Computing Individual Mobility Profiles from Mobile Phone Usage Traces

Miguel Á. Rodríguez-Crespo, Ana Armenta, Alberto Martín-Domínguez, Rocío Martínez-López, Rubén Lara
Telefónica I+D
Madrid, Spain
[miguel, aalv, amd, rml, rubenlh]@tid.es

***Abstract*—This paper describes the procedures used to automatically generate a complete individual mobility profile of users of mobile services, based exclusively on geo-located phone usage information and without requiring any customer interaction. It also presents the results of applying these procedures in a test with real data. The individual mobility description includes metrics such as area of activity, diameter of influence area, location and meaning of the user's points of interest, and frequent itineraries.**

Keywords— Individual mobility; mobile phone usage; center of mass; radius of gyration; points of interest; itineraries

I. INTRODUCTION

The study of human mobility patterns has received growing attention over the past few years, especially due to the increasing availability of location data coming from both global positioning systems (GPS) and mobile telephone usage, which leaves geo-located traces on the operators networks [1][2][3][4].

Understanding how and when human movements take place across our towns, cities or countries is of interest in many areas, such as traffic management, transport network design or diseases spread control [5][6][7]. However, not only a global view of population flows, but also individual mobility patterns of a user, are of great interest in a number of fields [8]. The knowledge of what locations a user periodically visits, over what period, with what frequency, what days of the week and at what times of the day [9][10] can be used for the provision of contextual services, relevant advertising, targeted offers to address the particular mobility needs of the user, itinerary planning… In general, knowing locations that are relevant for a user can enable the personalization of commercial communications or service interactions and improve their relevance.

This paper describes the set of procedures implemented to obtain a complete description of the mobility profile for mobile phone users. These procedures have been tested over several million customers from the same country, treated in an anonymous way. Future utilization of these procedures will imply customers give their previous authorization to treat their unanonymized id so they can be offered customized services or applications.

The paper is structured as follows. Section 2 describes the general framework and the kind of geo-located data we use to obtain these profiles. In Section 3, we explain how we compute certain areas for every customer. Section 4 explains how we detect and label the points of interest of each customer. Section 5 describes how we detect the frequent itineraries of a user. Finally, in Section 6, we present some conclusions and ideas for future work.

II. GENERAL FRAMEWORK

The method used for computing individual mobility profiles works on geo-located events generated by using mobile phones. These geo-located events are obtained in a non-intrusive way for the users. These events are traces that mobile phone usage leave on the network and include initiation and termination of voice calls, sending short text messages (SMSs) or multimedia messages (MMSs), signaling (as data sessions attach/detach events), and so on.

Geo-located events must contain, at least, the following information: an anonymized user id associated to the event, and the date, time and location of the event.

Location information is usually available at a Base Transceiver Station (BTS) level. BTSs are the places where the antennas that receive and transmit the radio signals from and to the mobile phone users are located. So the location information in fact refers to an approximate area, not an exact place. Even more, the customers' location is not known as a continuous function of time; it is only known at certain points in time when customers are interacting with the mobile phone network.

These kinds of events are collected over a period of time to obtain the dataset used to compute the mobility profiles of the users.

The work and results presented in this paper are based on data collected from Call Detail Records (CDRs) of voice calls of customers from a Latin American country over a six-month period. These data comprise about 7 billion CDRs, from more than 16 million customers. There are more than 5000 different locations (BTSs) over the whole country.

III. CENTER OF MASS, RADIUS OF GYRATION AND DIAMETER OF INFLUENCE AREA

First, some simple but informative descriptors are calculated. These descriptors give information about the areas where the customers are usually located over the time period considered.

For each customer, a set of locations is obtained, each with a count indicating the number of times (n_i) we observe the customer at that location. Each location is represented by a pair of 2-D planar coordinates (x_i, y_i). The center of mass (*CM*) for a customer is obtained as the location whose *CMx* and *CMy* coordinates are calculated as the weighted average of all the locations known for that user over the time period.

$$CMx = \sum_{i=1}^{N} n_i x_i / \sum_{i=1}^{N} n_i \quad (1)$$

 ISBN: 978-1-61208-226-4

$$CMy = \sum_{i=1}^{N} n_i y_i / \sum_{i=1}^{N} n_i \quad (2)$$

As the activity of the customers can be very different for workdays and weekends, two centers of mass are calculated: a workdays center of mass (CMx_{wd}, CMy_{wd}) from the activity observed for the user from Monday to Friday, and a weekends center of mass (CMx_{we}, CMy_{we}) from the activity observed on Saturdays and Sundays.

Once the two centers of mass are obtained, two radii of gyration (R_{wd} and R_{we}) are calculated for each one. A radius of gyration is computed as the weighted average of the distances from the customer registered locations to his/her center of mass.

$$R = \sum_{i=1}^{N} n_i \sqrt{(x_i - CMx)^2 + (y_i - CMy)^2} / \sum_{i=1}^{N} n_i \quad (3)$$

The combination of center of mass and radius of gyration defines a circular area where the customer concentrates most of his activity. Two circular areas are obtained for each customer, one for workdays, given by (CMx_{wd}, CMy_{wd}) and R_{wd}, and another one for weekends, given by (CMx_{we}, CMy_{we}) and R_{we}.

Fig. 1 shows an example of the activity areas for a user. This user clearly changes his/her known phone usage locations from workdays to weekends.

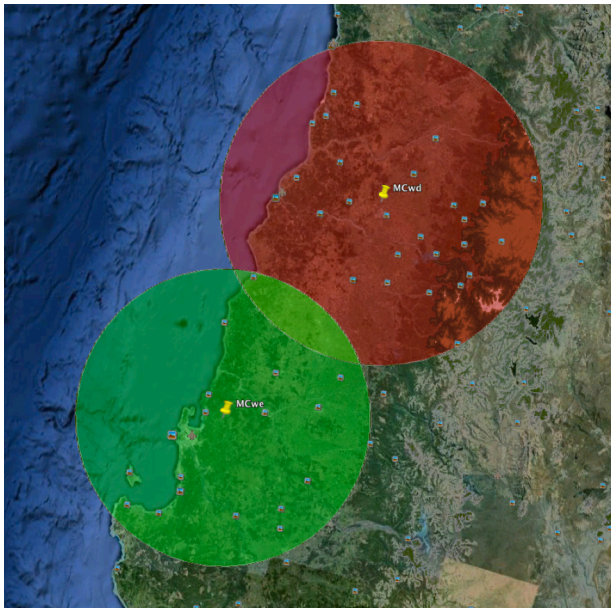

Figure 1. Workdays (red) and weekends (green) activity areas defined by centers of mass and radii of gyration for a user.

Another parameter is calculated to complement the information given by the centers of mass and radii of gyration: the maximum distance between two registered locations for a customer over the period of study (diameter of the influence area). This information allows identifying users that have gone to very distant places during the time period and users that have stayed in a more limited area.

From a general analysis of the locations visited by the customers, we can derive some information:

- 50% of the customers use less than 9 different BTSs during workdays and less than 7 on weekends. Just 5% of the customers visit more than 51 different BTSs during workdays.
- The radii of gyration are less than 4 km for 50% of the customers and are very similar for workdays and weekends. Just 5% of the customers have radii of gyration greater tan 17 km.
- 50% of the customers have an influence area diameter lower than 20 km in workdays and lower than 18 km in weekends.

IV. Detection and Labelling of Points of Interest

The second component of the individual mobility profiles is the detection and labeling of the users' points of interest (PoIs). The goal is to find the most relevant locations for every user, from the set of locations he/she had activity at, and to characterize them by labels that give information about what each PoI means for that user.

A. Detection of PoIs

First, the CDRs of users that have too many or too few calls during the time period are filtered out. This way we avoid the noise introduced by users not corresponding to an individual usage (i.e., switching devices or PBXs of a company or business), or corresponding to individuals whose activity is too low to extract a meaningful usage pattern. A lower limit of 200 calls and an upper limit of 5000 calls over the time period were established. More than 8 million customers are between these thresholds. For these customers, BTSs having at least 5% of the total number of each customer's calls are selected. The customer-BTS pairs that satisfy these requirements define the PoIs of that customer.

B. Labelling of PoIs

Every PoI is characterized by a communication vector that represents its temporal usage pattern. For one PoI (customer-BTS pair) this communication vector initially contains the number of calls the customer did in the BTS at every different hour of the different weekdays (from Monday to Sunday), aggregating the values for the same weekdays of the whole time period. As not every weekday has the same meaning in terms of life activity patterns, it was decided (after a previous analysis) to group Monday, Tuesday, Wednesday and Thursday in one single kind of weekdays (Monday-Thursday), and to maintain Friday, Saturday and Sunday as separate kinds of weekdays.

So, for every customer there are several communication vectors (curves), one for each of his/her PoIs, containing the number of calls made by that customer in that location at every hour of 4 kinds of weekdays (Monday-Thursday, Friday, Saturday and Sunday). Each vector has $4 \times 24 = 96$ values.

As the number of different kinds of weekdays is not the same, a first normalization is done dividing each vector value by the number of days of its kind of weekday that occurs in the time period. This allows comparing the 4 different parts of the curves.

 ISBN: 978-1-61208-226-4

In order to allow the comparison between different curves from different PoIs with very different activity levels (number of calls), a second normalization is done dividing each vector value by the sum of all the vector values. All normalized PoI communication vectors will have a resultant sum equal to 1. This way the differences between vectors focus on the curve shape itself reducing the importance of the curve amplitude levels.

Fig. 2 shows the normalized communication vectors of two PoIs, with different shapes that represent different kinds of phone usage. The first one is more uniform and regular over the different weekdays. The second only presents activity from Monday to Friday, especially in the evenings.

In order to label the communication vectors, it is assumed that different shapes (usage patterns over the week) imply activities of different nature at those locations. A set of clusters (groups) is obtained from the whole set of PoI communication vectors of all the customers. A meaningful label is assigned to each cluster, related to the usage pattern over time.

About 24 million PoIs (communication vectors) were obtained from the dataset. A representative sample of this set was selected to run the clustering algorithm (clustering training). A partitioning around medoids (PAM) algorithm was used. This algorithm allows the use of different types of distances to represent the dissimilarity between vectors. In particular, a distance based on the Pearson correlation coefficient between any two communication vectors a and b was used. The distance (dissimilarity) between two vectors was calculated as $1 - \rho_{ab}$, where ρ_{ab} is the Pearson correlation coefficient (similarity). So, the distances have a minimum value of 0 (exactly the same shape) and a maximum value of 2 (exactly the opposite shape).

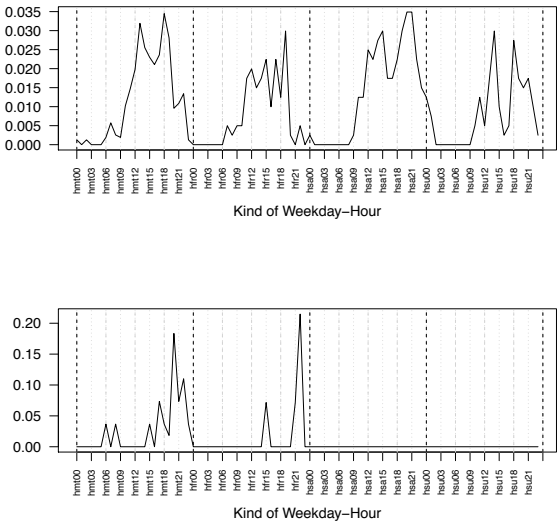

Figure 2. Normalized communication vectors of two PoIs.

The clustering over the sample of PoI vectors produces a vector cluster id (group) assigned to each vector in the sample. All the vectors with the same cluster id belong to the same group, and a representative vector for the cluster id is also obtained. The representative vector of each cluster id is the cluster medoid, that is one of the vectors assigned to that cluster whose average dissimilarity to all the vectors in the cluster is minimal. The cluster centroid can also be obtained as an average of all the sample vectors found inside the same cluster.

Once the medoids are obtained, the PoI vectors not used in the clustering (or any other PoI vector that could be built later) can be assigned a cluster id calculating the distances based on the Pearson correlation to the different medoids and choosing the cluster id of the representative that is closest to the new input vector (has the minimum distance to it).

The number of cluster or groups has to be given as an input to the training phase of the clustering algorithm. As we would like to detect many different groups for the PoI vectors to find enough different types of locations for a user, a value of 20 clusters was used.

Using the medoid and the centroid for each cluster, a label is assigned to them based on the knowledge of the social habits and cultural characteristics of the country. These are the level 0 labels (detailed set of labels).

Fig. 3 shows 4 examples of cluster representatives, along with the level 0 labels assigned to each. The percentage of vectors of the training sample found in each cluster is also shown.

The cluster representatives are later grouped into 5 level 1 labels, thinking on practical applications that will not need as much detail as given by some of the 20 level 0 labels. So, any PoI vector is assigned the level 0 label of its closest medoid, and the level 1 label is assigned based on a table that associates a level 1 label to each level 0 label.

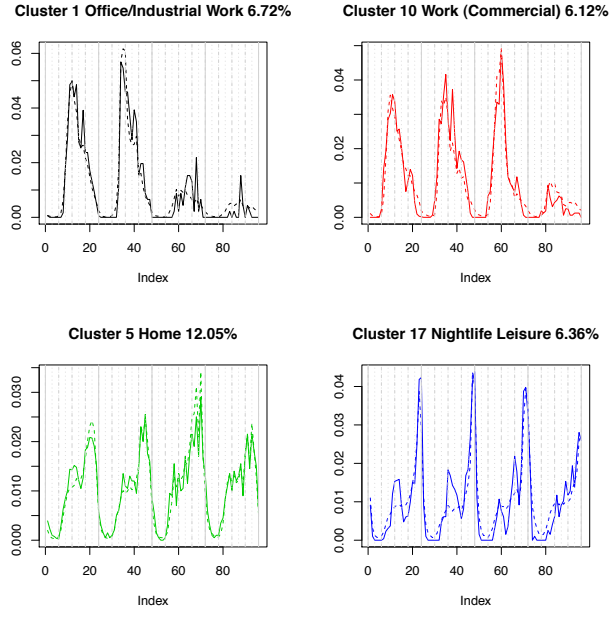

Figure 3. Example of medoids (continuous lines) and centroids (dotted lines) for 4 clusters, with their level 0 labels and percentages in the sample of PoI vectors used in training the clustering.

The level 1 labels and the percentage of training vectors in each of them are:

- Office/Industrial Work 10.03%.
- Commercial Work 16.67%.

 ISBN: 978-1-61208-226-4

- Home 39,00%.
- Night/Evening Leisure 9.31%.
- Afternoon Leisure / Shopping 24.99%

A labeling confidence flag (0/1 value) is obtained based on a distance threshold per cluster. The distances of the training vectors to the medoids of their assigned cluster are used to compute the distances mean and standard deviation for each cluster. The sum of the mean and standard deviation is the threshold distance for each cluster. If the distance of a PoI vector to its closest cluster representative is higher than the threshold distance for that cluster, then the label assigned to that PoI vector is not considered reliable enough (the labeling confidence flag is 0). Otherwise the labeling confidence flag is 1. About 15% of the PoI labels are not considered reliable enough using this criterion.

C. *Comments on the Results of PoI Detection and Labelling*

The output of the PoIs detection and labelling step is a set of locations (BTS's) of special interest for each customer. Each PoI is automatically labelled at two different levels (level 0 and level 1) including a labeling confidence flag. The labels express the particular meanings of the locations for each user.

Fig. 4 shows some distributions of the number of PoIs and labels for the whole set of customers whose data have been processed. More than 8 million customers have at least one reliable PoI (labeling confidence flag=1) and more than 20 million reliable PoIs are detected for those customers, giving an average number of 2.5 PoIs per customer. The black line shows that the highest number of PoIs reaches a value of 12. But, the highest number of different level 0 labels (red line) does not exceed the value of 9 (one customer can have more than one location with the same label or meaning). Most of the customers have 5 or less PoIs, and also 5 or less different level 0 labels. Regarding the number of different level 1 labels (green line), the upper limit of 5 is only reached by a very low percentage of customers.

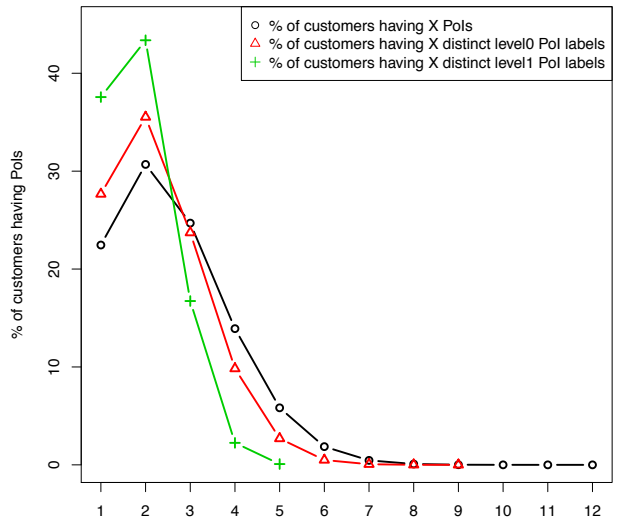

Figure 4. Percentage of customers that have X PoIs with labelling confidence flag=1 (black line), X distinct level 0 labels (red line) and X distinct level 1 labels (green line).

As each customer profile includes his/her centers of mass and radii of gyration, the customer's PoIs information is enriched by flags indicating whether a PoI is located inside or outside the activity area defined by one center of mass and radius of gyration. This is done both for workdays and weekends activity areas to observe if there are any variations depending on the type of PoI.

Fig. 5 shows the percentage of PoIs of a particular type (level 1 label) outside the workdays activity area (red line) and outside the weekends activity area (green line). It includes all the PoIs (even those with labeling confidence flag=0). The two horizontal dashed lines show the mean values of being outside the activity areas (workdays and weekends) for the whole set of PoIs.

As expected, most of the PoIs are found inside the customer's activity areas (almost 80%), as they represent important places for the customers, which concentrate most of their activity. Some types of PoIs are clearly above or below the mean value. There are variations in the percentages of the PoIs being outside depending on considering the workdays activity area or the weekends activity area. Fig. 6 illustrates those variations, showing the percentage of increment of a PoI type being outside the weekends activity area relative to being outside the workdays activity area. The PoI type labeled as "Office/Industrial Work" has the highest weekends/workdays variation (an 80% variation). It is a positive variation, indicating that is much more likely that an "Office/Industrial Work" PoI is outside the weekends activity area than it is outside the workdays activity area (this is consistent with the meaning of that label). The PoI labeled as "Home" has the highest negative variation. It is less likely that the "Home" PoI is outside the weekends activity area than it is outside the workdays activity area (people usually spend more time at home during weekends than during workdays).

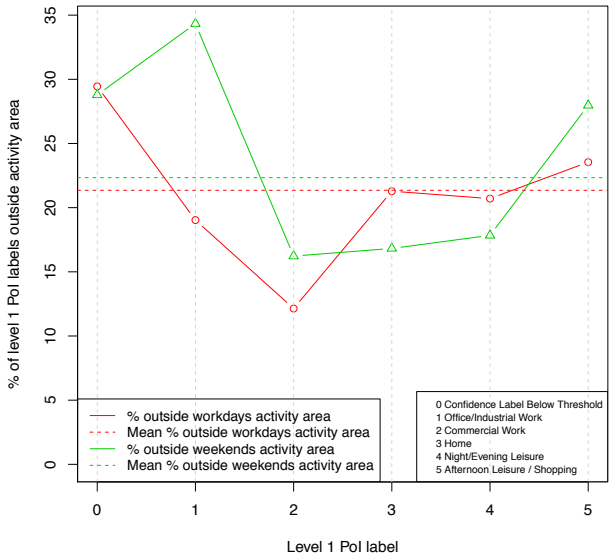

Figure 5. Percentage of level 1 PoI labels outside their customer's activity area (red, workdays activity area; green, weekends activity area)

 ISBN: 978-1-61208-226-4

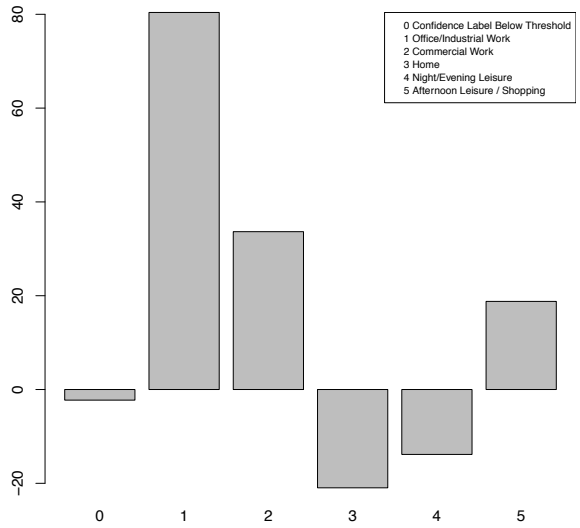

Figure 6. Percentage of increment in the percentage of level 1 PoI labels outside the weekends activity area relative to the percentage of level 1 PoI labels outside the workdays activity area

The relationship between PoI label meanings and their location inside/outside the user's activity areas can be further exploited to refine or even correct the initial PoI labels assignment. For instance, "Home" PoIs detected out of the workdays activity area and far enough from other existing "Home" PoIs inside the same user's workdays activity area can be labeled as "Second Residence".

V. FREQUENT ITINERARIES: MOVEMENTS BETWEEN POINTS OF INTEREST

Focusing exclusively on calls done or received by each customer at his/her PoIs, we consider a movement between two points A and B when we find two consecutive calls, the first one located at point A and the second one located at point B. We establish a maximum time difference of 6 hours between calls to consider the movement is valid. Each movement is spread as a probability over the time period between the two consecutive calls, because the exact time point when the movement takes place is unknown.

A movements curve is obtained as a sum of those probabilities over the six months period under analysis, describing the global probability that the customer moves from PoI A to PoI B at a certain hour of the week. When we find a minimum number of 6 movements between the same 2 points of interest of a given customer, we say we have detected a frequent itinerary between those 2 points. Finally, the itinerary is built as the parts (slots) of this curve above a percentage (20%) of its absolute maximum.

An itinerary description is composed of the peaks from the movements' curve that remain above the amplitude threshold, which represent the most probable time points when such itinerary takes place. The itinerary description also comprises the time-points where the curve intersects the threshold, which describe wider time intervals when the itinerary between the two PoIs implied would probably take place.

For the movements curve depicted at Fig. 7, the corresponding itinerary would contain a set of parameters describing the time positions of the 6 peaks that remain above the threshold (20% of the maximum amplitude), and also the parameters describing the width of those peaks.

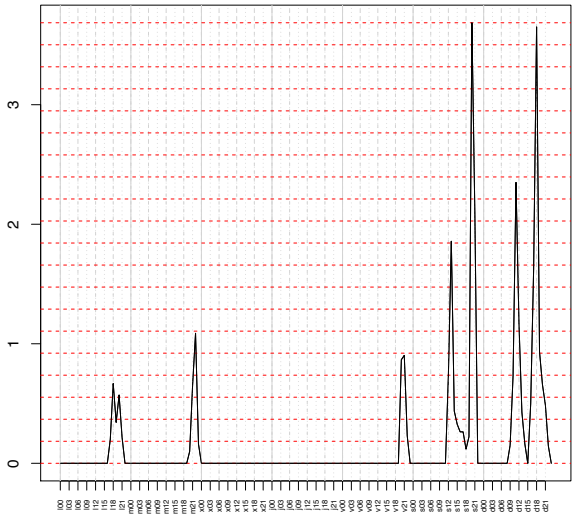

Figure 7. Movements curve between two PoIs of a given customer across the 168 hour intervals of the week (red dotted lines every 5% of the max. amplitude).

Under those assumptions, around 4.5 million customers have more than one frequent itinerary between two of his/her PoIs. The total number of itineraries detected is about 14 million, which gives an average ratio of 3 itineraries per customer. The total number of peaks that reach the 20% maximum amplitude threshold is about 125 million, which implies an average of 8 peaks per itinerary.

The real distributions are not uniform as the two curves in Fig. 8 clearly show. The red line represents the percentage of customers (from the 4.5 million customers that have at least one itinerary) that have X number of itineraries. It has a clear maximum at 2 itineraries per customer and decays quickly to 7-8 itineraries (higher number of itineraries have very low frequency). The green line represents the percentage of itineraries (from 14 million itineraries) that have X number of peaks that exceed the cutoff amplitude threshold. In this case the maximum is located between 5 and 7 peaks per itinerary.

Taking into account the kinds of PoIs that are destination of the detected itineraries, results are aggregated in order to validate internal coherence of the followed procedure. Fig. 9 shows the sum of itineraries across the 168=7*24 hours of the week that start at any kind of PoI with 5 curves for the 5 different categories of PoI as destination. Itineraries to home are majority and take place mainly at evenings of workdays, and at both mornings and evenings of weekends. Itineraries to "Office/Industrial Work" are mainly detected on mornings of workdays. Itineraries to "Commercial Work" are detected more likely from Monday to Saturday during the day hours, similarly to the itineraries to PoIs labeled as "Afternoon Leisure / Shopping". Itineraries to PoIs labeled as

 ISBN: 978-1-61208-226-4

"Night/Evening Leisure" are mainly detected on late night hours.

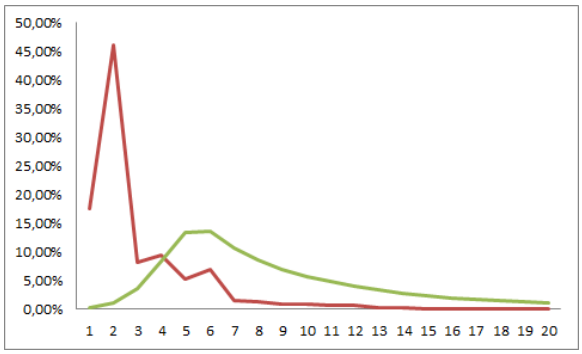

Figure 8. Percentage of customers with X number of itineraries (read line) and percentage of itineraries with X number of peaks (green line).

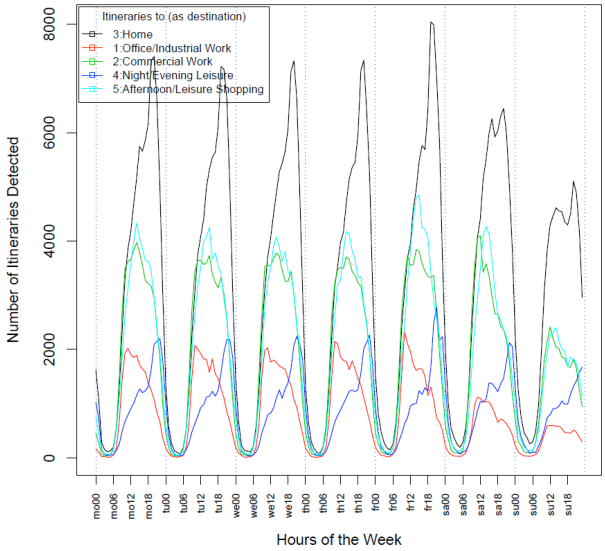

Figure 9. Itineraries detected from any PoI as origin to the different kind of level 1 PoI categories as destination (results based on a sample containing a portion of the total number of itineraries)

VI. Conclusion and Future Work

The mobility profiles we have presented provide a complete description of how mobile phone users move across space and time. The mobility profiles are obtained in a fully automatic way, without any customers' special interaction. Computed individual mobility profiles comprise workdays and weekends activity areas, influence area diameter, points of interest locations and labels, and frequent itineraries. These mobility profiles have been applied to around 16 million customers of a Latin American country.

Aggregated analysis of individual metrics show internal coherence between several independent procedures. The labels of the PoIs that fall outside workdays activity areas tend to be mainly related to leisure activities while labels of PoIs that fall outside weekends activity areas are much more related to work activities. The hour of the week when itineraries are detected match well with the label of the destination PoI. Itineraries to work PoIs are detected mainly on workdays in the mornings, while itineraries to home PoIs are more frequent in the evenings during the whole week.

Although the described method has only been applied to voice call events, other kinds of geo-located events could also be used to build the individual mobility profiles. The higher geo-located event density over space and time, the higher precision of the individual mobility profile in particular and the social dynamics in general.

This picture of social dynamics is of great interest for companies that want to customize services and applications, to better fit each individual's needs. Mobility profile based customer segmentation will probably be one of the direct applications of the mobility metrics.

Obviously, also governments and administrations can benefit from the knowledge of human dynamics seen as a whole. In that sense, our purpose is to link computed mobility profiles (both at individual and joint levels) with social variables, to be able to extract a clear picture of how population moves across space and time, and analyze the differences found for different ages, socioeconomic levels, regions, countries, and other segmentation criteria.

References

[1] M. C. González, C. A. Hidalgo, and A. L. Barabási, 'Understanding individual mobility patterns', *Nature 453, pp. 779-782* (Jun. 2008).

[2] C. Song, Z. Qu, N. Blumm, and A. L. Barabási, 'Limits of predictability in human mobility', *Science 327, pp. 1018–1021* (Feb. 2010).

[3] F. Calabrese, G. Di Lorenzo, and C. Ratti, 'Human mobility prediction based on individual and collective geographical preferences', *13th International IEEE Conference on Intelligent Transportation Systems (ITSC), pp. 312-317* (Sep. 2010).

[4] H. Firouzi, Y. Liu, and A. Sadrpour 'Mobility pattern prediction using cell-phone data logs', *EECS 545 Final Report* (2009).

[5] F. Calabrese, F. C. Pereira, G. Di Lorenzo, L. Liang, and C. Ratti, 'The geography of taste: analyzing cell-phone mobility and social events', *Proc. of the 8th International Conference on Pervasive Computing, pp. 22-37* (May, 2010).

[6] C. Cariou, C. Ziemlicki, and Z. Smoreda, 'Paris by night', *NetMob 2010, pp. 62-66* (May, 2010).

[7] R. A. Becker, R. Cáceres, K. Hanson, J. M. Loh, S. Urbanek, A. Varshavsky, and C. Volinsky, 'A tale of one city: using cellular network data for urban planning', *IEEE Pervasive Computing, Vol. 10, No. 4, pp. 18-26* (Oct.-Dec. 2011).

[8] B. Csáji, A. Browet, V. Traag, J.-C. Delvenne, E. Huens, P. Van Dooren, Z. Smoreda, and V. D. Blondel, 'Exploring mobility of mobile users', *NetMob 2011, pp. 35-37* (Oct. 2011).

[9] R. Becker, R. Cáceres, C. Han, K. Hanson, J. M. Loh, S. Urbanek, A. Varshavsky, and C. Volinsky, 'Finding meaningful usage clusters from anonymized mobile call detail records' *Netmob 2011, pp. 47-49* (Oct. 2011).

[10] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Vasharsvky, 'Identifying important places in people's lives from cellular network data', *Proc. of 9th International Conference on Pervasive Computing, pp. 133-151* (Jun. 2011).

 ISBN: 978-1-61208-226-4