describing the total and average daily number of geotagged tweets worldwide as well as individually for Manhattan, London and Madrid during the period under study in 2010. We also show the Twitter activity densities for each of the cities. The geographical area for London is defined by the area within Ringway 1. As for Madrid, we consider the urban area comprised within the $M-30$ highway.

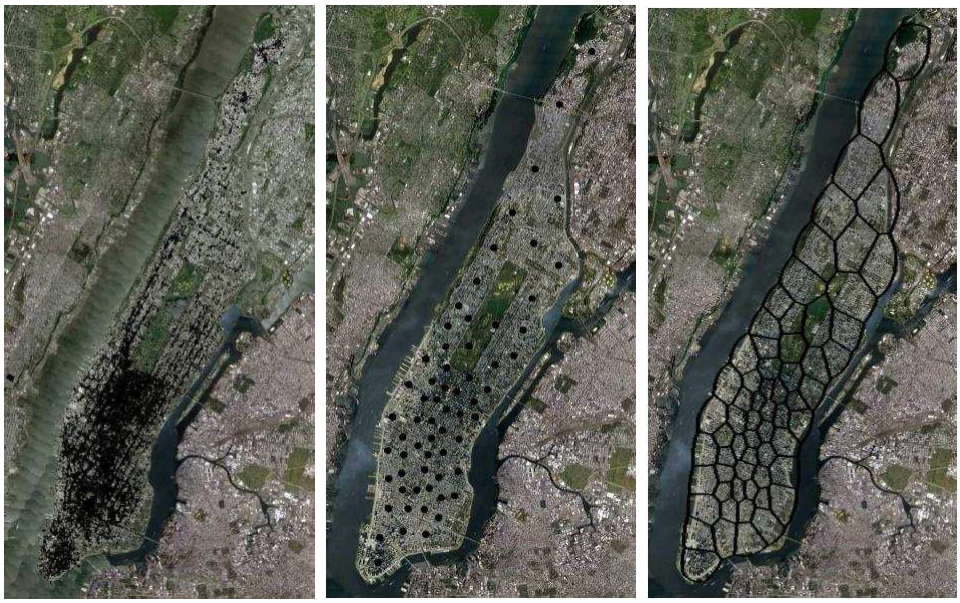
4.2. Land Segmentation and Land Use Clustering

Our method trains a SOM with the set of geolocated tweets to divide the urban area under study into different land segments s characterized by their tweeting activity vector X_s. Given the different geographies of the cities under study, we evaluated N in the range $N = [10, ..., 300]$ with N defined as $N = p \cdot q \quad p, q > 1,\ p, q \in \mathbb{N}$. The values of p and q define the number of neurons considered in each axis: p in the north-south axis and q in the east-west axis (we leave out the cases where N is a prime number). To adjust the neurons to the shape of each city, we only consider cases in which $p > q$ for Manhattan and Madrid and $p < q$ for London (Manhattan and Madrid

have a longer north-sout axis and London has a longer east-west axis). For example, in Manhattan $N = 10$ would define an initial grid with $p = 5$ and $q = 2$ and and $n = 12$ would generate $(p = 6, q = 2)$ and $(p = 4, q = 3)$.

Due to the randomized nature of the SOM training stage, 100 SOMs are trained for each city and each pair (p, q) with $N = p * q \in [10, ..., 300]$ and their average DB index is computed. The minimum DB index was associated to $N = 64$ for Manhattan with $p = 16$ and $q = 4$; $N = 168$ for London with $p = 12$ and $q = 14$; and finally $N = 91$ for Madrid with $p = 7$ and $q = 13$. As an example, Figure 1 shows the land segmentation for Manhattan. We observe that the Midtown area, where the best part of the tweets are geolocated (as shown in Figure 1(left)), shows a high density of neurons; whereas the north of Manhattan, with a scarce number of tweets, has a much smaller number of neurons. Finally, the land segmentation is computed by applying Voronoi tessellation [10] to each SOM centroid in the two-dimensional space as shown in Figure 1(right). Notice that areas with larger polygons represent areas with reduced Twitter activity.

Each one of the land segments identified in Manhattan, London and Madrid is characterized by its Twitter activity vector X_s which has 144 components, the first 72 describe the tweeting activity during an average weekday and the last 72 the activity during and average weekend day. Our method uses the set of X_s vectors to identify different land uses for each city separately identifying clusters of similar normalized activity using spectral clustering. Following the eigenvector detection approach, the best number of land segment clusters is identified as $k = 4$ for Manhattan, $k = 5$ for London and $k = 4$ for Madrid.

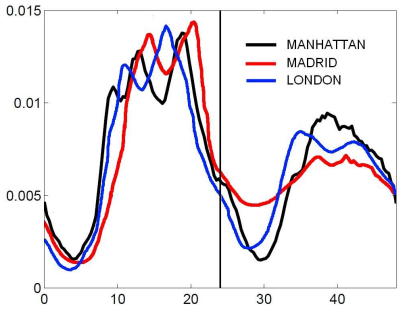


(a) (b) (c)

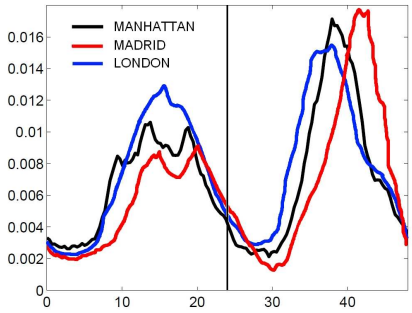
Figure 1: Land segmentation for Manhattan: (left) data points, (center) centers of activity computed with SOM and (right) Voronoi tessellation.

In order to understand the types of land uses identified by these clusters, we analyze the class representatives for each cluster together with its geographical distribution over the city map. A combined analysis can be used to provide a hypothesis about the potential types of land uses. Figure 2 presents the class representatives for each of the clusters identified across the three cities. Each representative or behavioral signature is computed as the average number of hourly tweets and is normalized per cluster and per city. For analytical purposes, we group the signatures across cities by euclidean similarity. We hypothesize that signatures that share similar shapes across cities represent comparable land use types.

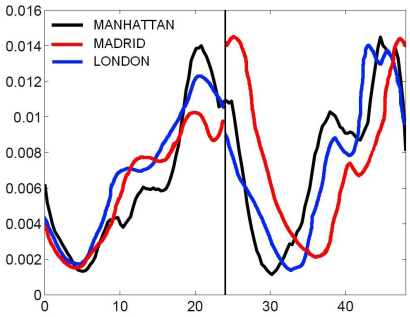
We observe that the activity vectors in **Cluster 1** are generally characterized by a larger tweetting activity during weekdays than weekends (see Figure 2(a)). During weekdays the highest tweeting activity is reached at around 9:30PM, 13:00PM and 8:30PM for Manhattan, which might be associated to the times at which people typically get to work, go for lunch, and leave work. The city of London shows similar peaks but slightly shifted in time. In the case of Madrid, the signature is shifted even more, suggesting that working hours might happen a little bit later during the day. The peak of the tweeting activity during the weekends is reduced by approximately 40% when compared to weekdays. In terms of geolocation of the clusters, these cover, among others, areas like Battery Park or Wall Street in Manhattan(see Figure 3 , the City and Canary Warf in London (see Figure 5 and the sorroundings of Castellana and the area of AZCA in Madrid (see Figure 4), all areas heavely associated with business/office activities. For these reasons, we hypothesize that the geographical area covered by this cluster represents



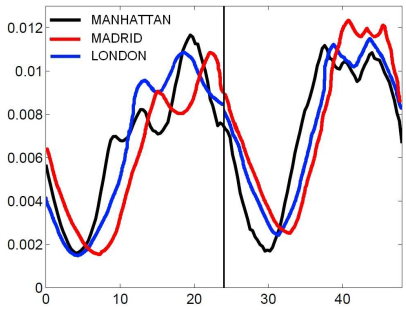
(a) Cluster Type1: Business



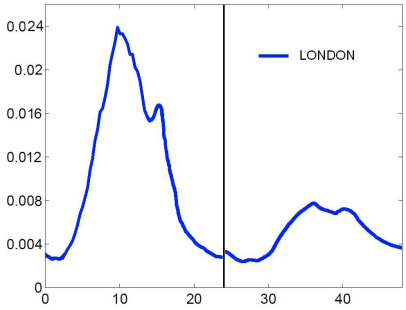
(b) Cluster Type 2: Leisure/Weekend



(c) Cluster Type 3: Nightlife



(d) Cluster Type 4: Residential



(e) Cluster Type 5: Industrial

Figure 2: Tweeting activity signatures per cluster for Manhattan, London and Madrid. The Y axis represents the normalized tweeting activity and the X axis two 24-hour periods, the first one for an average weekday and the second one for an average weekend.

Figure 3: Physical layout of business, nightlife and leisure clusters in Manhattan. Areas not marked with any color indicate residential land use.

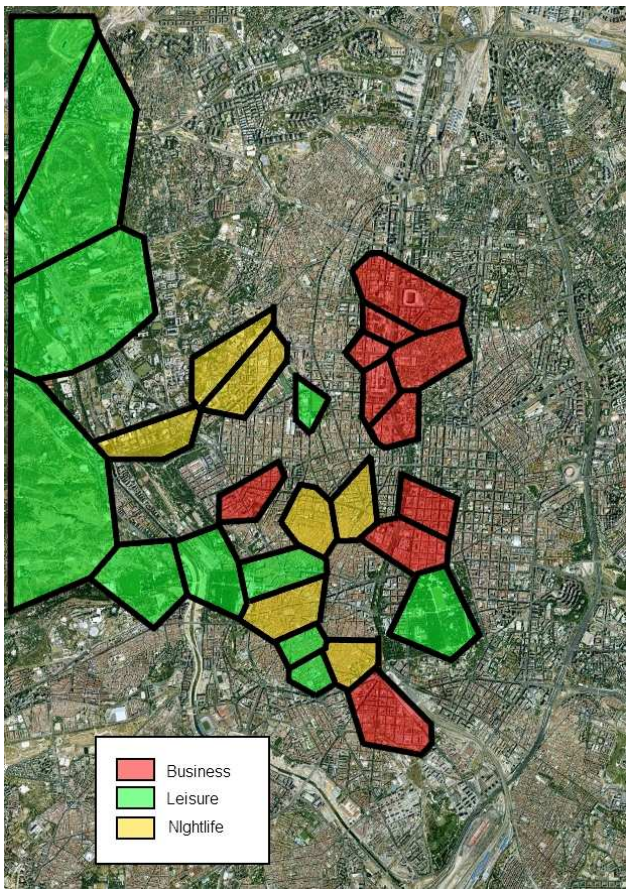


Figure 4: Physical layout of business, nightlife and leisure clusters in Madrid. Areas not marked with any color indicate residential land use.

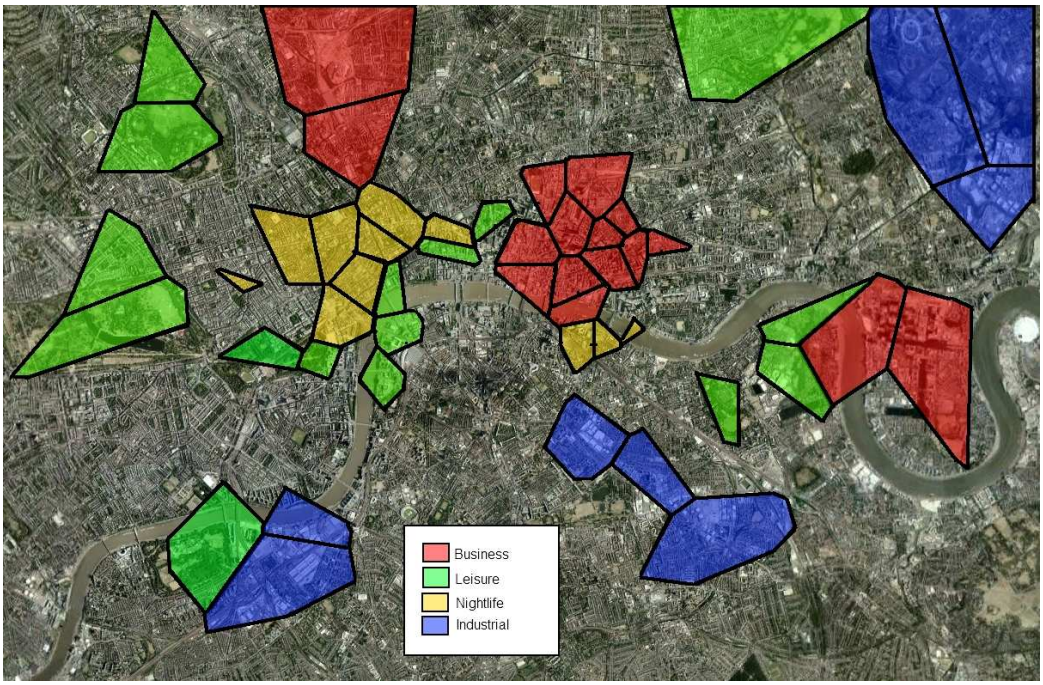


Figure 5: Physical layout of business, nightlife, leisure and industrial clusters in London. Areas not marked with any color indicate residential land use.

Business areas in Manhattan, London and Madrid.

Cluster 2 shows a large difference between weekend and weekday activity, in fact, the signature is almost doubled in volume (see Figure 2(b)). During weekends tweeting activity increases until the afternoon to continue in a constant decrease. Geographically, these clusters cover regions like Central Park and nearby museums in Manhattan, Hyde Park or Regents Park in London and El Retiro Park and Casa de Campo Recreational Park in Madrid. Also included are heavily touristic areas, like Sol and the Flea Market of El Rastro in Madrid, or the London Eye, Buckingham Palace and Covent Garden in London. Thus, we hypothesize that this cluster can be associated to Leisure or Weekend activities since users are active mostly during the weekends. However, we believe that it does not represent weekend nightlife since the tweeting activity highly decreases after 16:00PM during the weekends.

On the other hand, **Cluster 3** is associated to very large activity peaks at night (see Figure 2(c)). These peaks happen at around 20:00-21:00PM during weekdays and between 00:00-06:00AM during the weekends. We observe that the peaks happen earlier in London and Manhattan while a little bit later in Madrid suggesting that nightlife might continue until late hours in this city. Studying the physical layout of these clusters on the city maps, we observe that they cover areas like the East Village in Manhattan; the West End in London and Malasana/Chueca and Alonso Martinez in Madrid (see Figure 4), areas associated with restaurants, pubs and discos. All these elements suggest that this cluster might represent nightlife activities.

Cluster 4 shows a signature evently divided between weekends and weekdays, where, during weekdays, there is a peak of activity in the afternoon

between 6pm and 8pm depending on the urban area considered (6pm for Manhattan and London and 8pm for Madrid). Activity during weekends is of the same magnitude as in weekdays (see Figure 2(d)). This is the most important cluster considering the geographical area covered and the number of clusters included. The geographic layout of the clusters cover heavily residential areas in all cities: in Manhhatan the limits of the island, and in Madrid and London the outskirts of the areas considered. In Figures 3, 4 and 5, the areas include with this cluster are the ones not marked with any color. Our hypothesis for this type of signature is that it represents residential land use with citizens tweeting from home at any time during the weekends and after working hours during the week.

Finally, **Cluster 5** is only identified for London (see Figure 2(e)). Its signature is characterized by very little activity during the weekends. The weekdays show a peak in activity very early, at around 10am, after which a steady decrease happens showing little activity during the rest of the day. Looking at the physical layout, these clusters cover areas in the east and sout of the city, like the are around Battersea station and the Olympic park. As a result of that, we hypothesize that this cluster represents Industrial land use (see Figure 5).

Finally, it is important to clarify that we have only focused on identifying the main land use of each cluster (although there might be other minor ones), since this is the way urban planners typically compute land use

4.3. Land Use Validation

In order to validate our land use hypotheses, we compare the evaluation results against official land use data released by the NYC Department of

City Planning and the NYC Department of Parks&Recreation through the NYC Open Data Initiative [11]; against the ward profiles released by the London Datastore Open Data Initiative [12] and against the district land use information computed by the Urban Planning Department in Madrid's City Hall [13]. These catalogs are produced by city agencies typically through a combination of on-site inspections, interviews and questionnaires.

The NYC Department of City Planning considers four main land use types: (1) residential, (2) commercial, (3) industrial and (4) parks&recreation (see Figure 6(a) for details). On the other hand, the information provided by the *London Datastore* considers three types of wards: (1) domestic buildings, which we associate to residential areas, (2) non-domestic buildings, that we pair up with business and industrial land use wards and (3) greenspace and paths. Finally, the information provided by the Urban City Planning Department in Madrid provides land use information at a district level and considers four types: (1) residential at different density levels (which we group), (2) industrial, (3) services(commercial&business) and (4) greenspaces, as can be seen in Figure 6(b).

To understand how well the clusters we have identified using Twitter activity represent the official land use areas, we evaluate the percentage of overlapping that exists between the physical layout of the clusters and the official land use map for each city under study. Such analysis will give us an understanding of the accuracy of our approach to identify land uses as well as of the impact that the Twitter density might have on the quality of the results. It is important to highlight that the percentage of overlapping is an approximate measure to validate land use identification given that both

Official Land Use	Twitter Land Use			
	Business	Residential	Nightlife	Leisure&Weekend
Commercial	**81%**	13%	3%	2%
Residential	7%	**68%**	19%	4%
Industry	13%	77%	**0%**	6%
Parks&Recreation	6%	7%	6%	**81%**

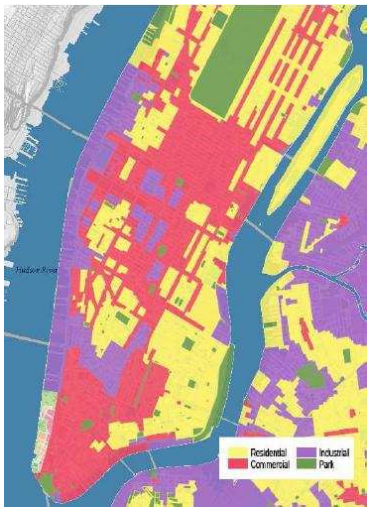
Table 2: Manhattan: Percentage of overlap between official land uses and Twitter land uses.

Official Land Use	Twitter Land Use				
	Business	Residential	Nightlife	Leisure&Weekend	Industrial
$Non-domestic\ buildings$	**61%**	9%	3%	2%	25%
Domestic buildings	9%	**56%**	23%	6%	6%
Greenspace&Paths	8%	11%	7%	**72%**	2%

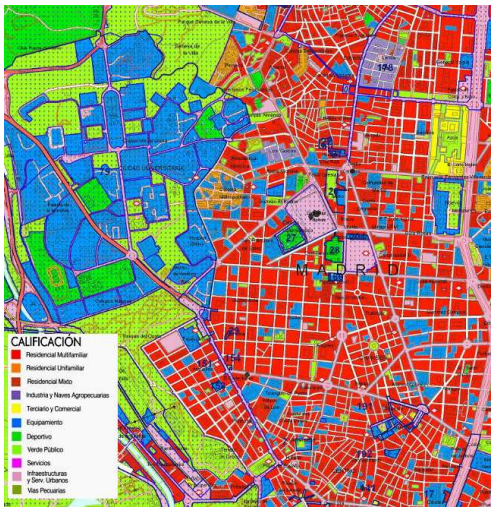
Table 3: London: Percentage of overlap between official land uses and Twitter land uses.

maps have different granularities: our cluster maps represent land segment clusters based on Voronoi and tweet density whereas the official land use maps show data at a block, ward or district level, depending on the city.

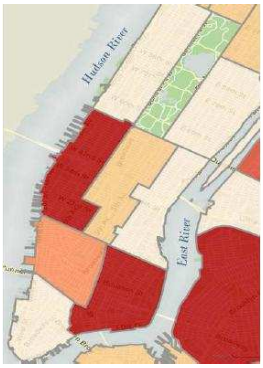
Tables 2, 3 and 4 show the percentages of overlap between the official land use maps for each city (rows) and our land use hypotheses (columns). Each element (i, j) in the tables represents the percentage of the official land use region that is covered by one of our land use clusters *i.e.,* Business, Residential, Nightlife, Leisure and Industrial. Note that in the case of Manhattan our Voronoi approximation to the island does not precisely cover all Man-



(a) Manhattan official land uses: Commercial, Residential, Industrial and Parks&Leisure official.



(b) Madrid official land uses: Comercial&Business, Residential (different densities considered), Industrial and Greespaces.



(c) Manhattan:Community Districts with Noise Complaints from the NYC 311 Service (red represents the highest number of complaints)

Figure 6: Official Land Use Maps for Manhattan (a) and Madrid (b) and (c) Number of noise complaints in Manhattan per district.

Official Land Use	Twitter Land Use			
	Business	Residential	Nightlife	Leisure&Weekend
Commercial&Business	**69%**	25%	4%	2%
Residential	11%	**61%**	18%	10%
Industry	58%	33%	**3%**	6%
Greenspace	7%	16%	6%	**71%**

Table 4: Madrid: Percentage of overlap between official land uses and Twitter land uses.

hattan land due to its irregularities, and as a result the percentages does not exactly sum up to 100%. Comparing our results across cities, we observe that Manhattan shows the highest percentages of official areas covered by our clusters, whereas London and Madrid share lower accuracies in terms of land use identification. It appears that the higher tweeting density that Manhattan has ($84.13/km2$) has, as expected, a positive impact on the quality of land use identification.

The official *Commercial* and *Business* land uses in the three cities are identified quite well by our business cluster with area coverages between $61\% - 81\%$. London is a special case in which the official non-residential land use is partially identified by our business cluster (61%) but also by our industrial cluster (25%). Similarly, the official *Residential/Domestic buildings* land use has a high overlap with our residential cluster with coverages between 56% and 68% of the official areas. However, we observe a generalized trend across the three cities whereby around a 20% of the official residential area is also covered by our nightlife clusters, probably highlighting residential areas with high densities of bars and restaurants. This is in fact common

in areas like the East Village in New York, Chelsea in London or Chueca in Madrid.

While in London we are able to detect Industrial land use, and compare it to the oficial non-residential land use, the official Industry land use, present in Manhattan and in Madrid, goes undetected. We consider that the main reason for that is that in both cases, within the area of the city considered, industrial land is minimum (less than 3% of the total area in Madrid and less that8% in Manhattam), and as a result they are included in larger Voronoi elements that has a different stronger land use. In fact, most of the official industrial land use is subsumed by our residential cluster in the case of Manhattan whereas in Madrid it is mostly covered by our business cluster. This might indicate that workers in the industrial areas are not using Twitter as much as people that live and/or work in that area, and as a result our technique captures the main land use, i.e. the oficial land use goes undetected due to lack of activity. Finally, the official *Parks&Recreation* and *Greenspace&Paths* land use is identified by our leisure cluster with overlaps between 71% and 81% of the official land use maps.

On the other hand, our method identified a Nightlife cluster that has no counterpart in any of the oficial land uses. Nighlife clusters mostly overlap with the official Residential land use. However, we wanted to understand whether the cluster is incorrect or whether it is modeling a different type of land use not accounted for by the city halls. For the particular case of Manhattan, we identified the number of noise complaints per community district made to the 311 on-line service during 2010 (see Figure 6(c)). Given that the community districts have much lower granularity than our land use

clusters, we computed the percentage of the Nightlife cluster that is included within the districts with the highest number of complaints, which corresponds to an 82% of overlap. Thus, it is fair to say that the Nightlife cluster detected by our method identifies a Nightlife land use that could be of interest for city halls to model potential areas of noise complaints. We did not find such validation information for London and Madrid.

Our evaluation and validation for three different cities with varied physical layouts shows two important results. First, our methodology constitutes a good complement to model and understand in an affordable and near real-time manner land uses in urban environments. In fact, we have shown that residential, commercial and parks&recreation areas are well identified with coverages above 60%. Also, our approach is able to identify a land use, nighlife activity, not being considered up to now by city halls. This has implications from a planing perspective as this areas usually cause noise and security problems and can move over time.

Second, the Twitter density or average number of tweets per square kilometer appears to impact the accuracy of our land use identification approach. As reported, coverage percentages of Manhattan, with a Twitter density of $84.13/km_2$ are slightly higher than those for London and Madrid with densities of and $42.51/km_2$ and $10.88/km_2$, respectively. However, although the accuracy of the land use detection is slightly lower, the results still offer significant information for land urban planners.

5. Conclusions

With the deployment of pervasive infrastructures and the increasing use of geolocated user-generated content, urban planning will have a relevant and real-time source of information for characterizing urban dynamics. We have shown that geolocated tweets can constitute a good complement for urban planners to model and understand in an affordable and near real-time manner land uses in urban environments. We will continue to address this challenge by expanding our methodology to other sources of geolocated information exploring how to meaningfully combine multiple data sources for land use identification.

References

[1] S. Wakamiya, R. Lee, and K. Sumiya, "Urban area characterization based on semantics of crowd activities in twitter," in *GeoSpatial Semantics*, ser. Lecture Notes in Computer Science, C. Claramunt, S. Levashkin, and M. Bertolotto, Eds. Springer Berlin / Heidelberg, 2011, vol. 6631, pp. 108–123.

[2] T. Fujisaka, R. Lee, and K. Sumiya, "Exploring urban characteristics using movement history of mass mobile microbloggers," in *Proceedings of the Eleventh Workshop on Mobile Computing Systems & Applications*. ACM, 2010, pp. 13–18.

[3] S. Kinsella, V. Murdock, and N. Oare, "I am eating a sandwich in glasgow modeling locations with tweets," in *Proc. of the 3rd Workshop on Search and Mining User-generated Contents, Glasgow, UK*, 2011.

[4] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil, "Exploiting semantic annotations for clustering geographic areas and users in location-based social networks," in *3rd Workshop Social Mobile Web (SMW 2011)*.

[5] J. Cranshaw, R. Schwartz, J. I. Hong, and N. Sadeh, "The livehoods project: Utilizing social media to understand the dynamics of a city," *Association for the Advancement of Artificial Intelligence*, 2012.

[6] T. Kohonen, "The Self-Organizing Map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, Sep. 1990.

[7] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 1, no. 2, pp. 224–227, Apr. 1979.

[8] J. Shi and J. Malik, "Normalized cuts and image segmentation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 8, pp. 888–905, 2000.

[9] Twitter, "Open twitter streaming api," https://dev.twitter.com/docs/streaming-api.

[10] G. Voronoi, "Nouvelles applications des parametres continus a la theorie des formes quadratiques," *Journal fur die Reine und Angewandte Mathematik*.

[11] NYC, "Nyc open data," https://nycopendata.socrata.com/.

[12] London, "London open data," http://data.london.gov.uk/visualisations/atlas/ward-profiles-summary/atlas.htm.

[13] Madrid, "Madrid open data," http://www.madrid.org/cartografia/idem/html/web/index.htm.