Robust Land Use Characterization of Urban Landscapes using Cell Phone Data

Víctor Soto and Enrique Frías-Martínez

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

Abstract. Mobile devices have become one of the main sensors of human behavior, and as such, can be used as proxies to study urban environments. In this paper we describe a method to automatically identify land uses from call detail record databases. Given the inherent diversity of human activities in urban landscapes, we use fuzzy clustering techniques to identify land uses in a robust way, characterizing only those geographical areas with well defined behaviors. Finally, we validate the results obtained using expert knowledge.

1 Introduction

Ubiquitous technologies, such as cell phone networks, geo-localized tagging, and more recently, apps running on smartphones, are proving to be excellent data sources for a variety of fields such as smart cities, urban planning and social network analysis [1, 2]. Of all these new data sources, cell phones traces are becoming increasingly important, as they contain valuable information on a variety of aspects of human dynamics (i.e. mobility, behavioral patterns, etc.).

In this paper we present a technique to automatically identify the uses that citizens give to the different parts of a city using the information contained in cell-phone records. Given the inherently fuzzy nature of both human behavior and urban landscapes, we propose a method to obtain robust land uses using fuzzy clustering techniques. Using this method, only sections of the city with a given minimum similarity to the land uses identified will be labeled. Although our technique is going to be presented for call detail records, it can also be used with other ubiquitous data sources, like geolocalized tweets, flickr or the logs of any service that includes geolocalization.

Several authors have already used cell phone data to carry out urban analysis studies. For example, [3] used aggregated cell-phone data to analyze urban planning in Milan with an interest in location-based services applications. In [2] the authors obtained behavioral patterns from the information obtained from individual phones. The authors of [4] use bluetooth to characterize pedestrian flow data. Focusing exclusively in land use analysis, some authors have already presented studies to solve related questions. [5] monitorized the dynamics of Rome and obtained clusters of geographical areas measuring cell phone towers activity using Erlangs, although little information is given about the characteristics of those areas. Another study is described in [6], where the authors analyze four different geographical spots at different times in Bangkok. Using eigendescomposition, [7] studied the time structure of the Erlangs, comparing the network activity and the commercial activity of the area. Nevertheless, to the best of our knowledge, none of the previous studies have focused on automatic identification of land uses.

2 Preliminaries

In order to automatically identify land use behaviors in urban environments we present a technique based on 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. The traffic handled by a BTS is shared by a number of sectors (tipically three). 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. For simplicity, we use Voronoi tessellation to approximate the area of coverage of a set of BTSs as non-overlapping polygons. Each time a user uses a service (SMS, MMS, voice), a Call Detail Record (CDR) is created with an associated timestamp and the sector that handled it, which gives an indication of the geographical location of the mobile phone at a given moment in time (no information about the position within a sector is known). The set of fields typically contained in a CDR include: (1) originating encrypted phone number, (2) destination encrypted phone number, (3) identifier of the sector that handled the originating phone number, (4) identifier of the sector that handled the destination phone number, (5) date and time of the call and (6) duration of the call.

Using the information contained in a CDR database we can characterize the uses given to specific urban areas. The geographical areas in which the city is going to be divided will be defined by the Voronoi tessellation of the set of BTSs, and each area will be characterized with the corresponding BTS activity (the signature of the BTS tower). The identification of land uses and the degree to which each geographic area is explained by them, can be automatically done by clustering the set of signatures using fuzzy c-means. The robust identification of land uses is implemented using as filters the degrees to which each geographical area is explained by each land use. Our study has been done using CDR data collected from Madrid during a period of 1 moth, from October 1st 2009 to October 31st 2009. The area covered by the city is of $400Km^2$, with more than 3 million inhabitants, and is served by 1100 towers that collected over 100 million interactions.

3 Activity 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. In our study we have measured the activity every five minutes. Although other intervals were tested, higher resolutions did not add any extra information (while

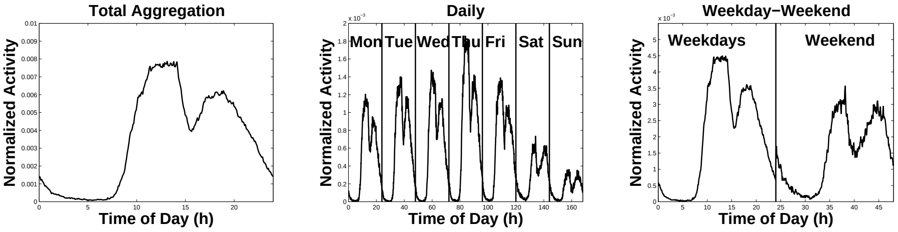


Fig. 1. Examples of (left) total aggregation of one day; (center) daily aggregation every day of the week and (right) weekday-weekend aggregation.

increasing the complexity), and lower resolutions affected greatly the results. In order to build the activity signature of each BTS, a 2-dimensional matrix A_n is defined for each BTS_n, $n \in \{1, \ldots, \|BTS\|\}$. Each element $A_n(\delta, \tau)$ contains the activity of BTS n during a 5-minute time interval τ on a given day δ where $\delta \in \{1, \ldots, 31\}$ and $\tau \in \{1, \ldots, 288\}$, with 288 the total number of measurements per each 24 hour period. Human dynamics are well differentiated between week days and weekend days [8], and those differences will translate into different BTS levels of activity. In order to preserve that information we opt to build each BTS signature as the concatenation of the aggregated actitivity of the BTS during weekdays (monday to friday) and weekends (saturday and sunday) (see Fig.1 right), producing a final signature of 576 elements. After that, the signature is normalized. Although other aggregations are possible, this representation allows to retain information otherwise lost using total aggregation and reduces computational complexity w.r.t. daily aggregation (Fig.1 left and center respectively). The weekday-weekend aggregation is computed as ($+\!\!+$ indicates concatenation):

$$Y_n(\tau) = \frac{1}{|\delta \in weekday|} \sum_{\delta \in weekday} A_n(\delta, \tau) \tag{1}$$

$$Z_n(\tau) = \frac{1}{|\delta \in weekend|} \sum_{\delta \in weekend} A_n(\delta, \tau) \tag{2}$$

$$X_n = Y_n +\!\!+ Z_n \tag{3}$$

4 Land Use Identification

We have used fuzzy c-means to cluster the BTS signatures and obtain the class representatives that define land uses. Fuzzy c-means needs as input the number of clusters and the fuzziness coefficient m. In order to find the optimal number of clusters we have used subtractive clustering [9]. This method assumes every item in the dataset might be a cluster center and assigns a potential value to them. The method iteratively picks the item with the highest potential, selects it as a cluster center, and decreases the potential of the surrounding items that are situated within a radius of influence r_{inf}. Low r_{inf} values produce a large number of small clusters, while the opposite happens with high r_{inf} values. Common values for

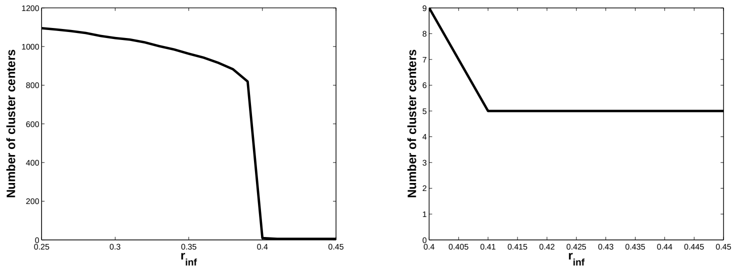
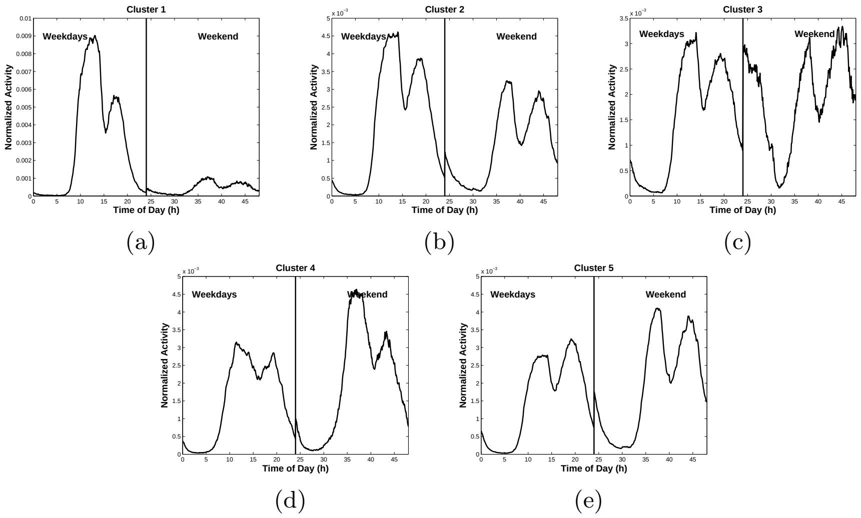


Fig. 2. (left) Evolution of r_{inf} vs. number of centroids in interval [0.25,0.45], and (right) in [0.4,0.45].



(a)

(b)

(c)

(d)

(e)

Fig. 3. Cluster signatures 1, 2 & 3 (a-c) and cluster signatures 4 & 5 (d-e).

r_{inf} are in the range $(0.25, 0.45)$. Figure 2(left) shows the number of clusters for each value in the mentioned interval with 0.01 increments. It can be observed that the curve decreases drastically in $r_{inf} = 0.4$ (Fig. 2(left)), and stabilizes in 5 clusters (Fig. 2(right)). The fuzziness coefficient m, $1 \leq m < \infty$, regulates the fuzziness of the partition. In general the value of m is data dependent. We used the method presented in [10] to estimate m, finding an optimum value of $m = 1.2$.

Figure 3 shows the land use representatives obtained after running fuzzy c-means with 5 clusters and $m = 1.2$. An analysis of these signatures allows to hypothesize aboute the land uses. Figure 3(a) describes a land use characterized by a high activity during weekdays and almost non-existent activity during weekends. Also, it shows that morning activity is higher than afternoon activity, indicating probably office or industrial parks activity. Figure 3(b) presents a similar behavior to 3(a), but in this case there is a relevant activity during weekends. This land use probably indicates business and/or commercial areas that probably include relevant residential parts. Figure 3(c) is characterized by a peak of night activity during weekends, implying nightlife areas. Figure 3(d) presents a land use where the main activity takes place during weekends, while during weekdays the activity is equally relevant during mornings and afternoons. The relevance of the weekend activity probably indicates leisure activity areas. Finally, in figure 3(e) weekend activity is higher than weekday activity and on weekdays the activity is higher during the afternoon than during the moorning. This behavior is typical of residential areas in which individuals come from work in the afternoon during weekdays and during weekends stay at home.

5 Robust Land Use Analysis

In general, land uses in urban landscapes are not necessarely well defined in the sense that one area has just one use. This is especially true in old cities where dowtown areas have a variety of commercial, office, residential and nightlife activities that are present in a reduced geographical space. In this paper we are interested in robust land use analysis, implying that we want to identify those areas that have a clearly defined land use.

The use of fuzzy c-means allows to capture for each BTS (and by extension for its coverage area) the degree to which each land use identified is present. Fuzzy c-means returns a membership index U_i^j for each pair of BTS tower bts_i and centroid c_j that meet the requirement $\sum_{j=1}^{5} U_i^j = 1 \; \forall i$. Values of U_i^j close to 1 indicate that the behaviour of that BTS is very close to the signature j. We run a crisp classification of the areas by assigning each BTS tower to the cluster with the highest membership index ($cluster_i = \arg\max_j U_i^j$) and with a membership degree of $m_i = \max_j U_i^j$.

In order to identify robust land uses we discard all BTS towers with a membership degree lower than a given threshold θ. Figure 4(left) displays the percentage of items per cluster that pass a given $\theta \in [0.7, 1]$. When $\theta = 1$ we are considerig BTSs for which fuzzy c-means has identified just one land use. It can be observed that in this case cluster 1 (representing probably industrial parks& office areas) and cluster 5 (representing probably residential areas) have 10% and 20% of its items with a membership degree of 1. For the rest of the clusters less than 2% of its items pass the $\theta = 1$ filter. This result implies that it is possible to find areas that are just residential or just industrial, while in general, areas whose main activity is commercial(cluster 2), nightlife (cluster 3) or week-end activities (cluster 4) are always combined with other use(s).

Because of the different effects of θ in the different clusters, in order to identify robust land uses, it is more intuitive to define the filter with a percentile common to all clusters. Figure 4(right) shows the minimum membership index value (Y axis) of the items above each p percentile (X axis) for each cluster. Considering these results, we define the concept of robust land use as the set of areas above the 60th percentile of each cluster, which implies a minimum membership degree of 0.99 for cluster 1, 0.94 for cluster 2, 0.92 for cluster 3, 0.82 for cluster 4 and

0.99 for cluster 5. We choose to filter on percentile 60, where cluster 4 (which obtains the worst results) has a minimum membership degree close to 0.8.

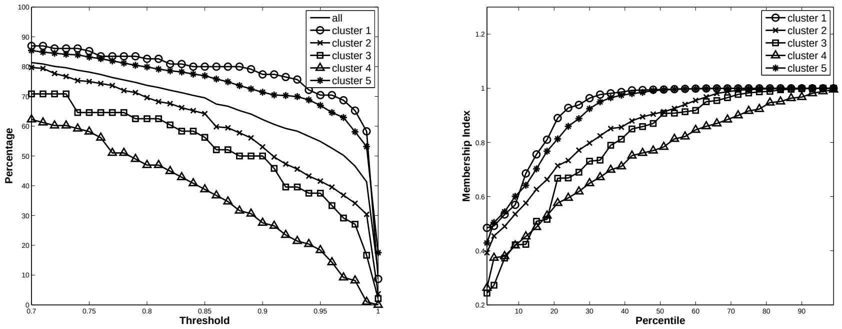


Fig. 4. (Left) evolution of the percentage of items with a minimum threshold and (right) evolution of the membership index per percentile.

6 Validation of Robust Land Uses

Figures 5 and 6 present the geographical representation of the areas that have a robust land use for each cluster (cluster 5 has not been presented due to space limitations). Madrid is defined by two concentric ring roads, M-30 and M-40. The inner ring road comprises the city center, all the touristic areas and the majority of the business and commercial activity. The main residential areas and industrial parks are situated in the area between the two ring roads. The validation consists on checking to which extent the interpretation of the signatures given in section 3 correlates with the infrastructures located in the geographical representation of each cluster. Since a database detailing the actual land use of the city (not the planned land use) and the different uses we have identified is not available, we use our expert knowledge of the city:

- Cluster 1 - Industrial Parks & Offices Areas (figure 5(left)): the geographical representation of this cluster comprises areas in the Paseo de la Castellana, which concentrates the main office areas of the city. Also, office areas and industrial parks located in the north east of the city are highlighted, including the Telefonica campus (Distrito C). This cluster also includes Madrid's most important public hospitals.
- Cluster 2 - Commercial & Business Areas (figure 5(right)): this is a hybrid cluster that appears more prominently in downtown, around the Paseo de la Castellana, in the districts of Salamanca, Chamberi and Chamartin. These districts concentrate the main commercial areas in the city, although they are also residential areas. It also includes small offices areas and part of the university campus.

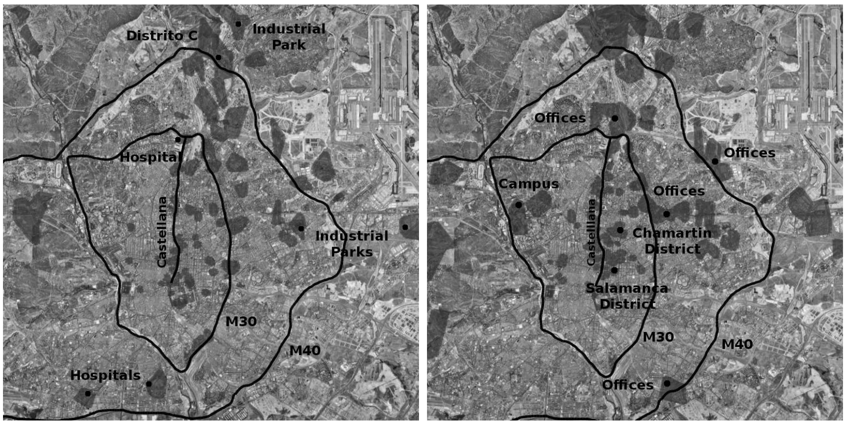


Fig. 5. Representation of (left) Industrial Parks&Offices and (right) Commercial clusters.

- Cluster 3 - Nightlife Areas (figure 6(left)): Madrid's most important nightlife areas appear in this cluster. The districts of Alonso Martinez-Tribunal-Chueca are grouped together in the city centre. Pub zones of Moncloa and Arguelles are situated in the northwest. The dorms of the university campus are also included. Other nightlife areas, such as Atocha and parts of Gran Via are part of the cluster.
- Cluster 4 - Leisure & Transport Hubs (figure 6(right)): this cluster comprises two kinds of activities that at first might seem very different. The first group of items is related to leisure activities on weekends. It contains the horse racing track, parks, golf clubs and country clubs situated in the west of the city. Also included in this cluster is the flea market, which takes place every sunday morning. The second group of items contain the main transport hubs: Madrid-Barajas airport, Atocha railway station and the main bus station.
- Cluster 5 - Residential Areas: the largest cluster contains the biggest residential districts in the city, all located between the ringroads, mainly in the southeast and southwest of the city.

The land use assumptions presented in section 3 are validated with the corresponding geographical representation, and in some cases (like hospitals in Cluster 1 and Trasport hubs in cluster 4), other uses have been identified.

7 Conclusions and Future Work

We have presented an automatic procedure to identify robust land uses using the information contained in call detail records. Our approach uses fuzzy-c means in order to capture the inherent fuzziness of human behavior and to implement a filter that defines the concept of robust land use. The results indicate that five different land uses can be identified, and their interpretation was validated with their geographical representation. For future work we are focusing on finding

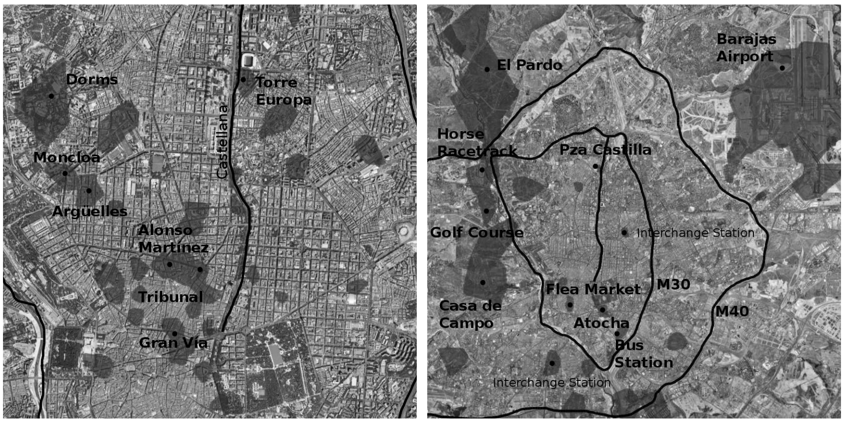


Fig. 6. Representation of (left) Nightlife and (right) Leisure & Transport clusters.

new sources of information that allows us to validate our approach with a ground truth and improving our technique in order to separate different land uses that are currently grouped under the same cluster.

References

1. Liao, L., Paterson, D., Fox, D., Kautz, H.: Learning an inferring transportation routines. In: Artificial Intelligence. Volume 171. (2007)
2. Eagle, N., Petland, A.: Reality mining: Sensing complex social systems. In: Personal and Ubiquitous Computing. Volume 10(4). (2006)
3. Ratti, C., Pulselli, R.M., Williams, S., Frenchman, D.: Mobile landscapes: using location data from cell phones for urban analysis. Environment and Planning B: Planning and Design **33**(5) (2006) 727–748
4. Ahas, R., Mark, U.: Location based services – new challenges for planning and public administration. In: Futures. Volume 37(6). (2005)
5. Reades, J., Calabrese, F., Sevtsuk, A., Ratti, C.: Cellular census: Explorations in urban data collection. IEEE Pervasive Computing **6**(3) (2007) 30–38
6. Horanont, T., Shibasaki, R.: Evolution of urban activities and land use classification through mobile phone and gis analysis. In: CUPUM. (2009)
7. Reades, J., Calabrese, F., Ratti, C.: Eigenplaces: analysing cities using the space time structure of the mobile phone network. Environment and Planning B: Planning and Design **36**(5) (2009) 824–836
8. Candia, J., Gonzalez, M., Wans, P., Schoenharl, T., Barabasi, A.L.: Uncovering individual and collective human dynamics from mobile phone records. In: J. Phys. A: Math. Theor. Volume 41. (2008)
9. Chiu, S.: Fuzzy model identification based on cluster estimation. Journal of Intelligent & Fuzzy Systems **2**(3) (September 1994) 267–278
10. Schwammle, V., Jensen, O.: A simple and fast method to determine the parameters for fuzzy c–means cluster analysis. Bioinformatics (Oxford, England) (2010)