How Long Are You Staying? Predicting Residence Time from Human Mobility Traces

Paul Baumann
Wireless Sensor Networks Lab
TU Darmstadt, Germany
paul.baumann@wsn.tu-darmstadt.de

Wilhelm Kleiminger
Institute for Pervasive Computing
ETH Zurich, Switzerland
kleiminger@inf.ethz.ch

Silvia Santini
Wireless Sensor Networks Lab
TU Darmstadt, Germany
santinis@wsn.tu-darmstadt.de

ABSTRACT

Predicting the arrival and residence time of individuals at their relevant places enables a plethora of novel applications. In this work we first analyze the theoretical predictability of arrival and residence times and then evaluate the performance of eight different residence time predictors. We show that these predictors tend to underestimate the time a user will spend at her relevant places.

Categories and Subject Descriptors

I.5.m [**PATTERN RECOGNITION**]: Miscellaneous

Keywords

Predictability of human mobility; Arrival time; Residence time

1. INTRODUCTION

Places where a person "*spends a substantial amount of time and/or visits frequently*" [2] are typically referred to as *relevant places*. On a typical day, an individual might move from one relevant place to the other and spend different amounts of time in each place. The time spent at each place is referred to as *residence time* [6, 10, 11]. The times of the day at which a person arrives at or leaves from a place are called *arrival time* and *departure time*, respectively [6, 10, 11]. The ability to predict when a person will arrive and how long she will stay at a specific place is fundamental to enable a number of applications like, e.g., smart heating control or urban navigation [5]. A number of algorithms that can perform these predictions have been presented in the literature [6, 11, 9]. This poster abstract presents our preliminary results on investigating both the theoretical and practical limits of the prediction performance achievable by arrival and residence time prediction algorithms.

To investigate the predictability of arrival and residence times we build upon recent work by Song *et al.* [1]. In their work Song *et al.* focus on the problem of predicting the next place that will be visited by a person, provided that the sequence of places she visited so far are known. In this context, they define the *predictability* Π as the "*upper bound that fundamentally limits any mobility prediction algorithm in predicting the next location based on historical records*" [7]. They also show how the value of Π can be computed from the entropy S_i of the sequence of places visited by a person – or *user* – i. Building upon this approach, we investigate the predictability of arrival and residence times. Our analysis allows to evaluate how close the performance of existing algorithms are to the theoretical limits. In particular, we investigate the actual performance achieved by eight algorithms in predicting residence times and show that these values are underestimated by most algorithms.

MobiCom'13, September 30–October 4, Miami, FL, USA.
ACM 978-1-4503-1999-7/13/09.
http://dx.doi.org/10.1145/2500423.2504583.

2. PREDICTABILITY OF ARRIVAL AND RESIDENCE TIMES

In this section we present the results of our analysis of the predictability of arrival and residence times. To this end, we first introduce the mathematical notation and describe the setup of our study. We would like to point out that we focus on the average predictability of a user over a given period of time – as in Song *et al.* 's work [1] – and not on the *momentary predictability* as done in [8, 4].

2.1 Terminology and notation

We indicate with L_j, $j = 1 : N_L$, the j-th relevant place of a user i and define the set $\mathcal{L} = \{L_1, L_2, \ldots, L_{N_L}\}$ as the set of N_L places relevant to user i.[1] The places are ordered according to the total amount of time spent at each place, e.g., L_1 is the place at which the user spends most of her time (L_1 is typically the home). Locations visited by the user that cannot be associated to any relevant place are assumed to be "irrelevant" to the user and are marked using the symbol L_x. We further define the *mobility trace* MT of a user i as the sequence of places visited by the user during the observation period T_{obs}. The observation period is thereby virtually divided in *slots* of length Δ_s. For instance, if the observation period is one day and the length of a slot is 15 minutes then the mobility trace will be a vector of 96 elements, whereby the elements of the vector take values in $\{\mathcal{L}, L_x\}$.

We further define the *arrival time trace* $AT(L_j)$ as the vector containing the ordered sequence of arrival times at place L_j. The length of the vector is not fixed a priori since the number of arrival events occurring in the mobility trace might vary from user to user and from place to place. The values of the elements of $AT(L_j)$ are the indexes of the time slots at which arrival events takes place. Using again the example above, in which $T_{obs} = 1$ day and $\Delta_s = 15$ minutes, the mobility trace has 96 elements. The first element (index 1) corresponds to the time slot spanning the period from 00:00 to 00:15 and the last elements (index 96) to the slot from 11:45 p.m to 12:00 p.m.. Consider for instance a user for whom L_1 is the home and that during the observation

[1]To simplify the notation, we omit the use of the subscript i. The quantities defined here however always refer to a specific user i.

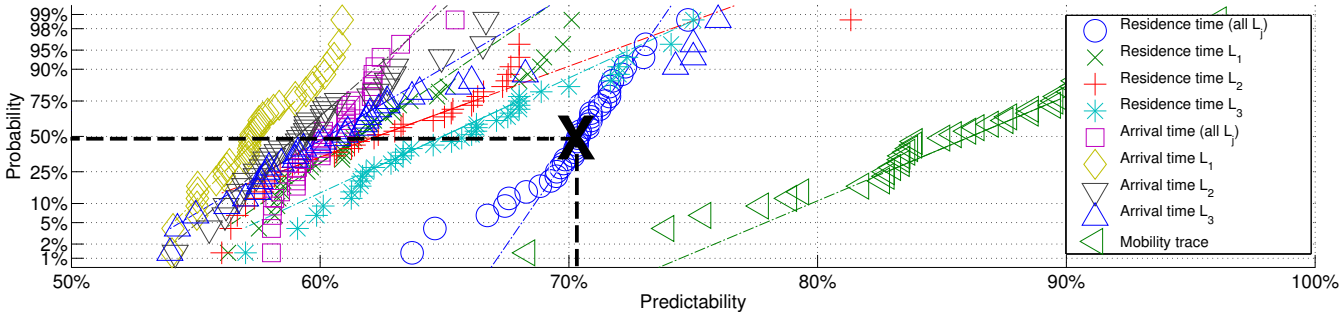


Figure 1: Normal probability plot of arrival time and residence time predictability vectors.

period had come back home at 17:00, then went out again and returned at 22:00. The corresponding arrival time trace for L_1 will be: $AT(L_1) = [68, 88]$. Arrival time traces of two different places can be combined: $AT(L_j, L_k)$ will for instance indicate the ordered union of the two vectors $AT(L_j$ and $AT(L_k)$. $AT(\mathcal{L})$ indicates the union of the arrival time trace of all L_j, $j = 1 : N_L$.

Finally, we define the *residence time trace* $RT(L_j)$ as the vector containing the sequence of residence times at place L_j. The length of a residence time is thereby indicated as the number of slots the user stays at place L_j before moving to another place. Using again the example mentioned above and assuming the user stayed at home between 17:00 and 18:30 and then again from 22:00 to the end of the observation period (midnight), then: $RT(L_1) = [6, 8]$.

2.2 Evaluation setup

We run our analysis on data collected using smart phones in the context of the Lausanne Data Collection Campaign (LDCC) [3]. The subset of LDCC data available for this study consists of records collected from 38 users over about 1.5 years including a large amount of location data. We use the Wi-Fi scans available in the data set as input to the PlaceSense algorithm by Kim *et al.* [2]. This way, we derive the set $\mathcal{L}$ of relevant places for each user. We use the same parameter settings as in [2] apart from the value of the *sensitivity* parameter which is set to 30% as in [9]. We then derive the mobility trace of each user using a Δ_s of 15 minutes. The observation period varies depending on the amount of data available for each user. Minimum, maximum, average, and median of the observation periods is 86, 407, 203, and 189 days, respectively.

From these mobility traces we derive the vectors $RT(\mathcal{L})$, $RT(L_1)$, $RT(L_2)$, $RT(L_3)$, $AT(\mathcal{L})$, $AT(L_1)$, $AT(L_2)$, and $AT(L_3)$. We compute the predictability of these traces obtaining one data point per vector and per user. For instance, we compute the predictability associated with the sequence of values in $RT(\mathcal{L})$ for each of the 38 users and combine these values in the *predictability vector* $\Pi_{RT(\mathcal{L})}$. Similarly, we compute the predictability vectors $\Pi_{RT(L_1)}$, $\Pi_{RT(L_2)}$, $\Pi_{RT(L_3)}$, $\Pi_{AT(\mathcal{L})}$, $\Pi_{AT(L_1)}$, $\Pi_{AT(L_2)}$, $\Pi_{AT(L_3)}$. The predictability of the arrival time traces and residence time traces is computed using the same method used by Song *et al.* to compute the predictability of mobility traces [1][2].

2.3 Evaluation results

Figure 1 shows the normal probability plot of the predictability vectors $\Pi_{RT(\mathcal{L})}$, $\Pi_{RT(L_1)}$, $\Pi_{RT(L_2)}$, $\Pi_{RT(L_3)}$, $\Pi_{AT(\mathcal{L})}$, $\Pi_{AT(L_1)}$, $\Pi_{AT(L_2)}$, $\Pi_{AT(L_3)}$ and of the predictability of the mobility traces. On a normal probability plot[3], data showing a normal distribution fits on a line. The x-axis indicates the predictability, computed as described above. The y-axis indicates the probability that the arrival time, residence time, or mobility trace of a user shows a predictability equal or lower than the corresponding value on the x-axis. For instance, the large 'X' marker in Figure 1 shows that the residence time of 50% of the users has a predictability of 71% or less (when all locations in $\mathcal{L}$ are considered). This implies that the prediction accuracy achievable by an algorithm that predicts the residence time does not exceed 71% for about 50% of the users.

Figure 1 also shows that the overall predictability of the mobility traces is higher than that of the arrival times and residence times. This means that it is in general easier to predict the next location of a user rather than the arrival or residence time at specific locations. The curves in Figure 1 further show that the predictability of the arrival times at location L_1 is low (about 60%) and lower than the predictability of the arrival times at locations L_2 and L_3. Furthermore, the predictability of the arrival times is in general lower than the predictability of the residence times.

Figure 2 shows the Cumulative Distribution Function (CDF) of the residence time at places L_1, L_2, L_3, and L_x averaged over all users in the data set. The curve corresponding to L_1 is significantly "smoother" than the L_2, L_3, and L_x. This indicates that the amount of time users spend at L_1 varies more than the time spent at other locations. This offers an explanation of the predictability values observed above: the high dispersion of residence times at L_1 increases the number of potentially predictable value, leading to an overall lower predictability.

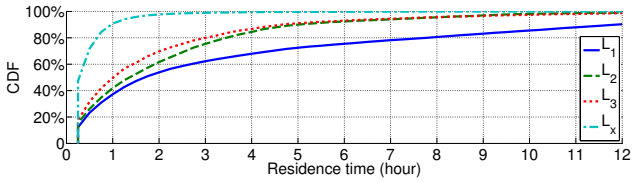


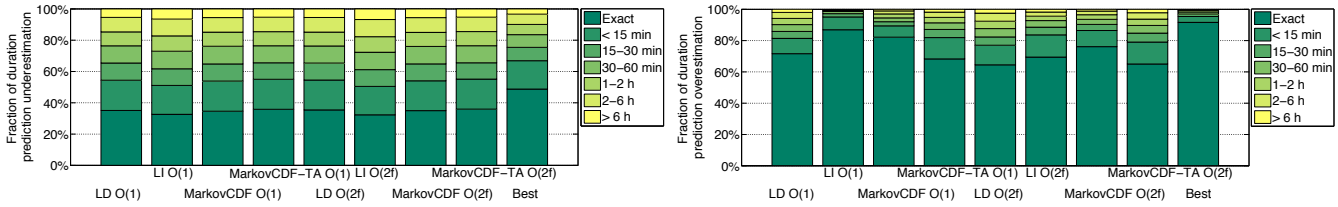
Figure 2: Cumulative distribution function of residence time for L_1, L_2, L_3, and L_x.

3. PREDICTION OF RESIDENCE TIME

After exploring the theoretical bounds for the predictability of users' arrival and residence times, in this section we focus on the accuracy achieved by existing, practical residence time prediction algorithms.

[2]The results by Song *et al.* are obtained under the assumption that the sequence of locations is the realization of a stationary ergodic process. This assumption is not likely to be fulfilled if long sequences (e.g., over several years) are considered. In our work, however, we consider shorter sequences (e.g., several months).

[3]http://www.mathworks.de/de/help/stats/normplot.html



(a) Correct and underestimated residence time predictions.

(b) Correct and overestimated residence time predictions

Figure 3: Fractions of residence time under- and overestimated predictions.

3.1 Prediction algorithms

We consider eight different residence time predictors selected from the literature [6, 11]. We use the location-dependent (LD), location-independent (LI), MarkovCDF, and MarkovCDF time-aided (MarkovCDF-TA) predictors. The Markov 1st O(1) and 2nd O(2) order location-dependent (LD) predictor with the fallback option O(2f) are described in [6, 11] and are used without further modifications. As a LI predictor we use the LD predictor by removing the explicit location dependency. For the predictors MarkovCDF and MarkovCDF time-aided (i.e., the residence time depends on the arrival time), we use the implementations proposed in [11]. We further consider a fictive algorithm – dubbed *Best* – which always takes, among the predictions computed by the other algorithms, the one known to result in the smallest prediction error. We use the same notation and data as described in Section 2 as well as the residence time traces $RT(L_j)$ ($RT(\mathcal{L})$ for the LI predictor).

3.2 Results

We use the algorithms described above to predict the residence time of each user at the locations L_j, $j = 1, \ldots, N_L$. We compute the average prediction error of each algorithm and split the results into correct predictions, overestimation, and underestimations, i.e., zero, positive, or negative error values. Figure 3(a) shows the percentage of both correct and underestimated predictions (the percentage is meant over the total of correct and underestimated predictions) for each of the nine considered algorithms. Figure 3(b) shows the same data when along with correct also overestimated predictions are considered.

Apart from the *Best* algorithm – which, as expected, always shows the best performance – all the predictors exhibit similar percentages of prediction error ranges when the error is underestimated. The LI O(1) predictor however generates the highest number of underestimated predictions. Overall, the number of overestimations with respect to the correct predicitons is much smaller than the number of underestimations. In particular, location-independent Markov 1st order predictor together with the fictive *Best* approach produce the smallest amount of overestimated predictions.

This allows us to conclude that the considered residence time predictors tend to underestimate the time a user will spend at the location L_j. Further investigations on whether this consideration can be generalized and on what are the reasons for this behavior are part of our future work.

4. CONCLUSIONS

This poster abstract describes preliminary results on the analysis of predictability of arrival time and residence time and on the evaluation of the actual performance of residence time predictors. We observe that arrival time traces have an overall lower predictability than residence time traces and that the higher the amount of time a user spends at a specific place, the lower is the corresponding arrival and residence time predictability. Our analysis of the performance of residence time predictors shows that most predictors tend to underestimate the time a user will spend at her relevant places.

5. ACKNOWLEDGEMENTS

This work has been partially supported by the Collaborative Research Center 1053 (www.maki.tu-darmstadt.de) funded by the German Research Foundation and by the Priority Program Cocoon (www.cocoon.tu-darmstadt.de) funded by the LOEWE research initiative of the state of Hesse, Germany.

6. REFERENCES

[1] C. Song *et al.* Limits of Predictability in Human Mobility. *Science*, 327(5968):1018–1021, 2010.

[2] D.H. Kim *et al.* Discovering Semantically Meaningful Places from Pervasive RF-Beacons. In *11th Intl. Conf. on Ubiquitous Computing (UbiComp'09)*, 2009.

[3] J. Laurila *et al.* The Mobile Data Challenge: Big Data for Mobile Computing Research. In *10th Intl. Conf. on Pervasive Computing (Pervasive'12)*, 2012.

[4] J. McInerney *et al.* Exploring Periods of Low Predictability in Daily Life Mobility. In *10th Intl. Conf. on Pervasive Computing (Pervasive'12)*, June 2012.

[5] J. Scott *et al.* PreHeat: Controlling Home Heating Using Occupancy Prediction. In *13th Intl. Conf. on Ubiquitous Computing (UbiComp'11)*, Sept. 2011.

[6] L. Song *et al.* Predictability of WLAN Mobility and its Effects on Bandwidth Provisioning. In *25th Intl. Conf. on Computer Communications (INFOCOM'06)*, 2006.

[7] M. Lin *et al.* Predictability of Individuals' Mobility with High-resolution Positioning Data. In *14th Intl. Conf. on Ubiquitous Computing (UbiComp'12)*, 2012.

[8] P. Baumann and S. Santini. On the Use of Instantaneous Entropy to Measure the Momentary Predictability of Human Mobility. In *14th IEEE Workshop on Signal Processing Advances in Wireless Communications (SPAWC'13)*, 2013.

[9] P. Baumann *et al.* The Influence of Temporal and Spatial Features on the Performance of Next-place Prediction Algorithms. In *2013 ACM Intl. Joint Conf. on Pervasive and Ubiquitous Computing (UbiComp'13)*, 2013.

[10] S. Scellato *et al.* NextPlace: A Spatio-temporal Prediction Framework for Pervasive Systems. In *9th Intl. Conf. on Pervasive Computing (Pervasive'11)*, 2011.

[11] Y. Chon *et al.* Evaluating Mobility Models for Temporal Prediction with High-granularity Mobility Data. In *10th Intl. Conf. on Pervasive Computing and Communications (PerCom'12)*, 2012.