Characterizing Urban Landscapes using Geolocated Tweets

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

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

***Abstract*—The pervasiveness of cell phones and mobile social media applications is generating vast amounts of geolocalized user-generated content. Since the addition of geotagging information, Twitter has become a valuable source for the study of human dynamics. Its analysis is shedding new light not only on understanding human behavior but also on modeling the way people live and interact in their urban environments. In this paper, we evaluate the use of geolocated tweets as a complementary source of information for urban planning applications. Our contributions are focussed in two urban planing areas: (1) a technique to automatically determine land uses in a specific urban area based on tweeting patterns; and (2) a technique to automatically identify urban points of interest as places with high activity of tweets. We apply our techniques in Manhattan (NYC) using 49 days of geolocated tweets and validate them using land use and landmark information provided by various NYC departments. Our results indicate that geolocated tweets are a powerful and dynamic data source to characterize urban environments.**

I. 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. As smartphones and data plans become more affordable, we are witnessing a worldwide shift towards mobile social media 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. The research presented in this paper focuses on understanding the usefulness of geolocated twitter datasets as a complementary information for urban planning applications.

Urban planning is a process that focusses on the control and on the design of urban environments in order to increase the well being of citizens. Two of the main processes concerning urban planing are the characterization of urban land use and the identification of urban landmarks. For that purpose, urban planners require, among other things, large amounts of data on urban land use and landmarks in order to make public policy decisions. Such information is typically gathered through direct observation or using questionnaires that attempt to capture how citizens interact with the 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 land use information through vision techniques. However, such techniques fail to provide real time information as images are not captured frequently. In order to overcome this issues, our research seeks a cost-effective approach to capture land uses and landmarks using the information provided by geolocated tweets.

The approach presented in this paper exclusively makes use of spatial (geo-tagged) and temporal (time-stamped) information, without accessing personal details or the content of the tweets. By doing so, our techniques preserve privacy and also can potentially be applied to any other mobile social media dataset with geolocation information. Our main contributions are: (1) a technique to automatically identify urban land uses *i.e.,* determine the type of activities that are most common in specific urban areas based on tweeting patterns; (2) a technique to automatically identify landmarks *i.e.* localize urban points of interest as places with high activity of tweets; and (3) a preliminary validation of our techniques in Manhattan(NYC) using 49 days of geolocated tweets and land use and landmark information provided by various NYC departments.

The rest of the paper is organized as follows: Section II presents related work in the characterization of urban land use and landmarks based on user-generated content. After that, we describe our technique to automatically identify land use and its evaluation in Manhattan in Sections III and IV, followed by our technique to detect landmarks and its evaluation in Sections V and VI. Finally, Section VII presents the conclusions and future research lines.

II. Related Work

The rise of location-based services, from social networks to microblogging sites, has opened a plethora of new research areas that take advantage of the location data. Researchers have explored how information propagates geographically [1], [2], have quantified influence across geographical areas [3], [4], have modelled trending topics in specific urban environments [5], and have studied the topological characteristics of the social networks that location-based services might create [6], [7].

Focusing on twitter, some authors have used geotagged datasets and its content to study and characterize human and crowd mobility. Wakamiya *et al.* [8] and Fujisaka *et al.* [9] 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. The authors propose models of aggregation and dispersion as a proxy to understand the bursty nature of human mobility. Similarly, Kinsella *et al.* [10] 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. Building on these results, we propose the use of twitter datasets to identify and characterize land uses and landmarks.

There exist interesting results using geotagged information from Foursquare and Flickr to model land use in urban environments. For example, Noulas *et al.* [11] have used the geolocated information provided by Foursquare to model crowd activity patterns in London and New York City using spectral clustering. The authors then 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. However, it's highly limited by the predefined categories described in Foursquare and it's not validated to understand the accuracy of the results. In a related work, Crandall *et al.* [12] used a dataset of geotagged photos from Flickr to perform landmark location throughout the world. The authors used the mean-shift algorithm to detect landmarks as areas with high numbers of geolocated pictures. The results were validated with an observational and qualitative approach that informally identified many of the landmarks as *well known* points of interest.

Our research builds on previous work and is similarly motivated. However, there are two significant novel contributions: (1) the use of geolocated tweets (without content and/or semantics) to automatically detect land use and landmarks, and (2) the validation of our results against *official* information on land use gathered by local governments, rather than using predetermined tags or evaluations based on popular wisdom.

III. Identifying Urban Land Uses

Urban land-use planning is a branch of public policy that focuses on regulating land use in an efficient way. Professional planners in the public and private sectors typically carry out research to understand land uses in the community under evaluation. Their main methods include public gatherings, questionnaires or GIS image analysis, among others. However, as mentioned earlier, such methods might involve high expenses as well as a lack of real time information. In the next two sections, we study the possibility of using geolocalized tweets to characterize urban land uses and explore whether these can be used as a complement to traditional land-use analytical approaches.

We present a method to automatically identify urban land uses from geotagged tweets using exclusively the spatial (localization) and temporal (timestamp) information. Our method consists of two main components: land segmentation and land use detection. Given that we want to identify land uses in different urban regions, we first need to partition the land into different segments (*land segmentation*), which can then be characterized by its tweet usage. The second component focuses on understanding common tweet uses across land segments and identifying how these behavioral patterns might relate to land use. The following two sections describe each phase in detail.

A. Land Segmentation with Geotagged Data

There are a variety of techniques that can be used to partition a geographical area into different land segments, ranging from administrative municipalities to grids or clustering. However, we seek a technique that preserves the topological properties of the geolocalized tweets, while respecting the actual shape of the geographical area under study. For that reason, we propose to use Self-Organizing Maps which have been shown to be very efficient for spatial clustering purposes [13], [14], [15], [16].

A Self-Organizing Map (SOM) is an unsupervised neural network (NN) that reduces the input data dimensionality to be able to represent its distribution as a map. As a result, SOM forms a map where similar samples are mapped close together and dissimilar apart. 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 in the time period under study. The SOM consists of a collection of N neurons where each neuron n is related to a weight vector $w(n)$ that represents the coordinates of the neuron in the map. Neurons are organized in a grid $[p, q]$, with $N = p * q$. The neurons are initially geolocalized at random within the boundaries of the map *i.e.*, the initial neuron weights are assigned randomly within its axis. During the SOM training, the neuron m most similar to a given geolocalized tweet x updates its weight. Similarly, neighboring neurons are also updated using a neighboring function $h(m, n)$. The training update rule is given by the following equation:

$$w_{t+1}(n) = w_t(n) + \alpha_t h_t(m, n)(x - w_t(n)) \qquad (1)$$

where α_t is the learning rate that decreases monotonically in time. The neighboring function $h_t(m, n)$ also decreases its influence in time and space: the further neuron n is from neuron m, the smaller will be the neighboring value, and thus less significant will be the update. For this reason, a common choice for the neighboring function is a Gaussian with its width parameter decreasing in time.

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

	Dataset	
	Total	Mean
World	24130423	492457.61
Manhattan	247381	5048.59

TABLE I
DATASET CHARACTERISTICS

	Weekdays		Weekends	
	Total	Mean	Total	Mean
Manhattan	184757	5278.77	62624	4473.14

TABLE II
DATASET CHARACTERISTICS FOR MANHATTAN CONSIDERING WEEKDAYS AND WEEKENDS

Davies-Bouldin clustering index [17]:

$$DB = \frac{1}{n}\sum_{i=1}^{n}\max_{i\neq j}\left(\frac{\sigma_i + \sigma_j}{d(w_i, w_j)}\right) \quad (2)$$

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.

At the end of the training, 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, we apply Voronoi tessellation considering the location of the neurons so as to compute the land segments that each neuron represents. These land segments are used as the elements for the characterization of land use.

B. Detecting Urban Land Uses

In order to detect different land uses in an urban region, we first characterize each land segment in the Voronoi tessellation by its average tweet activity. These tweet activities are then used to identify common land uses across land segments. Tweet-activity vectors are built following the ideas presented in [18], [19], where each land segment s is characterized by a vector X_s representing the average tweetting behavior as follows:

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 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 for land segment s is represented as the concatenation of weekday and weekend average activity vectors $X_s = \{X_{s,wkd}, X_{s,wkn}\}$ and is normalized as:

$$\hat{X}_s(t) = \frac{X_s(t)}{\sum_{t=1}^{72} X_{s,wkd}(t) + \sum_{t=1}^{72} X_{s,wkn}(t)}. \quad (3)$$

In the end, 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.

We use the activity vectors of all land segments to automatically identify and characterize urban land uses. In order to do so, we use the k-means algorithm to reveal clusters of common tweeting behaviors across land segments [20]. The land use of each cluster can be derived by analyzing the activity vectors of the regions comprised within the cluster. It is important to clarify that our research focuses on identifying the main land use of each cluster, although there might be other minor land uses associated to it. However, this is not a drawback of our method since land use maps computed by urban planners typically associate a unique land use to each region. Section IV will show evaluation details about how the method is used to identify and validate land uses in Manhattan.

Given that k-means depends on the initial random selected seeds and that it needs to specify beforehand the number of clusters k (land uses) to identify, we execute our method one hundred times for each value $k = 2, ..., 10$ and select the value of k that outputs the highest silhouette validity index [21]. The silhouette validity index is computed dividing a measure of intra-cluster similarity by a measure of inter-cluster dissimilarity. Since we seek well-separated clusters of similar samples, we aim to maximize the index to obtain the best partition of the data.

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 *i.e.*, $X_c = \frac{\sum_{s=1}^{m} X_s}{m}, c = 1, ..., k$ where m is the number of land segments in cluster c. Next section presents an evaluation of our method with tweeting activity from Manhattan and shows how to identify and validate land use.

IV. EVALUATION OF LAND USES IN MANHATTAN

In this section we first describe the twitter dataset we use to evaluate land use in Manhattan. Next, we describe how to apply our method to carry out land segmentation with the geolocated tweets and to identify clusters of common tweeting activity. We finish the section identifying possible land uses in Manhattan and validating our results against land use data retrieved from various open NYC datasets.

Fig. 1. Land segmentation process with Twitter: (left) data points, (center) centers of activity computed with SOM and (right) Voronoi tessellation.

A. Twitter Dataset

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 trough the GPS and not self-reported by the user. It is important to highlight that we are only interested in the spatial and temporal information of the tweets *i.e.,* latitude and longitude coordinates as well as timestamps. Thus, no personal identifiers or tweet content has been collected or is required to apply our method to identify urban land uses.

The process of collecting tweets was facilitated by the Twitter API. We used the Twitter Streaming API [22] 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. Finally, we relied upon the Tweepy [23] library for establishing the long-lived HTTP stream and for consuming the data received in JSON format. 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, we collected tweets worldwide mostly for sanity purposes. We observed that the dataset contained a considerable amount of fixed locations (probably GPS-enabled, non-mobile terminals) with large numbers of daily tweets. We posit that these might relate to mobile advertising companies sending commercial offers to mobile terminals. However, since these locations do not represent mobile users that can provide information regarding land use, we eliminate them from the dataset. In order to filter them, we apply the following filtering rule: any GPS location that generates more than 20 tweets per day is eliminated from the dataset (remember that for privay purposses we do not consider user identifiers, and as a result filtering is done at a GPS location level). As a result, approximately 10% of the tweets are eliminated.

Tables I and II show the general statistics for the dataset collected describing the total and average daily number of geotagged tweets worldwide and in Manhattan during the period under study. We can observe that Manhattan is responsible for approximately 1% of all the geolocalized tweets and that, on average, the tweeting activity in Manhattan is higher in the week than during the weekends. Finally, Figure 1(left) shows the geographical representation of all the tweets in our Manhattan dataset where each dot represents a geolocalized tweet.

k	2	3	4	5	6	7	8	9	10
S	0.488	0.496	**0.506**	0.491	0.471	0.457	0.451	0.455	0.463

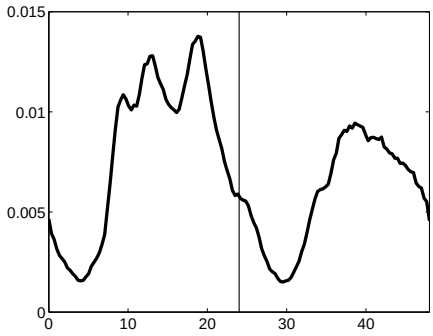
TABLE III
MEAN SILHOUETTE VALUES

B. Land Segmentation and Land Use Clustering

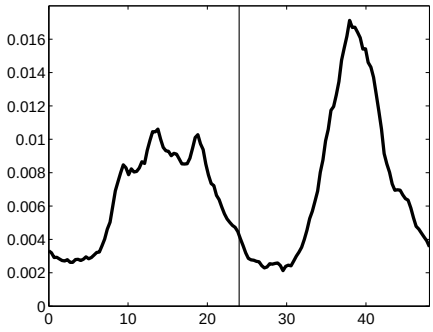
As explained earlier, our method first trains a SOM with the geolocalized tweets to divide Manhattan into different land segments s characterized by their tweet activity vector X_s. The SOM is trained with a varying number of neurons N and the value with the minimum DB index is selected as the best distribution on neurons (centroids of land segments). Since SOMs preserve the geographical information, neurons N must be geolocated within the area of Manhattan. For this case we considered N in the range $N = [10, ..., 100]$, 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). Given the rectangular shape of Manhattan, we only consider cases in which $p > q$. For example, $n = 10$ 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 pair (p, q) with $N = p * q \in [10, ..., 100]$ and their average DB index is computed. The minimum DB index obtained has a value of 0.3569 and is associated to $N = 64$ neurons with with $p = 16$ and $q = 4$, which adapts nicely to the geographical shape of Manhattan. Figure 1(center) shows the 64 SOM centroids after the training process. 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 [24] to each SOM centroid in the two-dimensional space as shown in Figure 1(right). The final land segmentation consists of 64 land segments. Each segment 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.

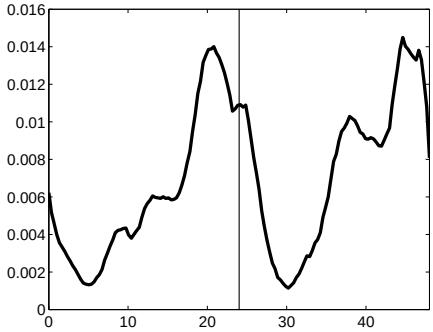
Next, our method uses the 64 X_s vectors to identify different land uses in Manhattan. For that purpose, it executes k-means to cluster land segments with similar activity vectors that could be associated to a common land use. Specifically, it executes k-means with k values in the range $[2, ..., 10]$ and selects the k with the largest mean silhouette value. Table III displays the mean silhouette validity index for each k. We observe that the best land segment clusters are computed for $k = 4$ which reveals four well differentiated land uses in Manhattan. Figure 2 presents the class representatives for these four land uses. Figures 3(a), 3(b), 3(c) and 3(d) presents the geographical representation of the land use clusters.



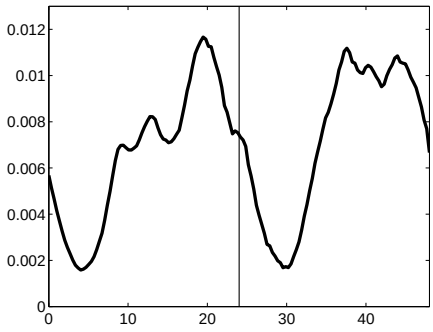
(a) Business Cluster



(b) Leisure/Weekend Cluster



(c) Nightlife Cluster



(d) Residential Cluster

Fig. 2. Tweeting activity signatures per cluster, where the Y axis represents the normilized 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.

Fig. 3. Geographical representation of Land Use Clusters: (a) Business, (b) Leisure/Weekend, (c) Nightlife and (d) Residential

Considering the cluster maps and the activity vectors, we can provide some hypothesis about the potential types of land use discovered:

- **Cluster 1** (Figures 2(a) and 3(a)): The activity vector of this cluster is characterized by a higher tweetting activity during weekdays than weekends. During weekdays the highest tweeting activity is reached at 9:30PM, 13:00PM and 8:30PM, which might be associated to the times at which people typically get to work, go for lunch, and leave work. The peak of the tweeting activity during the weekends is reduced by 33% when compared to weekdays.

Looking at the geolocation of the cluster in Figure 3(a), we observe that it includes areas like Battery Park and Wall street, and moving further north areas around Mid-Town. For these reasons, we hypothesize that the geographical area covered by this cluster might represent Business areas in Manhattan.

- **Cluster 2** (Figures 2(b) and 3(b)): This cluster shows almost twice as much tweeting activity during weekends that during weekdays. During weekends tweeting activity constantly increases until 16.00PM, when it peaks, and then constantly decreases. The geographical representation of this cluster covers Central Park and its surrounding areas, including the main NYC museums such as the Guggenheim or the Metropolitan Museum. Thus, we hypothesize that this cluster might 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.

- **Cluster 3** (Figures 2(c) and 3(c)): Unlike the first two clusters, this activity vector shows the highest peaks of activity at night. On weekdays the tweeting activity increases until a maximum is reached at 20:00PM. On weekends, we observe two peaks, a smaller one between 18:00PM and 23:00PM and the largest peak that happens between 00:00 and 04:00. This second peak shows a tweeting activity that doubles the tweeting activity of any other cluster for the same time period.

The geographical representation of this cluster (see Figure 3(c)) mostly focuses in the surroundings of the East Village and Broadway shows. For these reasons, we hypothesize that these tweeting behaviors might be associated to nightlife leisure activities, which during the week happen earlier (20:00PM) and during the weekends go on until late at night (04:00AM).

- **Cluster 4** (Figures 2(d) and 3(d)): This cluster has a signature that shows an almost constant tweeting activity between 10:00AM and 22:00PM during the weekends. During weekdays, we also observe a constant activity from 10:00AM until 16:00PM after which the tweeting activity increases until reaching a peak at 20:00. Figure 3(d) shows that the cluster mostly covers areas in the Upper-East and Upper/Lower-West sides. Thus, we hypothesize that this cluster might represent a residential land use, where people stay home during the weekends, and mostly return from work at night showing a

peak Twitter activity later inthe day (around 20:00).

As explained in Section III, we focus 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 maps.

C. Land Use Validation

In order to validate our land use hypothesis, we compare our 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 [25]. This initiative provides a catalog with hundreds of datasets of public data produced by City agencies typically through a combination of on-site inspections, interviews and questionnaires.

Figure 4 depicts the official land uses at a block level in Manhattan [1]. The NYC Department of City Planning considers four main land use types: (1) residential, (2) commercial, (3) industrial and (4) parks&recreation. Visually speaking, we want to understand the percentage of overlapping that exists between our land use clusters in Figures 3(a), 3(b), 3(c) and3(d) and the *official* land use areas declared by the NYC Departments in Figure 4. Such overlapping will give us an understanding of the accuracy that Twitter activity achieves in identifying land uses *i.e.*, the larger the overlapping areas, the more accurate tweeting activity is in modeling land use. It is important to highlight that the percentage of overlapping is an approximate measure to validate land use identification given that both maps have different granularities: our maps represent land segment clusters based on Voronoi and the density of tweets, whereas the NYC maps show data at a block level. However, we believe it constitutes a good preliminary approach to validate our results.

In order to analyze overlapping areas, we use ArcGIS [27]. ArcGIS allows, among many other GIS functions, to evaluate overlapping between the shapefiles of two given regions. In our case, one shapefile will represent the land use cluster we have obtained and the other one an official land use in the NYC Open Data map. The official land use areas are distributed by the NYC Open Data in the shapefile format; whereas our land use clusters are transformed from their latitude and longitude coordinates into shapefiles also using ArcGIS. Table IV shows the percentages of overlap between the official land uses (rows) and our land use hypothesis (columns). Specifically, each element (i, j) in the table represents the percentage of the official land use region $i = Commercial, Residential, Industry, Parks$ that is covered by one of our land use clusters $j = Business, Residential, Nightlife, Leisure$ (bear in mind that since our Voronoi tessellation does not precisely cover all Manhattan land, the percentages do not exactly need to sum up to 100%).

We observe that the official Commercial land use is identified with a coverage of 77% by our Business cluster. It also includes other minor land uses detected with Twitter like residential with a 14% of overlap. The official Residential land use is also well modeled by our Residential cluster, with an overlap area of approximately 62%, although it is also covered by a 14% of Nightlife land use. Focussing on the official Industrial land use, we see that in this case there is a strong overlap with our Residential land use. It seems that our method, using Twitter data, is unable to model Industrial land use which goes completely undetected and is a result is included within the Residential land cluster. This is probably due to the difference in granularity between the two land use maps: given that industrial areas in Manhattan are typically long, narrow and next to Residential land uses, it is harder for our clusters to separate them. Additionally, it might also be the case that workers in the industrial areas are not using Twitter and thus our method only captures the activity of citizens in the residential areas intertwined with the industrial zones. In order to clarify these issues, we plan to carry out future work to model user Twitter profiles and understand better tweeting behaviors based on job and location factors. Finally, the official Parks&Recreation land use is identified by our Leisure cluster

[1]Zoning map plotted with Oasis, an online free tool developed by the Graduate Center at CUNY [26]

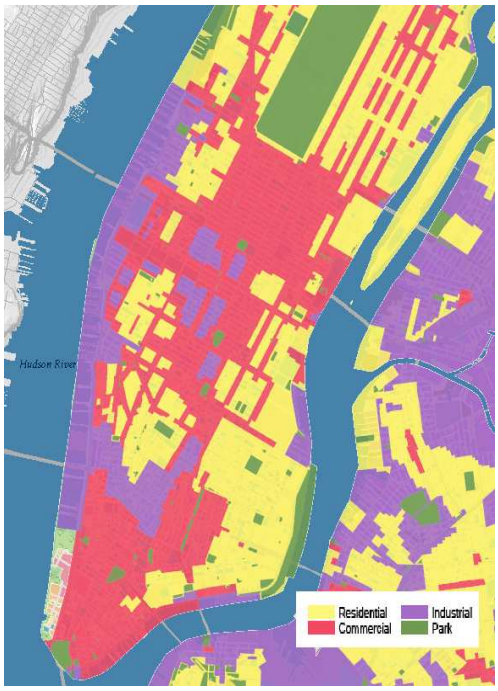


Fig. 4. Official Land Uses from NYC Department of City Planning: Commercial, Residential, Industrial and Parks&Leisure

Official Land Use	Twitter Land Use			
	Business	Residential	Nightlife	Leisure&Weekend
Commercial	**77%**	14%	4%	4%
Residential	8%	**62%**	22%	8%
Industry	7%	85%	**0%**	6%
Parks&Recreation	5%	6%	6%	**79%**

TABLE IV
PERCENTAGE OF OVERLAP BETWEEN OFFICIAL LAND USES AND TWITTER LAND USES.

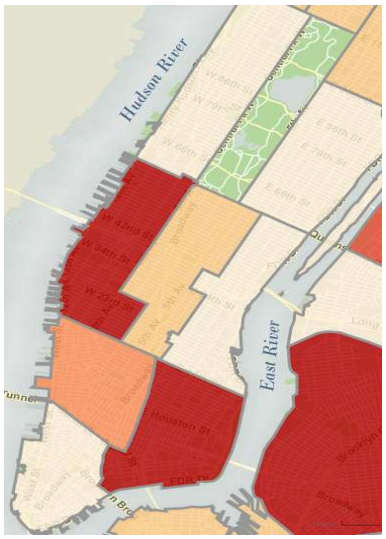


Fig. 5. Community Districts with Noise Complaints from the NYC 311 Service (red represents the highest number of complaints)

with an overlap of a 79%, although minor land uses are also included.

On the other hand, our method identified a Nightlife cluster which mostly overlaps 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 NYC Departments. Figure 5 displays the number of noise complaints per community district made to the 311 on-line service during 2010, where darker colors imply higher number of complaints. We observe that the two community districts with the highest average number of complaints (plotted in red) correspond to geographical areas covered by our Nightlife cluster. Given that the community districts have much lower granularity than our land use clusters, we compute 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.

To conclude, 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 have shown preliminary results for the identification of Residential, Commercial and Park&Leisure land uses. Although our method has failed to identify Industrial land uses, it has the ability to detect Nightlife land uses that might be of significance for city halls attempting to model sources of noise complaints.

V. Landmark Identification

Urban landmark identification constitutes an important component of policy making for Landmark Preservation Committees (LPC) or for Transportation Departments. Urban planners typically build maps of historic, popular or tourist areas and propose a wide range of policies for their preservation. Such policies might include walking tours through historic districts to promote awareness or the improvement of transportation routes to an urban area popular among citizens. Landmark maps are typically built gathering data from questionnaires or interviews with district commissions and local organizations, which has both time and cost limitations. In an attempt to help urban planners, we evaluate the usefulness of geolocated Twitter activity to identify historic and/or popular urban landmarks. The advantage of using Twitter as opossed, for example, to Flickr, is that, while Flickr is heavily influenced by visitors/tourists, Twitter is used by the population at large, thus facilitating the identification of more landmarks.

In this section, we propose a method to identify urban landmarks as areas with very high tweeting activity. In order to compute these areas, we need to use a clustering technique capable of detecting local maxima (landmarks) over a non-parametric distribution of geolocated tweets. Although one could apply techniques like k-means or DBSCAN, these have the limitation that knowing beforehand the number of clusters (urban landmarks) is, in general, not possible. In fact, unlike land uses in an urban area which might account for a few, urban landmarks can go anywhere from a few to a few hundreds. For that reason we use mean-shift, a clustering technique that does not require to specify the number of clusters beforehand [12].

A. Mean-shift Algorithm

Mean-shift is a non-parametric clustering technique that detects the modes of an underlying probability distribution from a set of discrete samples. As such, mean-shift can be used both as an algorithm to detect local maxima (modes) as well as a clustering technique (areas associated to the modes). In our setting, we assume that there exists an unobservable underlying probability distribution of where people tweet from. The modes of that distribution are determined to represent urban landmarks or points of interest in the city. Specifically, mean-shift estimates the gradient of the probability distribution from the set of tweets using a kernel function K and a bandwidth δ. The bandwidth represents the scale of observation *i.e.,* the scale associated to the spatial information of the samples. As such, larger values of δ will identify clusters that cover large geographical areas which could be associated to popular cities; whereas smaller values will identify clusters that cover smaller areas which could be associated to landmarks within a city.

Initially, mean-shift designates a given location x as the maximum and iteratively updates it following the direction given by the vector $m_{\delta,K}(x)$, which always points towards the direction of highest gradient. The procedure iteratively updates x until $m_{\delta,K}(x)$ converges to zero, and x is labeled as a maximum. All the points visited in the gradient ascent are marked as belonging to that maximum.

$$m_{\delta,K}(x) = \frac{\sum_{i=1}^{n} x_i k\|(x-x_i)/\delta\|^2}{\sum_{i=1}^{n} k\|(x-x_i)/\delta\|^2} \quad (4)$$

$$x^{(i+1)} = x^{(i)} + m_{\delta,K}(x^{(i)}) \quad (5)$$

	Weekdays			Weekends		
Rank	Places	Tweets	(lat, lon)	Places	Tweets	(lat, lon)
1	Penn Station	3532	40.750480, -73.993457	NYU	1053	40.728802, -73.999840
2	Rockefeller	2407	40.758597, -73.979010	Rockefeller	775	40.758616, -73.978972
3	NYU	2386	40.728801, -73.999828	Times Sq	755	40.757869, -73.985721
4	Times Sq	2178	40.756342, -73.986366	Penn	505	40.750123, -73.992414
5	Union Sq	1681	40.734215, -73.990600	Union Sq	577	40.736538, -73.990566
6	Herald Sq (Empire State)	1663	40.749757, -73.987731	MSG	559	40.750510, -73.993499
7	Theater District	1395	40.821546, -73.933991	Herald Sq	465	40.749761, -73.987992
8	Apple Store	1357	40.763919, -73.973101	Theater District	456	40.821546, -73.933991
9	Columbus Circle	1327	40.763919, -73.973101	Meatpacking	448	40.741375, -74.005089
10	East Village	1256	40.731259, -73.988741	Grand Central	446	40.752750, -73.977263

TABLE V
RANKING OF NYC LANDMARKS WITH HIGHEST NUMBER OF GEOTAGGED TWEETS FOR WEEKDAYS AND WEEKENDS.

Mean-shift algorithm is run for a set of different initial positions in order to identify all local maxima. At the end of the process, every geolocated tweet is assigned to a local maxima and a cluster representing a potential urban landmark. The larger the number of tweets assigned to a cluster, the highest the tweeting activity for that landmark. As a result, mean-shift algorithm applied to Twitter activity produces a list of local maxima/clusters which, if ranked according to number of tweets, represents the list of the most popular landmarks in the city.

VI. EVALUATION OF LANDMARKS IN MANHATTAN

In this section we evaluate whether we can identify the urban landmarks of Manhattan using geolocated tweets and validate our results against official data collected by the NYC Landmark Preservation Commission.

A. Landmark Detection

In order to detect Manhattan landmarks we apply the mean-shift landmark detection method to the Twitter dataset described in Section IV.*A*. In order to be able to explore landmarks at an urban scale, we set the bandwidth parameter to 0.001^{o}, which corresponds to approximately ≈ 85 meters at that latitude. Finally, we start mean-shift with 1000 randomly selected geolocations and iterate until convergence is reached.

Table V shows the output of the mean-shift algorithm applied to weekend and weekday geolocated tweets. It represents the top ten landmarks identified in Manhattan ranked by the largest amount of tweets during weekdays and weekends, respectively. Additionally, the table shows the number of tweets assigned to each location and its coordinates in WGS 84 format. We observe that the Penn Station area is detected as a popular landmark during weekdays, which seems logical given that it represents one of the most important commuting centers for thousands of New Yorkers on a daily basis. Additionally, the Rockefeller Center or Herald Square (Empire State Building area) which represent important historic and popular locations, make it to the top of the list. On weekends, we observe a shift whereby commercial/leisure areas like Times Square or the Meatpacking district are ranked higher when compared to weekend landmarks.

B. Landmark Validation

In order to validate our results (beyond common sense), we compare the top 50 landmarks detected with mean-shift against the official Manhattan landmark list retrieved from the NYC Open Data website. Note that the official list of landmarks is not ordered and does not give any indication of the relevance of the landmark, its importance or any other ranking factor. The data, collected by the Landmark Preservation Commission (LPC), covers over 800 locations grouped into historic districts like Greenwich Village (NYU area), Madison Square North or Herald Square (Penn Station area) [28]. NYC Open Data provides a list with the landmarks, their geolocation and the shapefiles for the different historic districts. Figure 6 shows the Manhattan landmarks focussing on the area south of Central Park.

The number of official landmarks co-located with our top 50 landmarks will determine the accuracy of the mean-shift method to reveal urban points of interest. Since the latitude and longitude of our landmarks represent the center of the cluster detected by mean-shift , we compute a *robust* location drawing an $100m$-diameter circle around each of our landmarks. Next, we determine the number of official landmarks that fall into the diameter of any of our detected landmarks. Following this procedure, approximately a 17% of the official landmarks are detected by the mean-shift algorithm.

In order to understand the types of landmarks that our method misses, we explored the official landmark list in depth. We observed that most of the landmarks that go undetected represent historic buildings that, although protected by the LPC, do not necessarily represent popular or tourist points of interest. For example, we detect Grand Central which is both a historic building and a popular landmark. However, we do not detect the Manhattan House Block (East $65th$), a 1947 New York Life Insurance Company building which occupies a full block but does not draw much attention. Additionally, we also detect new urban points of interest that are not considered historic or touristic by the LPC like the Meatpacking District, but which attract many new yorkers specially during the weekends. Thus, it is fair to say that Twitter activities can be used to detect urban popular/tourist landmarks.

To further support this point, we checked our top landmarks

Fig. 6. Historic Manhattan Landmarks determined by the Landmark Preservation Commission south of Central Park.

against the list of NYC landmarks published by Crandall *et al.* [12]. As explained in the related work, Crandall used the mean-shift algorithm to perform landmark location based on Flickr geotagged photos. The authors identified the top seven landmarks in NYC including the Empire State Building or Times Square among others. Our Manhattan landmarks in Table V include all the top ones detected by Crandall except for Liberty Island, which we have not considered as part of Manhattan. This analysis confirms the accuracy of our landmark detection and shows that Twitter appears to be as good as Flickr for the detection of touristic landmarks.

To sum up, we have shown that our mean-shift method can detect popular landmarks. However, it misses most of the historic landmarks outside the popular or tourist routes given that these probably do not receive a critical mass of tweeting visitors. We believe that our method might provide an affordable and near real-time tool for authoritative bodies to detect urban spots that are becoming landmarks.

VII. Conclusions and Future Work

We have presented techniques to automatically identify urban land uses and landmarks from geolocated information. Although our experiments have focussed on using Twitter datasets, our methods are potentially applicable to any user-generated dataset with geolocation information. Our results have shown that Twitter data can help urban planners to characterize commercial, leisure and residential areas, as well as to model new types of urban uses like Nightlife. In terms of landmarks, we have shown that our technique can help urban planners identify popular/tourist landmarks, although historic landmarks go highly undetected. Both elements have been validated with information collected by the NYC Open Data Initiative. Future work will evaluate our techniques on different geolocated datasets so as to understand the limitations and applicability of our methods at large.

References

[1] K. Lerman and R. Ghosh, "Information contagion: an empirical study of the spread of news on digg and twitter social networks," in *Proceedings of the fourth ACM international conference on Web search and data mining*, ser. ICWSM '10, 2010.

[2] J. An, M. Cha, K. P. Gummadi, and J. Crocroft, "Media landscape in Twitter: A World of New Conventions and Political Diversity," ser. ICWSM '11, Jul. 2011.

[3] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi, "Measuring User Influence in Twitter: The Million Follower Fallacy," ser. ICWSM '10, May 2010.

[4] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts, "Everyone's an influencer: quantifying influence on twitter," in *Proceedings of the fourth ACM international conference on Web search and data mining*, ser. WSDM '11, 2011, pp. 65–74.

[5] M. Naaman, H. Becker, and L. Gravano, "Hip and trendy: Characterizing emerging trends on Twitter," *J. Am. Soc. Inf. Sci.*, vol. 62, no. 5, pp. 902–918, 2011.

[6] H. Kwak, C. Lee, H. Park, and S. Moon, "What is twitter, a social network or a news media?" in *Proceedings of the 19th international conference on World wide web*, ser. WWW '10, 2010, pp. 591–600.

[7] D. Boyd, S. Golder, and G. Lotan, "Tweet, tweet, retweet: Conversational aspects of retweeting on twitter," in *Proceedings of the 2010 43rd Hawaii International Conference on System Sciences*, ser. HICSS '10, 2010, pp. 1–10.

[8] 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.

[9] 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.

[10] 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.

[11] 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)*.

[12] D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg, "Mapping the world's photos," in *Proceedings of the 18th international conference on World wide web*, ser. WWW '09, 2009, pp. 761–770.

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

[14] M. K. Allouche and B. Moulin, "Amalgamation in cartographic generalization using kohonen's feature nets," *International Journal of Geographical Information Science*, vol. 19, no. 8-9, pp. 899–914, 2005.

[15] M. Sester, "Optimization approaches for generalization and data abstraction," *International Journal of Geographical Information Science*, vol. 19, pp. 871–897, 2005.

[16] ——, *Self-Organizing Maps for Density-Preserving Reduction of Objects in Cartographic Generalization*. John Wiley and Sons, Ltd, 2008, pp. 107–120.

[17] 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.

[18] V. Soto and E. Frias-Martinez, "Robust land use characterization of urban landscapes using cell phone data," in *The First Workshop on Pervasive Urban Applications (PURBA)*, 2011.

[19] ——, "Automated land use identification using cell-phone records," in *The 3rd ACM International Workshop on Hot Topics in Planet-Scale Measurement (HotPlanet)*, 2011.

[20] J. Hartigan and M. Wong, "A k-means clustering algorithm," *Applied Statistics*, vol. 28(1), pp. 100–108, 1979.

[21] P. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, no. 1, pp. 53–65, Nov. 1987.

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

[23] Tweepy, "Open twitter streaming api," http://code.google.com/p/tweepy/.

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

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

[26] CUNY, Graduate, and Center, "Oasis maps," http://www.oasisnyc.net/.

[27] ESRI, "Arcgis," http://www.esri.com/software/arcgis/index.html.

[28] NYC, Open, and Data, "Landmarks preservation commission," http://www.nyc.gov/html/lpc/html/maps/maps_manh.shtml.