Automated Land Use Identification using Cell-Phone Records

Víctor Soto
Telefónica Research
Madrid, Spain
vsoto@tid.es

Enrique Frías-Martínez
Telefónica Research
Madrid, Spain
efm@tid.es

ABSTRACT

Pervasive large-scale infrastructures generate datasets that contain human behavior information. In this context, cell phones and cell phone networks, due to its pervasiveness, can be considered sensors of human behavior and one of the main elements that define our digital footprint. In this paper we present a technique for the automatic identification and classification of land uses from the information generated by a cell-phone network infrastructure. Our approach first computes the aggregated calling patterns of the antennas of the network and, after that, finds the optimum cluster distribution to automatically identify how citizens use the different geographic regions within a city. We present and validate our results using cell phone records collected for the city of Madrid.

Categories and Subject Descriptors

H.4 [**Information Systems Application**]: Miscellaneous; J.4 [**Social and Behavioral Sciences**]: Sociology—*Smart Cities, Urban Planning*

General Terms

Human Factors

Keywords

Land Use, Clustering, Classification, Call Details Records

1. INTRODUCTION

With the increasing capabilities of mobile devices, individuals leave behind footprints of their interaction with the urban environment. As a result, new research areas, such as urban computing and smart cities, focus on improving the quality of life in an urban environment by understanding the city dynamics through the data provided by ubiquitous technologies [6, 2].

HotPlanet'11, June 28, 2011, Bethesda, Maryland, USA.

Traditionally, urban analysis and the study of urban environments have used data obtained from surveys to characterize specific geographical areas or the behavior of groups of individuals. However, new data sources (including GPS, bluetooth, WiFi hotspots, geo-tagged resources, etc.) are becoming more relevant as traditional techniques face important limitations, mainly the complexity and cost of capturing survey data. One of the new data sources relevant for the study of urban environments are cell phone records, as they contain a wide range of human dynamics information (ranging from mobility, to social context and social networks) that can be used to characterize individuals or geographical areas.

In this paper we present an application of data mining techniques to automatically identify the uses that individuals make of the different parts of a city using the information contained in cell-phone records. Although our analysis focus on call-detail records the same methodology can be applied using other ubiquitous data sources such as geo-located data provided by Flickr or twitter, or logs collected by any location-based service. Our system is relevant for a variety of urban planning applications, mainly urban zoning. In the context of urban planning, urban zoning is defined as the designation of permitted uses of land based on mapped zones which separate one set of land uses from another (for example residential areas from industrial areas). One of the main problems of zoning is to actually evaluate to which extent the areas are being used as required or planned, because the collection of data has to be done on site. Our approach allows to compare the planned used of a city with the actual use that citizens give to the different areas of the city without the need of on-site data collection.

Some authors have already used cell phone traces to implement urban analysis studies. Among others, Ratti et al. [7] used aggregated cell-phone data to analyze urban planning in Milan, Eagle et al. [4] identified behavioral patterns from the information captured by phones carrying logging software, and in [1] the authors use bluetooth to characterize pedestrian flow data. Previous work on the use of the cell phone data for land use analysis is scarce, although some authors have presented studies to solve related questions. For example, [10] monitors the dynamics of Rome and obtains clusters of geographical areas measuring cell phone towers activity using Erlangs (being 1 Erlang 1 person using the phone for 1 hour). Another study is described in [5], where the authors analyze four different geographical spots at different times in Bangkok. Related to our work, Reades et al. [9] used eigendecomposition to study the time structure,

finding correlations between the number of Erlangs and the commercial activity of the area. To the best of our knowledge though, no previous publications attempted to study the problem of identifying and classifying the present land uses of the entirety of two urban cities from a CDR dataset with millions of interactions.

The paper is organized as follows: first we formally define the problem and the basic elements of a cell phone network. After that, we define the concept of land signature and define how to obtain it from the data captured by a cell phone network. The following section, applying unsupervised clustering, identifies the main land uses which are then validated using the city of Madrid. After that, the knowledge extracted is used to automatically classify land uses in the city of Barcelona. The paper finishes with future work.

2. PROBLEM DEFINITION

In order to automatically identify land use behaviors we present a technique that makes use of the information extracted from cell phone networks. Cell phone networks are built using base transceiver station (BTS) towers that are in charge of communicating cell phones with the network. A given geographical region will be serviced by a set of BTSs $BTS = \{bts_1, \ldots, bts_N\}$, each one characterized by its geographical coordinates (lat,lon). For simplicity, we assume that the area of coverage of each BTS can be approximated by a 2-dimensional non-overlapping polygon, and approximate it using Voronoi tessellation.

Call Detail Record (CDR) databases are populated whenever a mobile phone makes/receives a call or uses a service (e.g. SMS, MMS). Hence, there is an entry for each interaction with the network, with its associated timestamp and the BTS that handled it, which gives an indication of the geographical location of the mobile phone at a given moment in time. Note that no information about the position of a user within a cell is known. The set of fields typically contained in a CDR include: (a) originating encrypted phone number; (b) destination encrypted phone number; (c) identifier of the BTS that handled the originating phone number (if available); (d) identifier of the BTS that handled the destination phone number (if available); (e) date and time of the call; and (f) duration of the call.

Using the information contained in a CDR database generated from the BTS towers that give coverage to a city, we can characterize the use that citizens make of specific urban areas. In order to do so, the city is initially divided into the coverage areas defined by the Voronoi tessellation. Each area is then characterized by the activity associated to its corresponding BTS which is measured as the number of calls per time unit. This measure will be the *signature* of the BTS tower. Once all the signatures have been computed, they are clustered and class representatives of the land uses are identified. These representatives can also be used to automatically classify land use in other cities.

The study we present has been done using CDR data collected from the city of Madrid for a period of 1 moth, from October 1st 2009 to October 31st 2009. The area covered by the city is of 400 km^2, with more than 3 million inhabitants, and is served by 1100 towers that collected over 100 million interactions. Madrid will be used to identify and characterize land uses, knowledge that then will be used to automatically classify the land use for other cities.

3. OBTAINING LAND USE SIGNATURES FROM CDR DATA

We define the activity of a BTS, and by extension of its area of coverage, as the number of calls that are managed by that BTS over a given period of time. The activity associated to each BTS is represented as a matrix $v_n(\delta, \tau)$, with $n \in \{1, \ldots, N\}$ being the BTS identifier, $\delta \in \Delta = \{1, \ldots, 31\}$ representing each day and $\tau \in \{1, \ldots, 288\}$ representing each of the 5-minutes time slot of the day in which the measure of the number of calls is taken.

The signature X_n of each bts_n is defined by different aggregations of the information contained in the associated activity matrix v_n. Three types of aggregation have been studied: (1) total aggregation; (2) weekday-weekend aggregation and (3) daily aggregation. The total aggregation defines the signature of each BTS as the average value of the activities reported for each time slot τ throughout all days δ. Thus, each element $X_n(\tau)$ is computed as

$$X_n(\tau) = \frac{1}{\|\Delta\|} \sum_{\delta \in \Delta} v_n(\delta, \tau) \quad (1)$$

It is well known that human dynamics are very different between weekdays and weekend days [3], and those differences will translate into BTSs use. The weekday-weekend representation aggregates the calling activities reported for each time slot τ for two different types of days: weekdays (Monday through Friday, contained in Ω_1) and weekend days (Saturday and Sunday, contained in Ω_2). The final signature is represented as the concatenation of both components (represented as $+\!\!+$):

$$X_{n,\Omega_i}(\tau) = \frac{1}{\|\Omega_i\|} \sum_{\delta \in \Omega_i} v_n(\delta, \tau) \quad (2)$$

$$X_n = X_{n,\Omega_1} +\!\!+ X_{n,\Omega_2} \quad (3)$$

where $\Omega_i \subset \Delta$, $\Omega_1 \cap \Omega_2 = \emptyset$ and $\Omega_1 \cup \Omega_2 = \Delta$.

Finally, the daily aggregation can be considered as an extension of the weekday-weekend aggregation, in which each day of the week has its own component. Formally it is computed with Equation (2) but considering Ω_{day} with day $\in \{mon, tue, wed, thu, fri, sat, sun\}$. All signatures were normalized before identifying land uses.

Figure 1 graphically presents three examples of each one of the aggregation types. Figure 1(a) shows the typical curve for total aggregation with one peak around 12AM and another one at around 7PM. In this particular case because the 12AM peak is smaller than the 7PM peak the signal probably represents a residential location where individuals come home after work. Figures 1(b) and (c) present the same idea but concatenated 2 times and 7 times respectively representing weekday-weekend and daily aggregations.

4. AUTOMATIC IDENTIFICATION OF LAND USE

Once each area of coverage has been characterized with its BTS signature, k-means was applied to identify land uses considering the three aggregations presented in the previous section. Unsupervised clustering techniques require the number of clusters to be known beforehand. In the case of k-means, the different techniques used to determine an optimum number of clusters are based on identifying the clustering which results in compact clusters which are well

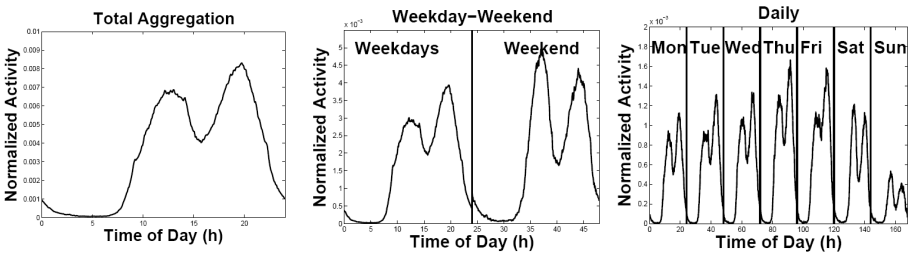


Figure 1: (Example of (left) total,(center) weekday-weekend and (right) daily aggregation

separated. From the variety of techniques available (Davies-Bouldin index, Dunn's index, etc.), we have chosen the validity measure presented in [8] as it is shown to outperform the previous techniques in a variety of applications. In this case, the quality of the partition, measured by the cluster validity measure, is given by computing the ratio between the inter-cluster and intra-cluster distances of the partition for each one of the k values (number of clusters) run. An ideal partition will minimize the intra-cluster distance and maximize the inter-cluster distance, therefore the best partition will be that which minimizes the proposed measure. Given the nature of the k-means algorithm, where every cluster is defined by its centroid c_i, the intra-cluster and inter-cluster distances are formally defined as:

$$intra - cluster = \frac{1}{N}\sum_{i=1}^{k}\sum_{X_n \in C_i} \|X_n - c_i\|^2 \quad (4)$$

$$inter - cluster = \min_{i \neq j} \|c_i - c_j\|^2 \quad (5)$$

where C_i is the set of signatures that belong to the cluster defined by centroid c_i. Due to the stochastic nature of k-means we executed the algorithm 500 times for each k and kept the results with better (minimum) validity values.

K-means clustering was implemented using two distance measures: Euclidean and a Dynamic Time Warping (DTW) based metric. DTW is specially useful when dealing with signals that can have slight time shifts, but experimental evidence showed that better index values were obtained with the Euclidean distance. Also the computational cost was considerably higher when using DTW.

5. CLUSTERING RESULTS

Clustering validity values were used to identify which combination of distance metric, land signature and number of clusters produced better results. Table 1 presents the minimum validity values when using k-means, with Euclidean distance and the total, weekend-weekday and daily aggregation for $k \in \{3, \ldots, 8\}$. In all cases the best value was obtained when using the weekday-weekend aggregation. The values of the clustering coefficients indicate that while the total aggregation loses information (it has higher values in all cases), the daily aggregation does not add any extra insight when compared with the weekday-weekend representation (index values are higher or similar). Within the weekday-weekend representation the top two coefficients are obtained for $k = 3$ and $k = 5$ respectively. The case $k = 3$ is obvious as our signatures are characterized by two peaks at each 24 hour period. Simplifying the interpretation, with $k = 3$ we observe three differentiated land uses: (1) when the first peak (12AM) is higher than the second one (7PM), which will indicate an area with commercial and/or industrial activity; (2) when the second peak is higher than the fist peak (a residential area) and (3) when the two peaks have the same height (a mixed used area). Nevertheless, we are interested in identifying a variety of land uses apart from the obvious ones identified by $k = 3$. The second case, $k = 5$ is much more interesting, as intermediate land uses are found. The rest of the paper focusses on analyzing the clusters obtained with k-means, using Euclidean distance, a weekday-weekend aggregation and $k = 5$.

	Number of clusters k					
	3	4	5	6	7	8
Wkday-Wknd	0.20	0.60	0.54	0.60	0.68	0.67
Daily	0.26	0.60	0.59	0.69	0.70	0.75
Total	0.30	0.84	0.80	0.81	0.84	0.85

Table 1: Cluster validity index

Figure 2 presents the signatures of the class representatives for the five land uses found. Obviously, these behaviors are heavily influenced by cultural characteristics, and as such, although similar land uses could be identified for other countries, the signatures would be shifted according to cultural and social routines. Also, Madrid is a traditional European city in the sense that the old part of the city concentrates commercial, office and residential areas, which should difficult the identification of land uses. The clusters are characterized by:

- Cluster 1: This cluster is characterized by the fact that the activity takes place mainly during weekdays, especially in working hours, and weekend activity is almost nonexistent. During weekdays the activity is heavily focused between 10AM and 14PM and another peak of activity between 16:00 and 19:00 hours. This cluster shows a clear work related activity and the hypothesis is that the BTS coverage areas included in this cluster are used as industrial parks and/or office areas.

- Cluster 2: The second cluster probably represents a hybrid land use. During weekdays there is activity during commercial hours, with two peaks of similar normalized intensity (0.005 and 0.004) at around 12AM and 19PM. There is a relevant weekend activity although of less intensity (both peaks at around 0.003). This behavior could be considered commercial. The lesser intensity of both peaks during weekends is probably caused because shops in Spain are generally closed on

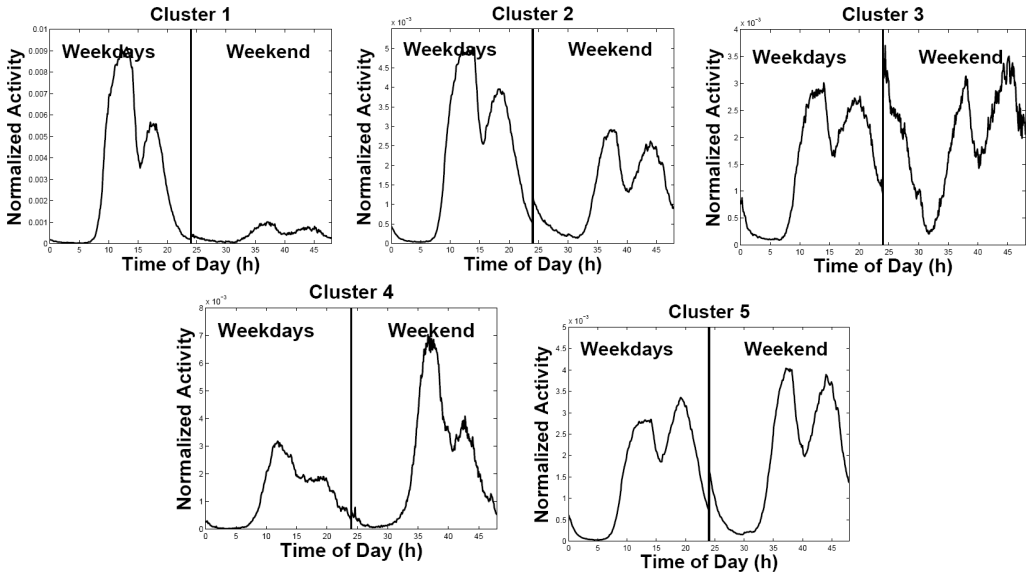


Figure 2: From left to right and top to bottom, class representatives of cluster 1 through cluster 5 of the land uses identified by k-means when using the weekday-weekend aggregation and Euclidean distance.

Sundays. Nevertheless, due to the mixed activities that take place on the city, the signature could also represent office and/or residential areas.

- Cluster 3: The third cluster has two elements that characterize its land use: (1) activity during weekdays and weekends has the same relevance and (2) during weekends there is a strong activity between 0AM and 5AM, indicating nightlife environments. The land use derived from this behavior is of nightlife areas: restaurants, bars, etc. This cluster is a good example of why the total aggregation produces worst coefficient values than weekday-weekend aggregation, as such differences are lost in the aggregation process.

- Cluster 4: This class representative shows that activity during weekends is more than twice that of weekdays, and that this activity is focused between 12PM and 5PM, i.e. during day light hours. This land use probably characterizes weekend leisure activities.

- Cluster 5: This signature shows that activity during weekdays and weekends is of the same magnitude. During weekdays, the second peak of activity is of higher magnitude than the morning peak, while during weekends both peaks have the same magnitude. These characteristics imply residential areas, where individuals come back after work and stay during weekends.

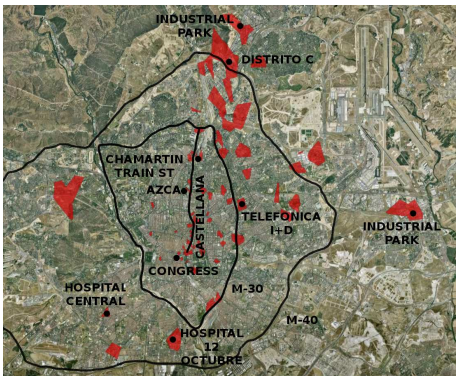
6. VALIDATION

Ideally, any validation should be done using a ground truth. Urban planning departments usually have some information available on urban land use. Nevertheless, in our experience, the information available has the following characteristics: (1) it is regarding how land use is planned not on the actual use of the land, which is not necessarily the same; and (2) urban planing focuses mainly on defining residential and industrial areas, not the variety of uses we have discovered. Considering the previous factors the validation has been done using our expert knowledge of the city.

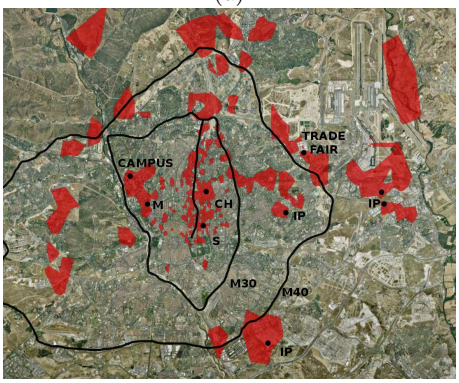
Figure 3 shows the geographical representation of clusters 1 through 4, in a map of Madrid. Generally speaking, the city is contained inside two concentric ring roads (M-30 and M-40). The area inside the M-30 (the smaller ring) contains the city center as well as the main business, tourist and commercial areas, all of them mixed with residential areas. The area comprised between the M-30 and M-40 contains mainly residential districts and industrial parks. Due to space constraints cluster 5 has not been presented, although it is geographically formed by the areas not included in the previous clusters. The validation of each cluster consists on checking to which extent the interpretation of the signatures given in the previous section actually correlates with the geographical location of the clusters. Our main findings are:

- *Cluster 1 - Industrial Parks & Office Areas* (Fig.3(a)): The geographical areas included are mainly around Castellana, which concentrates the main business areas of the city (like AZCA). Also included are industrial parks and office areas situated in the north and east of the city (including both Telefonica campus, known as Distrito C and Telefonica Research buildings), hospital complexes and some public buildings (such as the Congress or the Chamartin Train Station). The assumption made for this cluster seems to hold true, although not only for offices and industrial areas but also for hospital complexes and public buildings.

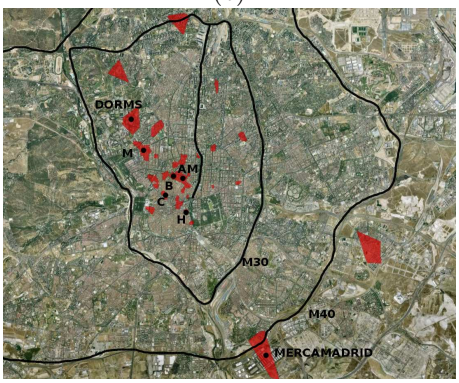
- *Cluster 2 - Commercial Areas* (Fig.3(b)): The areas included in this cluster focus mainly around the Salamanca (marked with S), Chamartin (Ch) and Moncloa (M) districts. These districts have a strong commercial activity, but are also densely populated residential areas. It also contains the university campus (which includes residential areas for students), some Industrial Parks (IP) with a strong commercial activity, and ar-



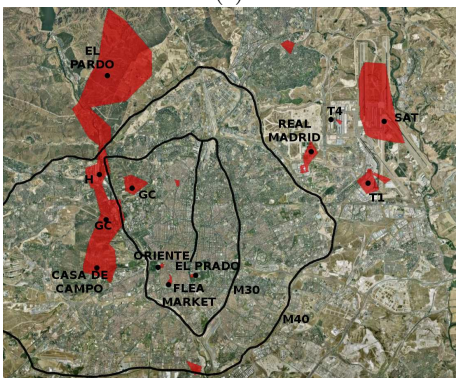
(a)



(b)



(c)



(d)

Figure 3: Geographical representation in a map of Madrid of: (a) Cluster 1 (Industrial Park - Office Areas), (b) Cluster 2 (Commercial-Residential), (c) Cluster 3 (Nightlife) and (d) Cluster 4 (Leisure).

eas along the M40 probably caused by the traffic in the road. Our hypothesis is partially validated as the areas included are mainly commercial-residential.

- *Cluster 3 - Nightlife areas* (Fig.3(c)): The correspondence in this case with nightlife areas is straightforward, as the main nightlife areas of the city are included: Gran Via, Callao (C), Bilbao (B), Moncloa (M), Alonso Martinez (AM), Huertas (H). Part of the Campus, which contains a high concentration of dorms, is included in this cluster. Also there is activity in Mercamadrid, the wholesale food market of the city, whose activity takes place early in the morning, and as such is clustered as a nightlife area.

- *Cluster 4 - Leisure areas* (Fig.3(d)): As in the previous case the behavior of this cluster is highly correlated with the geographical location of the cluster. Good examples are parks (Casa de Campo), golf clubs (GC), the Horse Racing Track (H), Plaza de Oriente, El Prado Museum (including the botanical gardens) and the flea market (which happens Sunday mornings). Also included in this cluster are the grounds of the Ciudad Deportiva del Real Madrid, where the team trains in public, and part of the airport (terminals T1-T2-T3 and the Satellite terminal), indicating a strong activity of these terminals during weekends.

- *Cluster 5 - Residential areas*: The area covered by this cluster accounts for approximately 60% of the geographical area. It is highly correlated with residential areas, mainly in the south and west of the city.

We can conclude that the land use assumptions made in the previous section with the cluster representatives are validated because there is a strong correspondence between those uses and the infrastructures included in the geographical representation of each cluster.

7. AUTOMATIC CLASSIFICATION OF LAND USE

The knowledge provided by the class representatives can be used to automatically classify the land use of other cities. Given that k-means assigns every signature to the cluster that minimizes the Euclidean distance to its centroid, the natural classifier to build is one that assigns each signature X_n to the class that minimizes the Euclidean distance to its class representative c_i. Formally the class of a BTS signature X_n, y_n, is defined as:

$$y_n = \arg\min_i d(X_n, c_i) \quad (6)$$

Considering the inherent cultural factors captured by the land uses identified, we tested the classifier using data from Barcelona. The city contains around 900 BTSs in an area of around 100 km^2. Figure 4 presents the geographical areas classified in each one of the land uses identified. The areas not present in any of those figures were classified as residential. The Industrial Parks&Office Areas cluster includes the commercial part of the port and the port free zone ($ZonaFranca$) and different areas around Diagonal avenue, the main business area of the city where the main offices are located. Hospital Complexes were also identified. The second land use, commercial-residential includes more areas

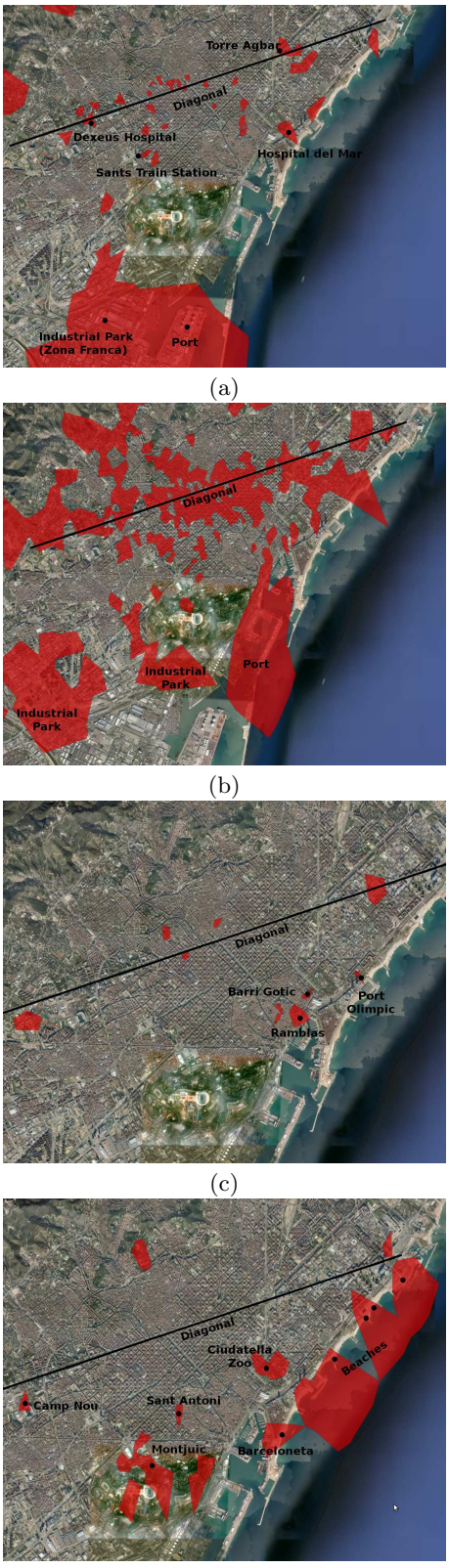


Figure 4: Automatic classification of land use in Barcelona: (a) Industrial Park - Office Areas, (b) Commercial-Residential Areas, (c) Nightlife areas and (d) Leisure activities.

around the Diagonal avenue, the recreational part of the port and some industrial parks with commercial activity. The land classified as Nightlife is present around the Port Olimpic (with a high concentration of bars and nightclubs), La Rambla (specially the part closest to the port), el Barri Gotic and part of the University Campus. The leisure areas identified include the beaches (including Barceloneta), the area around Parc Montjuic and the Ciutadela Gardens (where the Zoo is located), the Guell park (not shown in the map), and an area around the Sant Antoni market that has an open market on Sundays. Also the Barcelona soccer stadium (Nou Camp) is included under this land use.

8. CONCLUSIONS AND FUTURE WORK

We have presented a data mining approach to the problem of identifying land use. While traditional approaches are based on questionnaires, which implies cost and time limitations, our solution overcomes these problems and brings new advantages like the capability of tracing land use evolution over time or the ability to focus the study in a particular social background(elders, tourists, socio-economic levels, etc.). Our approach is not intended to substitute traditional urban analysis approaches but to complement and improve them. For future work we plan to find other sources of information to have an improved validation and to construct other classifiers for land use detection.

9. REFERENCES

[1] R. Ahas and U. Mark. Location based services – new challenges for planning and public administration. In *Futures*, volume 37(6), 2005.

[2] D. Brockmann. Human mobility and spatial disease dynamics. In *Review of Nonlinear Dynamics and Complexity - Wiley*, 2009.

[3] J. Candia, M. Gonzalez, P. Wans, T. Schoenharl, and A.-L. Barabasi. Uncovering individual and collective human dynamics from mobile phone records. In *J. Phys. A: Math. Theor.*, volume 41, 2008.

[4] N. Eagle and A. Petland. Reality mining: Sensing complex social systems. In *Personal and Ubiquitous Computing*, volume 10(4), 2006.

[5] T. Horanont and R. Shibasaki. Evolution of urban activities and land use classification through mobile phone and gis analysis. In *CUPUM*, 2009.

[6] L. Liao, D. Paterson, D. Fox, and H. Kautz. Learning an inferring transportation routines. In *Artificial Intelligence*, volume 171, 2007.

[7] C. Ratti, R. M. Pulselli, S. Williams, and D. Frenchman. Mobile landscapes: using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design*, 33(5):727–748, 2006.

[8] S. Ray and R. H. Turi. Determination of number of clusters in k-means clustering and application in colour image segmentation. In *ICAPRDT*, 1999.

[9] J. Reades, F. Calabrese, and C. Ratti. Eigenplaces: analysing cities using the space-time structure of the mobile phone network. *Environment and Planning B: Planning and Design*, 36(5):824–836, 2009.

[10] J. Reades, F. Calabrese, A. Sevtsuk, and C. Ratti. Cellular census: Explorations in urban data collection. *IEEE Pervasive Computing*, 6(3):30–38, 2007.