Human mobility prediction based on individual and collective geographical preferences

senseable city lab:.::

Human Mobility Prediction based on Individual and Collective Geographical Preferences

Francesco Calabrese, Giusy Di Lorenzo, Carlo Ratti

***Abstract*— Understanding and predicting human mobility is a crucial component of transportation planning and management. In this paper we propose a new model to predict the location of a person over time based on individual and collective behaviors. The model is based on the person's past trajectory and the geographical features of the area where the collectivity moves, both in terms of land use, points of interests and distance of trips. The effectiveness of the proposed prediction model is tested using a massive mobile phone location dataset available for the Boston metropolitan area. Experimental results show good levels of accuracy in terms of prediction error and prove the advantage of using the collective behavior in the prediction model.**

I. Introduction

Understanding and modeling people's mobility is a crucial component of transportation planning and management. Methods currently being used are divided in two categories (see for instance [1]):

- Trip-based, where aggregated mobility is considered (trips between areas are usually estimated from surveys, and generate the Origin Destination matrices);
- Activity-based, where individual mobility is considered (each person is given a set of resources he/she has to access, and trips are generated as a consequence of this).

During recent years, it has also been of interest to model and predict individual mobility, both to provide guidance for cars in the form of smart GPS navigation systems (see e.g. AIDA[1]) or for personal recommendation systems (see e.g. iTour[2]).

Recently, a very large scientific community, from transportation to computer science and physics, has been working in this area. Individual mobility patterns has been studied in [2]–[4] using personal GPS data. Car trip destination prediction using GPS data has been studied in [5], [6]. Predicting people's location from WiFi and Bluetooth data has been one of the study conducted on the Reality Mining dataset and obtained encouraging results [7], [8]. All the studies mentioned above have typically required end-user consent, and so have relied on sample sizes of, at most, several hundred simultaneous users, limiting scalability and generalization of derived results. Recently, however, massive mobile phone location data have been studied and shown to have great potential to model human mobility [9], [10]. In particular in [10] the authors considered the theoretical limits of predictability of mobile phone users location based on their individual temporal patterns. The authors found that temporal patterns of locations are consistent over a large population, and then mobility prediction is indeed possible. The authors however did not provide any prediction algorithm, and did not consider another important parameter which is crucial in each individual mobility choice: the reasons for moving. In this paper we propose an individual mobility predictor that combines the person's past mobility choices and collective behavior in terms of:

- Propensity to change location;
- Type of geographical areas that are of interest for the collectivity at a give time, both in terms of land use, points of interest (POI) and distance of trips. This feature is assumed to be affecting the mobility choices as a proxy for activities.

The idea of using collective behavior is not new, and was already used in car trips prediction [11]. However, no information about geography has been combined so far. Trajectory patterns are instead considered to see whether different cars are moving in the same direction. Propensity to change location has already been considered in [12], considering a daily routine, and used to simulate human mobility. However, no prediction has been considered in that case, and no temporal considerations have been done in evaluating the simulation errors. Finally, no geographical information has been considered to improve the mobility simulation. The proposed model instead makes explicitly use of the geographical features of an area, as well as the habits of the collectivity.

The paper is structured as follows. Section II introduces the predictive model. Section III presents the data used in the proposed case study, and general statistics. Section IV presents the results of experiments made on the avaialble individual trajectories. Finally, some conclusions are given in Section V.

Notations. The following notation will be used in the rest of the paper: $N_n = \{1, 2, \ldots, n\}$, $\lfloor a \rfloor = max\{n \in N : n < a\}$, $\forall a \in R$.

II. Predictive model

Let us divide the space in n grid cells $i \in N_n$ and denote the time with k, which can be chosen for instance as days, hours, minutes. Let us hypothesize that we have a population of individuals for which we know their past trajectories. Let us denote each of the m considered individuals as u, and the location of a person u at time k as $x(u)_k = i$.

Authors are with the Senseable City Laboratory, Massachusetts Institute of Technology, 77 Massachusetts avenue, Cambridge, MA, USA fcalabre@mit.edu, giusy@mit.edu, ratti@mit.edu

[1]http://senseable.mit.edu/aida/

[2]http://www.itourproject.com/

Our problem is to predict the person's next location $x(u)_{k+1}$ given historical data.

We follow a probabilistic approach and define a probability for each grid cell j to be the next location of the person as function of the person and collective past behaviors. The cell with greatest probability is then chosen as predicted next location of the person. We hypothesize that the behavior is periodic over time, with a period T, e.g. weekly behavior. Then to predict the person's location at time k we use information about previous choices at times $k-T, k-2T, \ldots$. Periodic behavior is in line with what has been shown in [13], [14].

A. Individual behavior

We model the individual behavior as follows:

$$P_I(x_{k+1}=j|x_k=i)=\frac{\sum_{k_t=1}^{\lfloor k/T \rfloor} f_I(x_{k-Tk_t+1}=j|x_{k-Tk_t}=i)}{\lfloor k/T \rfloor} \quad \forall j \in N_n, \tag{1}$$

where the frequency on the right hand side is defined as:

$$f_I(x_{k+1}=j|x_k=i)=\begin{cases} 1 & if\ x_{k+1}=j\ and\ x_k=i \\ 0 & elsewhere \end{cases}.$$

The model says that the probability of a cell j to be the next destination of a person is equals to the frequency of visiting that cell starting from cell i during all previous periods $k-T+1, k-2T+1, \ldots$. If the person has never been in cell i at those times, the frequency is then computed as follows:

$$f_I(x_{k+1}=j|x_k=i)=\begin{cases} 1 & if\ x_{k+1}=j \\ 0 & elsewhere \end{cases}.$$

B. Collective behavior

We use the collective behavior to help predicting the likelihood the person changes location, and in that case, the type of place he/she will visit. We take into account the collective behavior in two elements:

- distances being traveled;
- types of places being visited. Since from the mobility traces we are not able to directly infer the activities that people make (many activities could be performed at a same location) we use information about an area's resources as a proxy for it.

In other terms, we design the probability to choose a given destination to be a function of the distance of the destination, the presence of points of interests similar to the ones the collectivity has visited, and the type of land use the collectivity has been in. The next subsections explain the three different contributions.

1) Distance: It is usually assumed that people tend to travel short distances, following a gravity-like model (the probability of a trip of length d is inversely proportional to d^2), or more sophisticated distance-based probability distributions [9]. It is also important to note that the length of trips that people make might depend on the time of the day or of the week when people start their trip. Using collective information, we can define a distance-based probability

$$P_D(x_{k+1}=j|x_k=i)=f_d(d_{ij},k),$$

where d_{ij} is the distance between cells i and j, and $f_d(d,k)$ is the normalized frequency of collective trips at distance d at times $k-T, k-2T, \ldots$. For instance,

$$f_d(0,k)=\frac{1}{m\lfloor k/T \rfloor}\cdot \cdot \sum_{u=1}^{m}\sum_{k_t=1}^{\lfloor k/T \rfloor} P(x(u)_{k-Tk_t+1}=x(u)_{k-Tk_t}).$$

where $P(x(u)_{k-Tk_t+1}=x(u)_{k-Tk_t})$ is the non-moving probability, e.g. the probability that a person does not change location.

2) Points of Interest: To each cell i a list of POIs is associated, belonging to Q categories. We can then characterize each cell with a vector $poi_i=\{poi_i(1),\ldots,poi_i(Q)\}$, where $poi_i(q)$ is the number of POIs of category q that are found in cell i. By analyzing the collective traces, we can infer the probability to find a person at a give time k close to a POI of category q:

$$f_{POI}(q,k)=\frac{\sum_{k_t=1}^{\lfloor k/T \rfloor}\frac{\sum_{u=1}^{m} poi_{x(u)_{k-Tk_t}}(q)}{\sum_{u=1}^{m}\sum_{q'=1}^{Q} poi_{x(u)_{k-Tk_t}}(q')}}{\lfloor k/T \rfloor}.$$

It then results, that given a cell j, the probability to find a person in that cell as function of the POIs available in that cell can be written as follows

$$P_{POI}(x_k=j)=\sum_{q=1}^{Q} f_{POI}(q,k)\frac{poi_j(q)}{\sum_{q'=1}^{Q} poi_j(q')}.$$

3) Land use: An analogous argument can be made for the land use, once we define the percentage of land use belonging to a grid cell i as $lu_i=\{lu_i(1),\ldots,lu_i(R)\}$:

$$P_{LU}(x_k=j)=\sum_{r=1}^{R} f_{LU}(r,k)\frac{lu_j(r)}{\sum_{r'=1}^{R} lu_j(r')},$$

where

$$f_{LU}(r,k)=\frac{\sum_{k_t=1}^{\lfloor k/T \rfloor}\frac{\sum_{u=1}^{m} lu_{x(u)_{k-Tk_t}}(r)}{\sum_{u=1}^{m}\sum_{r'=1}^{R} lu_{x(u)_{k-Tk_t}}(r')}}{\lfloor k/T \rfloor}.$$

Combining the three components, we obtain the following:

$$P_C(x_{k+1}=j|x_k=i)=P_{LU}(x_{k+1}=j)P_{POI}(x_{k+1}=j)\cdot \cdot P_D(x_{k+1}=j|x_k=i)\cdot \left(\sum_{j'=1}^{n} P_{LU}(x_{k+1}=j')P_{POI}(x_{k+1}=j')P_D(x_{k+1}=j'|x_k=i)\right)^{-1} \quad \forall j \in N_n. \tag{2}$$

Please note the scaling factor being used to ensure that $\sum_{j=1}^{n} P_C(x_{k+1}=j|x_k=i)=1$.

C. Combined behavior

We define the model to predict an individual behavior as a combination of individual and combined models (1) and (2):

$$P(x_{k+1}(u) = j | x_k(u) = i) = \quad (3)$$
$$(1 - \alpha(k)) P_I(x_{k+1}(u) = j | x_k(u) = i) +$$
$$+ \alpha(k) P_C(x_{k+1} = j | x_k = i),$$
$$\forall j \in N_n,$$

where the combination parameter $\alpha \in [0, 1]$ can change over time to model periods where individual behavior is more important, and periods where collective behavior is better able to model future decisions.

D. Observations

The proposed model could be further extended to take into account the following:

1) Temporal component: The prediction model starts working from time $k > T$, since it requires information from the first time period in order to predict what could happen in following periods. It might be useful to add a forgetting factor, to take into account that the most recent data might contain more useful information compared to older one. This element could be implemented by introducing a forgetting factor λ, and modifying the formula (1) as follows

$$P_I(x_{k+1} = j | x_k = i) = \frac{1}{\lfloor k/T \rfloor \sum_{k_t=1}^{\lfloor k/T \rfloor} \lambda^{\lfloor k/T \rfloor - k_t}} \cdot$$
$$\cdot \sum_{k_t=1}^{\lfloor k/T \rfloor} \lambda^{\lfloor k/T \rfloor - k_t} f(x_{k-Tk_t+1} = j | x_{k-Tk_t} = i),\ \forall j \in N_n.$$

Similar changes have to be made on the collective probabilities. λ generally ranges from 0 to 1 and the closest it is to 1, the more the older samples are considered. Samples older than $\tau = \frac{1}{1-\lambda}$ carry a weight that is less than about 0.3.

2) Collectivity selection: The prediction model is based on the collectivity's habits. It is then of importance to choose the right sample of people that represent the collectivity. The model indeed can be refined by customizing, for each person, the group of users to be considered as collectivities. Those people can be selected so that they behave similarly (based on past data). As example, if we consider a person who is a business man, we might want to chose a collectivity with similar habits in terms of working hours. At the same time, if the person is a retired person, it might be better to chose a collectivity of people who do not travel for work. It might also be possible to select different collectivity groups and give them different weights in the combined model, with higher weights to groups with behaviors more similar to the individual one. For instance, in the case of 2 collectivity groups, the combined model (3) will become:

$$P(x_{k+1}(u) = j | x_k(u) = i) =$$
$$(1 - \alpha_1(k) - \alpha_2(k)) P_I(x_{k+1}(u) = j | x_k(u) = i) +$$
$$+ \alpha_1(k) P_{C_1}(x_{k+1} = j | x_k = i),$$
$$+ \alpha_2(k) P_{C_2}(x_{k+1} = j | x_k = i),$$
$$\forall j \in N_n.$$

III. Case study

We performed a case study of the proposed prediction model using a massive mobile phone mobility dataset available for the Boston metropolitan area.

A. Datasets

1) Mobile phones location: The dataset consists of anonymous location estimations collected by AirSage[3] and generated each time a device connects to the cellular network, including:

- when a call is placed or received (both at the beginning and end of a call);
- when a short message is sent or received;
- when the user connects to the internet (e.g. to browse the web, or through email programs that periodically check the mail server).

In the remaining of the paper we will call these events *network connections*. These events represent a superset of the ones contained in the Call Details Records, previously considered in [9], [15]. Moreover, not only the id of the cell the mobile phone is connected to is available, but also an estimation of its position within the cell is generated through triangulation by means of the Airsage's Wireless Signal Extraction technology.

Each location measurement $m_i \in M$ is characterized by a position p_{m_i} expressed in latitude and longitude and a timestamp t_{m_i}. For each user, the locations measurements are then connected into a sequence $\{m_1 \rightarrow m_2 \rightarrow ..... \rightarrow m_n\}$ according to their timestamp.

From a spatial point of view, mobile phone-derived location data estimated by Airsage has a greater uncertainty range than GPS data, with an average of 320 meters and median of 220 meters as reported by AirSage based on internal and independent tests. Moreover, some peak errors appear when the user is connected to the network not using the closest cell phone tower. In these cases it can appear that the user travels for several kilometers in just a few seconds.

Based on the area covered by the mobile phone locations dataset, we analyzed the movements among areas in 8 counties in east Massachusetts (Middlesex, Suffolk, Essex, Worcester, Norfolk, Bristol, Plymouth, Barnstable) with an approximate population of 5.5 million people. The available dataset consists of 829 millions of anonymous location estimations - latitude and longitude - from close to 1 million devices (corresponding to a share of approximately 20% of the population) in 4 months.

[3] $http://www.airsage.com/$

For this dataset, we extract traces for 2,000 users, who make at least 100 network connections per day (with individual inter-event time below 1 hour in 75 percent of the cases). The raw mobile phone data is then processed expanding the methodology in [16] to obtain traces with sampling rate of 1 hour. Since localization errors might generate fictitious trips, we propose a pre-processing step in which we manipulate the data applying the same methodology used for analyzing GPS traces, see [5], [6]. The methodology is composed of the following steps:

- We infer measurement series $M_s = m_q, m_{q+1}, ..., m_z \in M^{z-q-1}$ where the user makes network connections over a certain time interval $\Delta T = t_{m_z} - t_{m_q} > 0$ into an area within the radius ΔS, i.e.

$$\max distance(p_{m_i}, p_{m_j}) < \Delta S \quad \forall \quad q \leq i, j \leq z$$

  The spatial threshold has been defined as $1km$, to take into account the localization errors estimated by Airsage.
- The points $M_s = m_q, m_{q+1}, ..., m_z \in M^{z-q-1}$ are fused together so that a single geographic region $p_s = (z-q)^{-1} \sum_{i=q}^{i=z} p_{m_i}$ (centroid of the points) can be regarded as a virtual location characterized by a group of consecutive location measurements. This location becomes the origin or destination of a trip.
- Once the virtual locations are detected, we can evaluate the stops (virtual locations) and trips as paths between user's positions at consecutive virtual locations.
- Each location is associated to a $500m \times 500m$ cell of a grid covering the whole Boston metropolitan area.

We consider the 2,000 users as collectivity, and compute the non-moving probability $f_d(0,k)$ for k corresponding to 1 hour period and $T = 168$ corresponding to 1 week, as shown in Figure 1. The one week period has been chosen according to results presented in previous work [13], [14]. The distance-depended normalized frequency $f_d(d_{ij}, k), d_{ij} > 0$, is shown in Figure 2 averaged over all values of k.

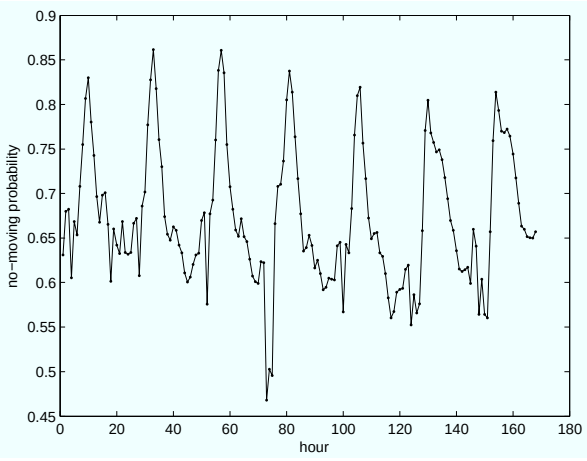


Fig. 1. Collective non-moving probability in one week period (starting Monday 12am).

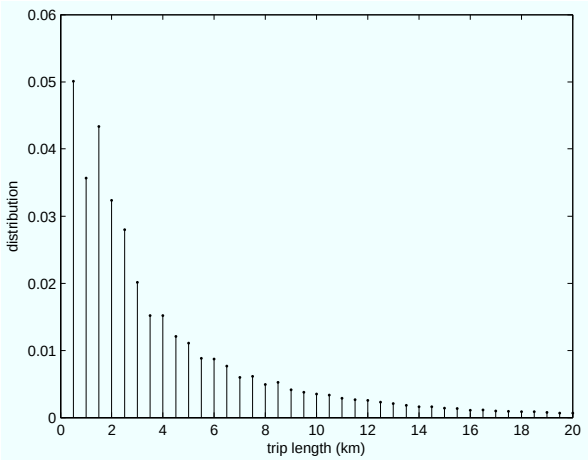


Fig. 2. Collective normalized frequency of trips as function of distance.

2) Landuse data: Landuse data has been collected from MassGIS[4] and grouped at the level of a $500m \times 500m$ cells grid. Different landuses have been grouped in 33 categories. Figure 3 shows the spatial distribution of land use, while Figure 4 shows the average preference of users for the different categories $1/T \sum_{k=1}^{T} f_{LU}(r,k), \quad r = 1, \ldots, R$. Looking at the most visited areas, Multi-family, high density residential and commercial combined have more than 50% of the preferences, followed by transportation and recreational areas.

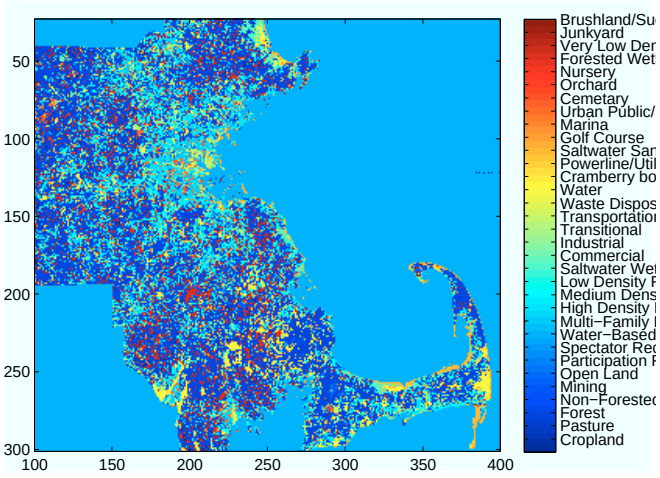


Fig. 3. Map of land uses (each cell in the map associated to the LU which covers the largest area in the cell).

3) POI data: POIs have been extracted from Yelp[5] at the same cells grid level, and grouped in 22 categories. Figure 5 shows the spatial distribution of points of interest, while Figure 6 shows the average preference of users for the different categories $1/T \sum_{k=1}^{T} f_{POI}(q,k), \quad q = 1, \ldots, Q$. Looking at the most visited areas, Food-related POIs cover

[4]http://www.mass.gov/mgis/lus2005.htm
[5]http://www.yelp.com/boston/

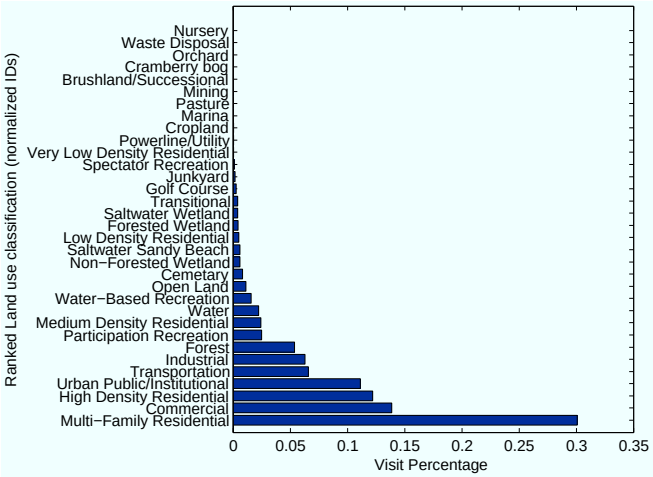


Fig. 4. Collective Land use preferences.

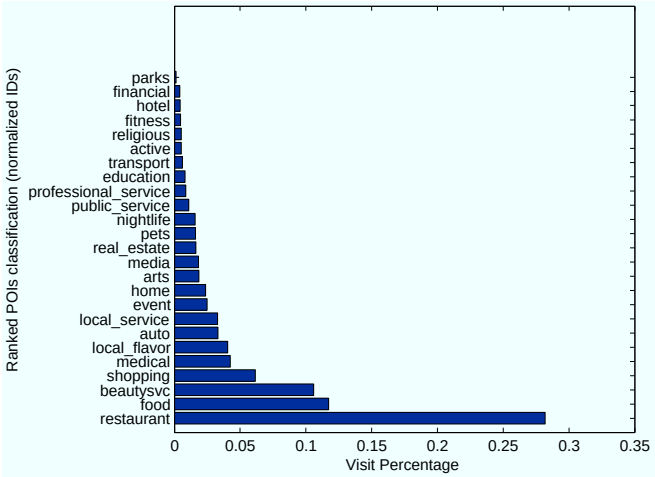


Fig. 6. Collective Point of Interest preferences.

almost 50% of the preferences, followed by beauty, shopping and medical. Other categories have very low impact.

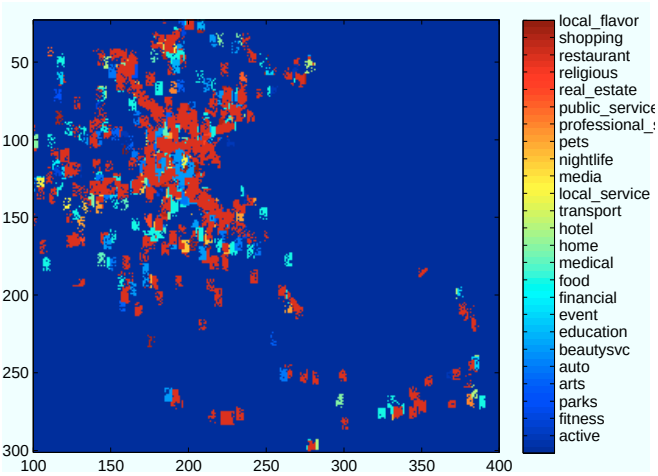


Fig. 5. Map of points of interest (each cell in the map associated to the POI which is most present). Dark blue corresponds to no points of interest.

IV. Experiments

To test the accuracy of the proposed prediction model, we implemented it for all 2000 traces, and evaluated the accuracy of the individual location prediction as error between predicted location $x^P_{k+1}(u) = P(x_{k+1} = j|x_k = i)$ and observed one $x_{k+1}(u)$. We measured the errors as:

$$e(k) = \left| x_{k+1}(u) - x^P_{k+1}(u) \right|.$$

We made different experiments for different values of α (kept constant over time) from 0 (individual only) to 1 (collective only), as shown in Figure 7. It results that $\alpha = 0.8$ allows obtaining the smallest mean error (1.34 km). All errors are less than half the mean error in case we always predict the individual most visited cell (mean error 2.8km). For the optimal value of α, Figure 8 shows that 60% of the errors are zero (we are able to correctly estimate the user's next location). We also compare results of the prediction made considering a collectivity composed only by the same user, so no information from other users. Results show an increased error due to the absence of global preferences from other users (see Figure 7).

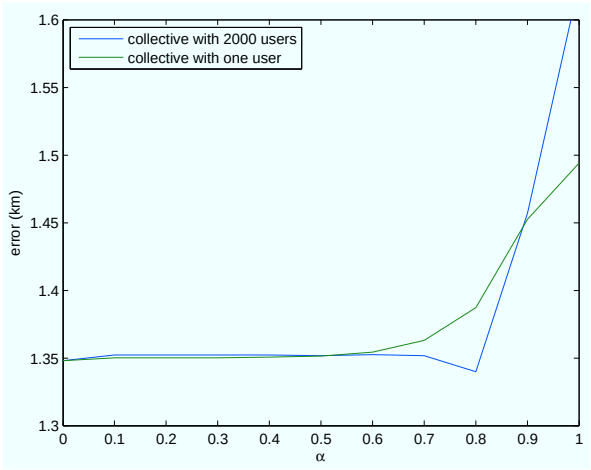


Fig. 7. Mean prediction errors $e(k)$ as function of α, considering collective behavior of 2000 users, or just the analyzed one.

A. Recommending places

To show an application of the proposed predictive model, we implemented a location-based service in the form of personal places recommender system. We created a service that recommends a list of places where he/she would possible go in the future and measure the error as minimum distance between recommended places and real one. As the length of the list increases, the error decreases (see Figure 9(a)). Already for lists of 3 elements, we are able to achieve almost the smallest error. In fact, the best prediction is almost always in the first three recommended places (see Figure 9(b)).

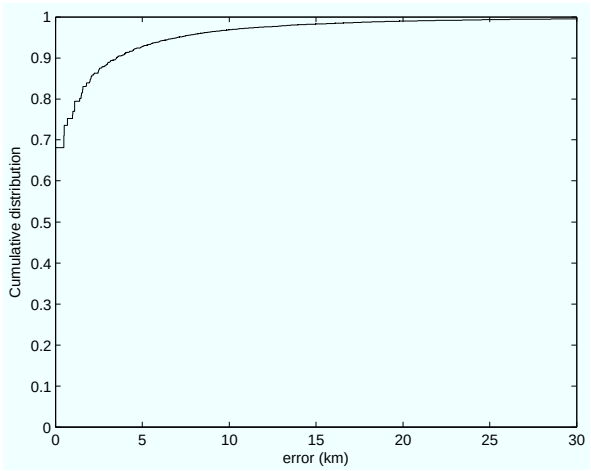


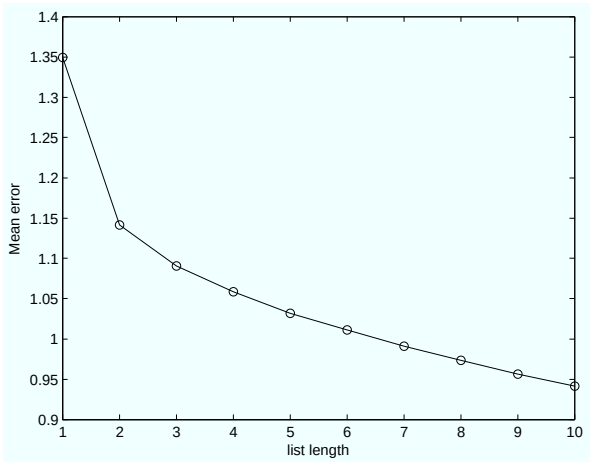
Fig. 8. Cumulative distribution of the prediction errors $e(k)$ for different values of α.

V. Conclusions

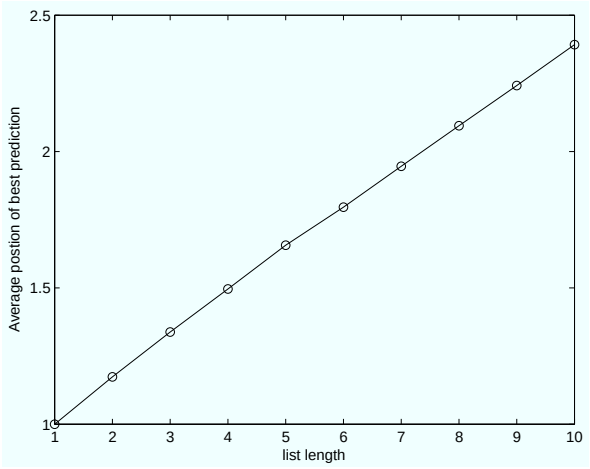
In this paper we have proposed a new model to predict the individual location of a person based on individual and collective behaviors. The model is based on the geographical features of the area where the person moves, both in terms of land use, points of interests and collectivity's habits. Using a massive mobile phone location dataset, we have tested the model for users living in the Boston metropolitan area. Experimental results show good levels of accuracy in terms of prediction error and prove the advantage of using the collective behavior in the prediction model. Future work will concentrate on improving the model and integrate it in current activity-based models.

VI. Acknowledgments

The authors gratefully acknowledge Airsage for providing the mobile phone location dataset.



(a) Mean errors of recommender system as function of list length.



(b) Mean position of best prediction as function of list length.

Fig. 9. Recommender system performance.

References

[1] M. Gendreau and P. Marcotte, *Transportation and Network Analysis: Current Trends*. Springer, 2002.

[2] Q. Li, Y. Zheng, X. Xie, Y. Chen, W. Liu, and W.-Y. Ma, "Mining user similarity based on location history," in *GIS '08: Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*. New York, NY, USA: ACM, 2008, pp. 1–10.

[3] Y. Zheng, Q. Li, Y. Chen, X. Xie, and W.-Y. Ma, "Understanding mobility based on gps data," in *UbiComp '08: Proceedings of the 10th international conference on Ubiquitous computing*. New York, NY, USA: ACM, 2008, pp. 312–321.

[4] L. Liao, D. J. Patterson, D. Fox, and H. Kautz, "Learning and inferring transportation routines," *Artif. Intell.*, vol. 171, no. 5-6, pp. 311–331, 2007.

[5] J. Krumm and E. Horvitz, "Predestination: Inferring destinations from partial trajectories," in *UbiComp*, 2006, pp. 243–260.

[6] J. Krumm, "Real time destination prediction based on efficient routes," in *Society of Automotive Engineers (SAE) 2006 World Congress*, 2006.

[7] N. Eagle and A. (Sandy) Pentland, "Reality mining: sensing complex social systems," *Personal Ubiquitous Comput.*, vol. 10, no. 4, pp. 255–268, 2006.

[8] A. Pentland, "Eigenbehaviors: identifying structure in routine," *Behavioral Ecology and Sociobiology*, vol. 63, pp. 1057–1066(10), May 2009.

[9] M. Gonzalez, C. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, 2008.

[10] C. Song, Z. Qu, N. Blumm, and A.-L. Barabasi, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, February 2010.

[11] A. Monreale, F. Pinelli, R. Trasarti, and F. Giannotti, "Wherenext: a location predictor on trajectory pattern mining," in *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM, 2009, pp. 637–646.

[12] P. Wang, M. Gonzalez, C. Hidalgo, and A.-L. Barabasi, "Understanding the spreading patterns of mobile phone viruses," *Science*, vol. 324, no. 5930, pp. 1071–1076, 2009.

[13] J. Reades, F. Calabrese, A. Sevtsuk, and C. Ratti, "Cellular census: Explorations in urban data collection," *IEEE Pervasive Computing*, vol. 6, no. 3, pp. 30–38, July-September 2007.

[14] F. Calabrese, J. Reades, and C. Ratti, "Eigenplaces: segmenting space through digital signatures," *IEEE Pervasive Computing*, 2010.

[15] J. White and I. Wells, "Extracting origin destinationn information from mobile phone data," in *Road Transportation and Control*, 2002.

[16] F. Calabrese, F. Pereira, G. DiLorenzo, L. Liu, and C. Ratti, "The geography of taste: analyzing cell-phone mobility and social events," in *International Conference on Pervasive Computing*, 2010.