An Empirical Study of Human Mobility Patterns

Douglas do Couto Teixeira[1], **Jussara M. Almeida**[1]

[1]Departamento de Ciência da Computação – Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte – MG – Brazil

{douglas,jussara}@dcc.ufmg.br

Abstract. *This paper documents our efforts towards understanding which factors are more relevant in human mobility prediction. Our work is divided into two phases. First, we characterize a dataset consisting of more than 200,000 user check-ins in the Foursquare social network, inferring important patterns in human mobility. Second, we use factorial design to quantify the importance of several types of contextual information in human mobility prediction. Our results show that the proximity of the users possible next check-in to his or her home and work location are the most important factors (among the ones we analyzed) to be used by mobility prediction models.*

1. Introduction

The interaction of Computer Science with other realms of knowledge has become more intense over the years. One example of this interaction is Urban Computing, an interdisciplinary research area concerned with applying Computer Science techniques to solving urban problems. In a broad sense, Urban Computing can be defined as the process of acquisition, integration, and analysis of large volumes of heterogeneous data produced by several sources (sensors, mobile devices, cars, buildings, people, etc.) in urban spaces [Zheng et al. 2014].

The increased amount of information available in urban areas poses several challenges for the Urban Computing community. One of these challenges is understanding people's mobility patterns. By understanding such patterns, it is possible to create models to predict a person's location at a given moment in time. The knowledge of people's locations has applications in many areas, such as urban planning, traffic engineering, epidemiology, recommender systems, and advertising [Zheng et al. 2014, Ma et al. 2013, Hasan et al. 2013].

Several human mobility prediction models have been developed in the past years [Lu et al. 2013, Munjal et al. 2011, Silveira et al. 2016, Zheng et al. 2014, Dong et al. 2013]. The underlying assumption of these models is that the more information we have about the users' mobility patterns, the more precise the models are. To the best of our knowledge, there is no work measuring how much impact each type of contextual information has in human mobility prediction. To try to remedy that, we conducted an empirical study of real world data to try to determine the impact of different types of information in human mobility prediction, namely, we investigate the impact of temporal data (weekday/weekend and day/night information), and distance from the person's home and work. Previous works [Song et al. 2010, Lu et al. 2013, L. Silveira 2015, Silveira et al. 2016] used mobile phone records to study human mobility. However, it has been shown [Jurdak et al. 2015] that online social networks can be a reliable source for studying human mobility. In this paper, we used a dataset of check-ins in the Foursquare social network to conduct an experimental study of human mobility patterns.

Our goal in this paper is to analyze data from the Foursquare online social network to try to quantify the impact of two types of contextual information (temporal and geographical) in the precision of prediction models. In particular, we will propose a simple mobility prediction model that, given the user's current location and the places she visited in the past, tries to infer her next location. We also evaluate our model using temporal information, i.e., whether the user's next check-in will happen during the day or at night, and geographical information i.e., whether the user's next check-in will happen near her home or work location. To measure the impact of these types of information in mobility models, we conduct an experimental design and compute the effect of each type of information in the precision of our model.

The rest of this report is organized as follows. In Section 2 we discuss related work and explain the different approaches to human mobility prediction. In Section 3, we describe the dataset used in this work and try to get insights that will later help us determine what types of information are more important to mobility models. Section 5 analyzes the impact of two types of contextual information (temporal and geographical) on mobility prediction. To conduct such analysis, we propose a simple prediction model that uses a person's location preferences (a history of her previously visited location) to infer the location she will visit next. In Section 6, we use factorial designs to infer the impact of several categories of contextual information in human mobility prediction. Section 7 summarizes the lessons we learned from our study and discusses future works.

Our work makes the following contributions:

- We study mobility patterns using a real world dataset of more than 200,000 user check-ins in the Foursquare social network.
- We quantify the importance of several types of contextual information on the precision of mobility models.
- We show that the knowledge of a person's work location is the single, most important factor in influencing the precision of mobility prediction models.
- We propose several variations of a mobility prediction model that uses contextual information to predict where a person will go next given her current location.

2. Related Work

2.1. Human Mobility Prediction

The process of urbanization has increased the number of urban areas and modernized people's lives. There are estimates saying that by 2050, 70% of the people will live in cities [Zheng et al. 2014]. However, urbanization brings along several problems: pollution, increased energy consumption, and traffic jams, to name a few. With the process of urbanization and the advancements in technology, we have more information available about urban areas. And the growing volume of data creates the need for techniques to analyze such data and extract useful information from it. Urban Computing tries to develop such techniques to help identify and solve problems brought up by urbanization.

The applications of Urban Computing are many, but one area in particular, namely human mobility, has been the target of steady efforts by the academic community.

Zheng *et. al.* [Zheng and Xie 2011] used GPS data from vehicles to identify transportation problems in Beijing. Toole *et. al.* [Toole et al. 2012] used call record data to measure the concentration of people in regions of a city through time. Herring *et. al.* [Herrera et al. 2010] show how people move from one place to another in a city. Hasan *et. al.* [Hasan et al. 2013] used social networks data to understand patterns of mobility and human activity. And these are only some examples of works that try to explain patterns of human mobility. Our goal in this study is to contribute towards these same efforts not by implementing a sophisticated model, but by creating a simple one that will allow us to analyze how much precision each type of contextual information adds to prediction models. In other words, our focus is on the factors that affect the precision of mobility models, and not on specific prediction techniques. We find that although other types of information are relevant in mobility prediction, people prefer to visit places that are near their current location, their home, or their work.

We have mentioned several applications of Urban Computing in identifying and solving problems in urban areas. There are, however, many other studies that try to infer more general, theoretical results from mobility data. Jiang *et. al.* [Jiang et al. 2013] established general patterns of mobility: people tend to make many short trips during the week, and a few longer ones on weekends and holidays —these patterns are sometimes called Lèvy flights[1]. To do that, they divided a city in regions and analyzed how people move from one place to another during a period of time.

The goal of this paper is to analyze data from the Foursquare online social network to try to determine the impact of two types of contextual information in the precision of prediction models. We use a simple mobility prediction model that, given the user's current location and the places she visited in the past, tries to infer his/her next location. To measure the impact of these types of information in mobility models, we conduct an experimental design and compute the effect of each type of information in the precision of our model.

2.2. Experimental Designs

The goal of an experimental design is to gather the maximum amount of information with the minimum number of experiments [Jain 1991]. This goal can be achieved in different ways, depending on what we wish to measure and how many experiments we are willing to perform. There are several types of experimental designs that present a compromise between the number of variables being measured and the number of experiments required.

The most common type of experimental design is called factorial project. In this type of experiment, there is a *response variable*, which is the result that being measured in the experiment, and k other variables, referred to as *factors*, that affect the response variable. Each of the factors has two alternative values, called *levels*. Two factors are said to interact if the effect of one depends on the level of the other.

In general, a factorial project helps in the analysis of the effects of factors that might affect the performance of a system. It also allows us to determine if a factor's effect is significant or if the observed results may be attributed to random variations caused by measurement errors and uncontrolled parameters [Jain 1991]. Furthermore, by computing

[1] https://www.nature.com/articles/srep09136

a factor's effect on the observed results, a factorial design helps in the decision of whether the difference in performance of two levels of a factor is significant enough to justify its further examination.

In this work we quantify the importance of several factors in determining a person's next location. To accomplish that, we measure how much precision the use of each one of several types of contextual information adds to a simple prediction model[2].

3. Dataset Characterization

In this paper, we use a dataset provided by Yang *et. al.* [Yang et al. 2015], consisting of long-term check-in data from to New York city collected from Foursquare during the period of April 2012 to February 2013[3]. The dataset contains a total of 227,428 check-ins, 1,083 users, and 38,333 venues, distributed over 251 categories. Our decision of using this dataset was based on two facts: first, it consists of real-world data, spanning a long period of time; and second, it gathers data at the user level (as opposed to aggregated data). This second fact is of more importance because in 2015, many social networks created mechanisms to prevent the gathering of data of users. For instance, Foursquare prohibits the gathering of a user's check-in without the user's consent, and the use of mechanisms to circumvent this prohibition is considered a violation of Foursquare's API terms of service.

The distribution of check-ins under each category varies broadly. The average number of check-ins per category is about 2,000, but the most popular categories have as many as 16,000 check-ins. Analyzing the distribution of check-ins per category gives us an overview of the places people tend to check-in at. However, to better understand patterns of human mobility, we need to take a closer look into the data. That is, we also need to analyze the distribution of check-ins per venue and per user. This analysis suggests that there are a few venues with a high number of check-ins (more than 1,000 check-ins), but the majority of venues have relatively few check-ins (100 check-ins, on average).

So far we have looked at how the check-ins are distributed in terms of categories and venues. Looking at how the users check-in at the venues will give us a complimentary look into the data and will help us understand the users behavior in a location-based social network such as Foursquare. Figure 1 shows the cumulative distribution function of the number of check-ins per user. The red vertical line shows the average number of check-ins among all users in our study.

4. Patterns of Human Mobility

In the previous section, we analyzed the dataset and extracted general information about users and venues. In this section, we will focus on understanding patterns of human mobility. In particular, we are interested in discovering the pieces of information that make mobility models more precise.

Mobility prediction models use a variety of information to infer a person's next location, which is a function of where said person is at a particular moment, whether it

[2] The scripts used in this paper can be found at: https://github.com/dougct/mob-fact

[3] The dataset is available here: https://sites.google.com/site/yangdingqi/home/foursquare-dataset

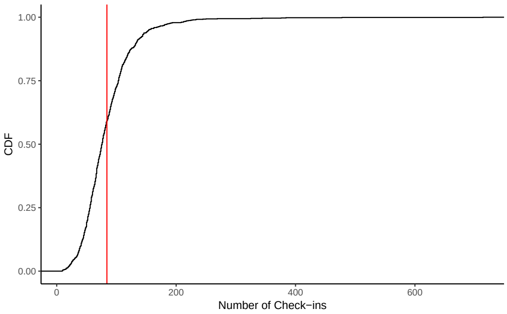

Figure 1. Number of Check-ins per User.

is a weekday or a weekend, whether it is day or night, and how far is the next location from the person's home or work. In this section we will study what role these types of information play in mobility models.

One of the pieces of information mobility prediction models use to infer the next location of a user is distance from where she currently is to her next possible location. Previous work [Jiang et al. 2013] has shown that people tend to make many short trips during the week (probably going from home to work and back) and a few longer trips (usually on weekends or holidays), which suggests that, most of the time, they will travel small distances between check-ins. We compute the average distance each user traveled between check-ins using the same procedure as Xin Lu *et. al.* [Lu et al. 2013]. According to this procedure, such distance is given by:

$$\overline{D}(i) = \frac{1}{n-1}\sum_{j=2}^{n} \mid m_j - m_{j-1} \mid,$$

where n is the total number of locations visited by user i, $\overline{D}(i)$ is the mean distance traveled by user i, and m_j is the jth latitude/longitude location visited by user i. We assume that m_{j-1} and m_j were visited one after another. If m_j and m_{j-1} occur on different days, and assuming the user goes back to her house every day, $m_j - m_{j-1}$ provides a lower bound on the actual distance traveled by the user between those two check-ins.

Figure 2 shows that, on average, people travel between 2.5 and 3 kilometers between each check-in, and the majority of check-ins are less than 5 kilometers far from the previous one.

Another important piece of information used by mobility prediction models is whether the prediction of the user's next location is to happen on a weekday or a weekend. It has been shown that the precision of mobility prediction models is inversely proportional to the number of previously unvisited venues in a dataset [Lu et al. 2013]. In theory, when a person visits a certain location she has not been to before, it is impossible for a prediction model to correctly guess that location. Thus, every previously unvisited location a person goes to generates a miss-prediction for the mobility model.

Our hypothesis is that the number of previously unvisited venues is higher on weekends due to people's less structured routine in those days. Therefore, we analyzed our dataset to check whether the number of previously unvisited venues is higher on week-

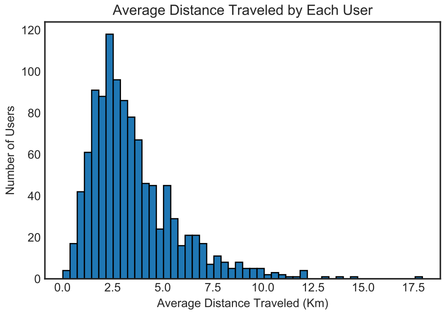

Figure 2. Average Distance Traveled Between Check-ins.

days or weekends. We found that the probability of a user visiting a previously unvisited is about 30% on weekdays and about 20% on weekends. However, previous work has shown that this difference in the number of previously unvisited venues does not impact the precision of prediction models [Song et al. 2010]. Thus, one our goals is to check the impact of weekday/weekend on the precision of mobility models using our experimental design framework. As we will show later, our results agree with the literature regarding this subject.

Yet another important piece of information used by mobility prediction models is the knowledge of the user's home and work locations. There are several heuristics for determining these locations based on check-in data. In this work we use the same strategy used by Jiang *et. al.* [Jiang et al. 2013]: we assume that the place where the user checks-in more often between 9pm and 6am is her home and the place where she checks-in more often between 6am and 9pm is her work place. Our analysis indicates that not only do the user's check-ins tend to be near each other, but they also tend to be near the user's home or work location. Thus, mobility models should favor places that are near the user's current location, home, and work when making predictions about her next venue.

Figure 3 and Figure 4 show the distribution of check-ins according to the distance from the user's home and work location. As we see from these figures, people tend to visit places that are near where they live or work. Our study tries to determine not only whether this is true, but also if knowing people's home location is more important to prediction models than knowing the person's work location.

In this paper, we analyze the impact that two types of contextual information, namely temporal and geographical information, have on the precision of mobility prediction models. To accomplish that, we create a simple prediction model and conduct an experimental design to measure the impact of these types of information on its precision. We argue that our model is simple enough to allow us to analyze the importance of the features decoupled from a specific prediction technique.

5. Mobility Prediction Model

Given a person's location preferences (the history of the locations she has been to before) and the her current location, mobility prediction models try to foresee where said person will go next. The person's location preferences are usually inferred from the places she visited before, and the person's current location is given by the latitude and longitude of

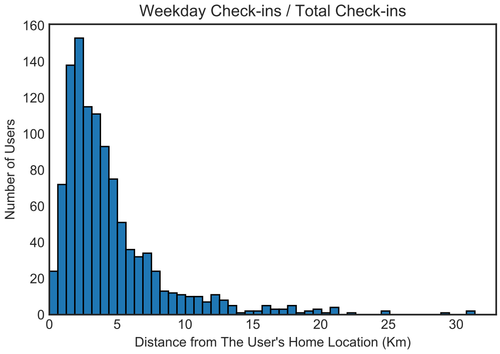

Figure 3. Check-in Distance from the User's Home Location.

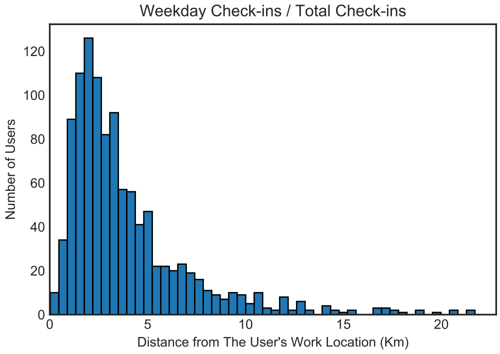

Figure 4. Check-in Distance from the User's Work Location.

the place she is at a given moment. Our model uses this information, along with contextual information, to try to predict the person's next move.

Although our dataset provides the exact location (venue) users visit, it is a common practice in mobility prediction models to try to infer the next region a given person will visit. We use the a similar strategy here: we consider that when the next visited venue is within a given distance (radius), our model makes a correct prediction. The size of the radius determines a region, and the size of such region varies among mobility models. For instance, Song *et. al.* used a region of size $3km^2$ in their seminal paper [Song et al. 2010]. In this paper we experimented with several region sizes to see their impact on the precision of prediction models.

To conduct our experiments, we split the dataset in two parts: train and test. The training set contains 90% of the data, and the test set contains 10%. We use the training set to built the users' location preferences. When evaluating our model, given a user's current location in the test set, we predict her next location using the information we learned in the training set. If the person is at location x, we consider as possible next locations only those locations that the person visited after visiting x before. If there is no record of a location being visited after x, our model makes a miss-prediction. When considering a given factor, we further filter the possible next locations to account for that factor. For instance, given that location x was visited at night, we would only consider locations visited after x *and* at night. The factors work both as contextual information and as filters.

Our prediction model works as follows. Given the person's current location and the contextual information associated with that location, we compute the most likely region she will visit next. To perform this computation, we analyze the places said person visited in the past and build a list S of tuples $\langle l_i, l_j \rangle$, where l_i is where the person was before she went to l_j. Note that the frequency with which the person visits a given location is already encoded in this computation. For instance, if the person always checks-in at her house after getting back from work, the tuple $\langle work\_location, home\_location \rangle$ will appear many times in the list S.

For each check-in $\langle x, w \rangle$ in the test set, we take the person's current location x, and filter the list S to contain only tuples of the form $\langle x, y \rangle$. Thus, we have a list of candidate locations $S_c = (\langle x, y_1\rangle, \langle x, y_2\rangle, \cdots, \langle x, y_n\rangle)$. Finally, we randomly select a tuple $\langle x, y_i\rangle$ from the training set according to the probability distribution of the user's location history. If the distance between y and w is less than a given radius, our model makes a correct prediction. Otherwise, it miss-predicts the user's next location. It is important to point out that this distance calculation is an adaptation to our model, since our dataset consists of *exact* locations (venues) and not regions. It is common practice to work with regions because most mobility models use data from call detail records, in which a person's location is said to be within the same region as the antenna that registered the call.

Though simple, our model allows for the addition of contextual information to enhance the predictions. If we want to use contextual information when making predictions, we only need to change the way we build the set of candidate locations S_c. Before, given a person's current location x, we built $S_c = (\langle x, y_1\rangle, \langle x, y_2\rangle, \cdots, \langle x, y_n\rangle)$. Now, suppose we want to add the information of whether the check-in will happen during the day or at night. For instance, given that a person is at location x during the day, we build S_c as follows: $S_c = (\langle x_{day}, y_1\rangle, \langle x_{day}, y_2\rangle, \cdots, \langle x_{day}, y_n\rangle)$. That is, we only consider as candidates the tuples that have x as the current venue and that the transition from x to the next venue occurred during the day. The same reasoning is applied when considering the other types of contextual information.

Although simple in its formulation, our model allows us to measure the impact of contextual information in mobility prediction. Its simplicity makes it possible for us to distance ourselves from the intricacies of a single prediction technique and makes it possible to focus on the features, i.e., the types of information models use in their predictions.

In our prediction model every new (previously unvisited) venue will cause a miss-prediction, because we only consider as a next venue candidate those venues that have been visited before. In fact, previously unvisited venues hamper the precision of every mobility prediction model because these models rely on the user's location preferences, which in turn are obtained from her previously *visited* places. Thus, as explained in Section 4, the amount of previously unvisited venues is inversely proportional to a model's precision.

6. Factorial Project

In this section we try to measure the importance of the several types of contextual information in understanding human mobility patterns. To accomplish that, we use factorial designs [Jain 1991] and measured the impact of several factors (as well as the interaction among them) in human mobility prediction. In particular, we are interested in studying how much precision the use of each factor adds to a mobility prediction model.

Our study is based on a 2^k factorial design. The factors in our design are the types of contextual information that affect the precision of mobility models, namely: (i) the frequency with which a person visits a particular venue, (ii) the time of the visit (day or night), and (iii) the distance of the venue to the person's home or work location. The levels of each factor indicate whether the information it represents was used. To make the presentation easier, we will use letters to refer to the several types of contextual information (factors), according to Table 1. In the experiments for our factorial design, we used a region size of 2 km, and we consider as near home or work a check-in that occurs at a distance less than 1 km.

Information (factors)	**Symbol**
Frequency of visits	Baseline
Same Day (weekday/weekend)	A
Same Time (day/night)	B
Near Home	C
Near Work	D
Same Day & Same Time	AB
Same Day & Near Home	AC
Same Day & Near Work	AD
Same Time & Near Home	BC
Same Time & Near Work	BC
Near Home & Near Work	CD
Same Day & Same Time & Near Home	ABC
Same Day & Same Time & Near Work	ABD
Same Day & Near Home & Near Work	ACD
Same Time & Near Home & Near Work	BCD
Same Time & Same Time & Near Home & Near Work	ABCD

Table 1. Types of Contextual Information.

The idea behind a 2^k factorial project is that the value of the response variable can be obtained from the values of the variables x_A, x_B, and x_C using a non-linear additive model as follows:

$$\begin{aligned} y = q_0 + q_A x_A + q_B x_B + q_C x_C + q_D x_D + q_{AB} x_{AB} + q_{AC} x_{AC} \\ + q_{AD} x_{AD} + q_{BC} x_{BC} + q_{BD} x_{BD} + q_{CD} x_{CD} + q_{ABC} x_{ABC} \\ + q_{ABD} x_{ABD} + q_{ACD} x_{ACD} + q_{BCD} x_{BCD} + q_{ABCD} x_{ABCD} \end{aligned} \tag{1}$$

In the equation above, y is the response variable (the percentage of correct predictions), q_0 is the average precision of our prediction model independent of factor levels, and each of the q_* variables is the effect of one or more factors.

Using both Table 1 and the equation above, we evaluate our prediction model using a 2^k factorial design, following the procedure described by Raj Jain [Jain 1991]. According to this procedure, we build Table 2 with the grand mean (column I) and columns for all combinations of factors. The grand mean of the predictions is the average prediction obtained without considering specific levels of each factor. The relative importance of each factor is based on how the results for a specific factor deviates from the ground mean. Each of the other columns contains the results for a specific level of a factor, where a value of 1 indicates that the type of information the factor represents was used by the prediction model and a value of -1 says that the model did not use that information. In other words, this is a binary procedure: either we use a type of contextual information or we do not.

In addition to having one column for each of the factors alone, Table 2 also contains one column for each of the possible interaction among factors —two factors interact if the level of one changes the effect of the other. The last column of the table shows the response variable, i.e, the percentage of correct predictions made by our model when a given arrangement of factors was used. Our results are shown in Table 2.

I	**A**	**B**	**C**	**D**	**AB**	**AC**	**AD**	**BC**	**BD**	**CD**	**ABC**	**ABD**	**ACD**	**BCD**	**ABCD**	*y*
1	-1	-1	-1	-1	1	1	1	1	1	1	-1	-1	-1	-1	1	33.06
1	1	-1	-1	-1	-1	-1	1	-1	1	1	1	1	1	-1	-1	33.49
1	-1	1	-1	-1	-1	1	-1	1	-1	1	1	1	-1	1	-1	33.41
1	1	1	-1	-1	1	-1	-1	-1	-1	1	-1	-1	1	1	1	34.15
1	-1	-1	1	-1	1	-1	-1	1	1	-1	1	-1	1	1	-1	50.89
1	1	-1	1	-1	-1	1	-1	-1	1	-1	-1	1	-1	1	1	52.10
1	-1	1	1	-1	-1	-1	1	1	-1	-1	-1	1	1	-1	1	51.00
1	1	1	1	-1	1	1	1	-1	-1	-1	1	-1	-1	-1	-1	52.10
1	-1	-1	-1	1	1	1	1	-1	-1	-1	-1	1	1	1	-1	53.00
1	1	-1	-1	1	-1	-1	1	1	-1	-1	1	-1	-1	1	1	53.03
1	-1	1	-1	1	-1	1	-1	-1	1	-1	1	-1	1	-1	1	52.83
1	1	1	-1	1	1	-1	-1	1	1	-1	-1	1	-1	-1	-1	52.61
1	-1	-1	1	1	1	-1	-1	-1	-1	1	1	1	-1	-1	1	57.31
1	1	-1	1	1	-1	1	-1	1	-1	1	-1	-1	1	-1	-1	57.00
1	-1	1	1	1	-1	-1	1	-1	1	1	-1	-1	-1	1	-1	57.17
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	57.69

Table 2. Factorial Project. Column I represents the ground mean of the predictions. Each of the other columns contains the results for a specific level of a factor.

The next step in the procedure is to compute the values of variables q_0, q_A, q_B, q_C, q_D, q_{AB}, q_{AC}, q_{AD}, q_{BC}, q_{BD}, q_{CD}, q_{ABC}, q_{ABD}, q_{BCD}, and q_{ABCD}, which are the average precision obtained for each factor (or interaction of factors). We show the results of this computation in the table below:

q_0	q_A	q_B	q_C	q_D	q_{AB}	q_{AC}	q_{AD}	q_{BC}	q_{BD}	q_{CD}	q_{ABC}	q_{ABD}	q_{ACD}	q_{BCD}	q_{ABCD}
48.80	0.218	0.067	5.605	6.2775	0.0487	0.096	0.0150	-0.216	-0.0725	-3.392	0.041	0.023	-0.046	0.127	0.093

Table 3. Values of the variables in our factorial project.

Now we compute the *effects* of each of the factors, that is, the importance of the factor relative to the grand mean, as follows:

$$SST = 2^k \times q_*^2 = 1319.6$$
$$SSA = 2^k \times q_A^2 = 0.765$$
$$SSB = 2^k \times q_B^2 = 0.072$$
$$SSC = 2^k \times q_C^2 = 502.6$$
$$SSD = 2^k \times q_D^2 = 630.5$$
$$SSAB = 2^k \times q_{AB}^2 = 0.038$$
$$SSAC = 2^k \times q_{AC}^2 = 0.148$$
$$SSAD = 2^k \times q_{AD}^2 = 0.003$$

$$SSBC = 2^k \times q_{BC}^2 = 0.748$$
$$SSBD = 2^k \times q_{BD}^2 = 0.084$$
$$SSCD = 2^k \times q_{CD}^2 = 184.1$$
$$SSABC = 2^k \times q_{ABC}^2 = 0.027$$
$$SSABD = 2^k \times q_{ABD}^2 = 0.009$$
$$SSACD = 2^k \times q_{ACD}^2 = 0.034$$
$$SSBCD = 2^k \times q_{BCD}^2 = 0.260$$
$$SSABCD = 2^k \times q_{ABCD}^2 = 0.140$$

Now we normalize the computations made above, which gives us the the variance explained by each factor. This variance tells us how much, a factor influences the response variable.

$$\begin{aligned}
SSA/SST &= 0.058\% & SSBD/SST &= 0.006\% \\
SSB/SST &= 0.005\% & SSCD/SST &= 13.95\% \\
SSC/SST &= 38.09\% & SSABC/SST &= 0.002\% \\
SSD/SST &= 47.77\% & SSABD/SST &= 0.0006\% \\
SSAB/SST &= 0.002\% & SSACD/SST &= 0.002\% \\
SSAC/SST &= 0.011\% & SSBCD/SST &= 0.019\% \\
SSAD/SST &= 0.0002\% & SSABCD/SST &= 0.010\% \\
SSBC/SST &= 0.056\% & &
\end{aligned}$$

From the factorial project, we conclude that the distance from person's home or work are the two most important factors in determining the next venue this person will visit. The other two factors, namely the information about whether the check-in happens on a weekday or weekend, and during the day or at night, have less importance than the distance from home and work. It is important to emphasize that our results do not imply that other factors are not important for the precision of prediction models. We only show that weekday/weekend and day/night information are less important to determine a person's next location than the information about the person's home or work place.

Our experiments, in agreement with previous work [Wang et al. 2015], suggest that, in general, people will visit regions that a) they have visited before, and b) that are close to where they live or work. Looking at the variance explained by the factors, one notices that the distance from people's work place (variable SSD) is the single most important factor (among the four factors analyzed here) in determining the region this person will visit next. This result indicates that people's social interactions at work may play an important role in their life outside of work. Previous works either did not quantify the importance of each factor or placed the knowledge of the user's home location and work location as equally important for mobility prediction models. Here we quantity the importance of each factor and show that the knowledge of a person's work location is the single most important factor influencing the precision of mobility prediction models.

To further investigate the impact of the distance in the precision of our model, we varied the distance from the user home and work to the next check-in and measured how it affected the precision of our prediction model. To produce these results, we kept the region radius fixed at 2 km, and measured the precision of our model varying the distance from the user's home and work.

Figure 5 shows that varying the distance from the user's home does affect the precision of the model. In our simple model the variation in this distance affected the precision of the model by as much as 4%. We believe that this value would increase considerably if we were using a more sophisticated prediction model instead of our basic model. Our results also show that increasing the distance from the user's home when making predictions does not always increase the precision of the model. After a certain point, increasing this distance actually decreases the model's precision.

Figure 6 shows the results for when we vary the distance from the user's work location. Similarly to what happens when we vary the distance from the user's home,

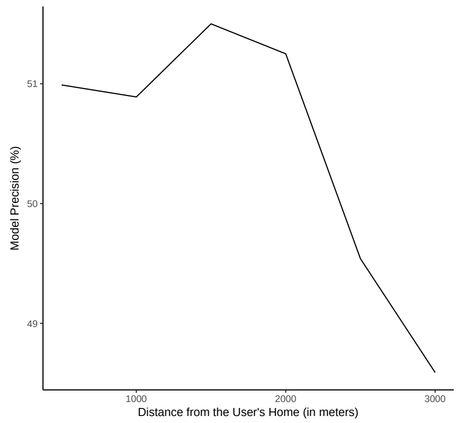

Figure 5. Precision versus Distance form the User Home.

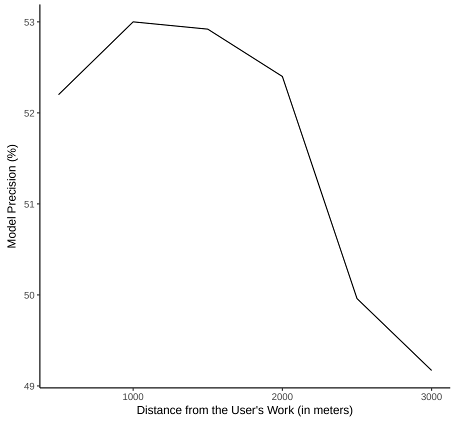

Figure 6. Precision versus Distance form the User Home.

varying the distance from the user's work location impacts the precision of the prediction model. Also, increasing the distance from the user's work only increases the precision of the model up to a point, after which the precision drops as we increase the distance.

Knowing that distance (from home or work) is the most important factor in mobility prediction, we went ahead and also analyzed how much the radius of the region impacts a model's precision. In Figure 7, we increase the size of the region and see whether our model becomes more precise. We find that the larger the size of the region, the more precise the model. However, there is a trade-off between increasing the region size and getting useful information from the model. As an extreme example, suppose we make the region as large as the whole city. This would make our model very accurate, in the sense that it would guess the correct region almost always, but the information that the user's next check-in will happen within the confines of the city is not very useful.

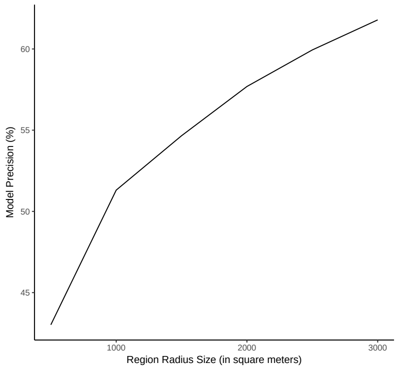

Figure 7. Model Precision versus Region Radius.

7. Conclusion

We have studied a dataset of more than 200,000 user check-ins in the Foursquare social network to try to understand human mobility patterns. Our work validates some conclusions drawn in previous studies: (i) users make many short trips during the week and a few long trips, usually on weekends or holidays; (ii) there are important factors to consider when trying to predict a user next location, such as whether it is a weekday or weekend, day or night, and whether the place is near the user's home or work.

But we have also found a couple of nuances on such conclusions. First, we claim that, unlike mentioned in previous work that deals with regions [Song et al. 2010], the precision of mobility models should be lower on weekends compared to weekdays. This phenomenon happens because the rate of previously unvisited venues is higher on weekends, hence reducing the precision of mobility models. Second, we found that, among the factors considered in mobility models, both the size of region considered by the prediction model and the distance the candidate venue is to the person's home and work location are the most important, though the other factors (weekday/weekend, and day/night information) are also important. Mobility models that aim to be precise cannot afford to discard any of these informations when making predictions.

In the future we would like to study other factors that may improve the precision of mobility models, namely contacts among users and heterogeneous sources of data. It is important to consider the user's contacts (one's friends on Twitter, for instance) because people tend to visit the places their friends visit [Cho et al. 2011]. Heterogeneous information is important because it provides complimentary views on how people move through a city. There have been studies showing the impact of information from multiple sources on mobility models before [Silveira et al. 2016]. The work of Silveira *et. al.* [L. Silveira 2015, Silveira et al. 2016] shows that adding more sources of data increases the precision of mobility models, which raises the question: To which degree does the addition of information from multiple sources increase the precision of prediction models?

References

Cho, E., Myers, S. A., and Leskovec, J. (2011). Friendship and mobility: User movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 1082–1090, New York, NY, USA. ACM.

Dong, W., Duffield, N., Ge, Z., Lee, S., and Pang, J. (2013). Modeling cellular user mobility using a leap graph. In *Proceedings of the 14th International Conference on Passive and Active Measurement*, PAM'13, pages 53–62, Berlin, Heidelberg. Springer-Verlag.

Hasan, S., Zhan, X., and Ukkusuri, S. V. (2013). Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. In *Proceedings of the 2Nd ACM SIGKDD International Workshop on Urban Computing*, UrbComp '13, pages 6:1–6:8, New York, NY, USA. ACM.

Herrera, J. C., Work, D. B., Herring, R., Ban, X. J., Jacobson, Q., and Bayen, A. M. (2010). Evaluation of traffic data obtained via gps-enabled mobile phones: The mobile century field experiment. *Transportation Research Part C: Emerging Technologies*, 18(4):568 – 583.

Jain, R. (1991). *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling*. Wiley.

Jiang, S., Fiore, G. A., Yang, Y., Ferreira, Jr., J., Frazzoli, E., and González, M. C. (2013). A review of urban computing for mobile phone traces: Current methods, challenges and opportunities. In *Proceedings of the 2Nd ACM SIGKDD International Workshop on Urban Computing*, UrbComp '13, pages 2:1–2:9, New York, NY, USA. ACM.

Jurdak, R., Zhao, K., Liu, J., AbouJaoude, M., Cameron, M., and Newth, D. (2015). Understanding human mobility from twitter. *PLOS ONE*, 10(7):1–16.

L. Silveira, J. Almeida, H. N. A. Z. (2015). Mobdatu: Um novo modelo de previsao de mobilidade humana para dados heterogeneos. *Simposio Brasileiro de Redes de Computadores*.

Lu, X., Wetter, E., Bharti, N., Tatem, A. J., and Bengtsson, L. (2013). Approaching the limit of predictability in human mobility. *Scientific Reports*, 3.

Ma, S., Zheng, Y., and Wolfson, O. (2013). T-share: A large-scale dynamic taxi ridesharing service. In *Data Engineering (ICDE), 2013 IEEE 29th International Conference on*, pages 410–421.

Munjal, A., Camp, T., and Navidi, W. C. (2011). Smooth: A simple way to model human mobility. In *Proceedings of the 14th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, MSWiM '11, pages 351–360, New York, NY, USA. ACM.

Silveira, L. M., Almeida, J. M., Marques-Neto, H. T., Sarraute, C., and Ziviani, A. (2016). Mobhet: Predicting human mobility using heterogeneous data sources. *Computer Communications*, 95:54–68.

Song, C., Qu, Z., Blumm, N., and Barabási, A.-L. (2010). Limits of predictability in human mobility. *Science*, 327(5968):1018–1021.

Toole, J. L., Ulm, M., González, M. C., and Bauer, D. (2012). Inferring land use from mobile phone activity. In *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*, UrbComp '12, pages 1–8, New York, NY, USA. ACM.

Wang, H., Xu, F., Li, Y., Zhang, P., and Jin, D. (2015). Understanding mobile traffic patterns of large scale cellular towers in urban environment. In *Proceedings of the 2015 Internet Measurement Conference*, IMC '15. ACM.

Yang, D., Zhang, D., Zheng, V. W., and Yu, Z. (2015). Modeling user activity preference by leveraging user spatial temporal characteristics in lbsns. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 45(1):129–142.

Zheng, Y., Capra, L., Wolfson, O., and Yang, H. (2014). Urban computing: Concepts, methodologies, and applications. *ACM Trans. Intell. Syst. Technol.*, 5(3):38:1–38:55.

Zheng, Y. and Xie, X. (2011). Learning travel recommendations from user-generated gps traces. *ACM Trans. Intell. Syst. Technol.*, 2(1):2:1–2:29.