Technical University of Denmark

DTU

Understanding predictability and exploration in human mobility

Cuttone, Andrea; Jørgensen, Sune Lehmann; González, Marta C.

Link back to DTU Orbit

DTU Library
Technical Information Center of Denmark

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE

Open Access

CrossMark

Understanding predictability and exploration in human mobility

Andrea Cuttone[1,3*] , Sune Lehmann[1,2] and Marta C. González[3,4,5]

*Correspondence:
andreacuttone@gmail.com
[1]DTU Compute, Technical University
of Denmark, Richard Petersens Plads
Building 324, Kgs. Lyngby, Denmark
[3]Department of Civil and
Environmental Engineering and
Engineering Systems,
Massachusetts Institute of
Technology, 77 Massachusetts
Avenue, 1-290, Cambridge, 02139,
United States
Full list of author information is
available at the end of the article

Abstract

Predictive models for human mobility have important applications in many fields including traffic control, ubiquitous computing, and contextual advertisement. The predictive performance of models in literature varies quite broadly, from over 90% to under 40%. In this work we study which underlying factors - in terms of modeling approaches and spatio-temporal characteristics of the data sources - have resulted in this remarkably broad span of performance reported in the literature. Specifically we investigate which factors influence the accuracy of next-place prediction, using a high-precision location dataset of more than 400 users observed for periods between 3 months and one year. We show that it is much easier to achieve high accuracy when predicting the time-bin location than when predicting the next place. Moreover, we demonstrate how the temporal and spatial resolution of the data have strong influence on the accuracy of prediction. Finally we reveal that the exploration of new locations is an important factor in human mobility, and we measure that on average 20-25% of transitions are to new places, and approx. 70% of locations are visited only once. We discuss how these mechanisms are important factors limiting our ability to predict human mobility.

Keywords: human mobility; next-location prediction; predictability

1 Introduction

Billions of personal devices, ranging from in-car GPS to mobile phones and fitness bracelets, connect us to the cloud. These ubiquitous interconnections between the physical and the digital world open up a host of new opportunities for predictive mobility models. Each user of a device produces rich information that allows us to measure their daily mobility routine. This type of knowledge, especially when arising from large numbers of individuals, is expected to impact a wide range of areas such as health monitoring [1], ubiquitous computing [2, 3], disaster response [4], or smart traffic management [5]. Recent contributions to mobility modeling come from computer science [6–8], transportation engineering [9, 10], geographic information sciences [11, 12], and complexity sciences [13–15]. The state-of-the-art for mobility modeling has developed rapidly over the past decade, but further work is needed, especially to tackle the problem of individual predictability.

In the literature, human mobility has been studied using a multitude of proxies (for example call detail records (CDR), GPS, WiFi, travel surveys), and a variety of techniques

Springer

have been suggested for predictive models, such as Markov chains, Naive Bayes, artificial neural networks, time series analysis. Studies report varying results for the predictive power of these models, with accuracy from over 90% to under 40%.

In this paper we set out to uncover the reasons behind these surprisingly large differences in performance via a systematic investigation of the factors that may influence estimates of mobility predictability. The key contributions of this paper are:

1. We investigate which factors influence performance in the reported cases of predictability. Showing that the differences in results are driven by the following questions: (a) Does the analysis concern an *upper bound* of predictability, or actual next-place prediction? (b) How is the prediction problem formulated? E.g. is the goal to predict the next location, or is the goal to identify location in the next time-bin? (c) What is the spatial resolution of the data source? E.g. is the analysis based on GPS vs. CDR data? (d) What is the temporal resolution of the data source e.g. minutes, hours?
2. We quantify the amount of explorations and locations visited only once, and show that these are key limiting factors in the accuracy of predictions for individual mobility.
3. We measure the predictive power of a number of contextual features (e.g. social proximity, time, call/SMS).
4. We study the problem of predictability of human mobility using a novel, longitudinal, high-precision location dataset for more than 400 users.

The rest of the paper is organized as follows. We first provide an overview of related work in the field of predictions of human mobility. Next, we introduce the dataset and describe the preprocessing steps. In the subsequent section we describe the baseline models, and compare their performances. Finally we introduce the exploration prediction problem and report the performance of the predictive models.

2 Related work

In a seminal paper Song *et al.* [13] investigate the limits of predictability of human mobility, using Call Detail Records (CDR) as proxy for human movement. In their analysis, the authors discretize location into a sequence of places, and estimate an upper limit for the predictive performance using Fano's inequality on the temporal entropy of visits. Their results show that for a majority of users, this upper bound is surprisingly high (93%). This framework has been further explored to refine the estimate of the upper limit. Specifically, Lin et al. [16] study the effects of spatial and temporal resolution on the predictability limit, Smith et al. [17] consider spatial reachability constraints when selecting the next place to visit, and obtain a tighter upper bound of 81-85%, and Lu et al. [4] analyze the predictability of the population of Haiti after the earthquake in 2010, and find an upper limit of predictability of around 85%. The work described above focuses on estimating an upper limit of predictability for an individual based on an estimate of the entropy their trajectory.

When the topic is actual prediction performance, the most studied models are Markov chains, where the probability of the next location is assumed to depend only on the current location. Markov chains have been applied to a variety of data sets. Lu et al. [18] applied Markov chain models to CDR-based locations in Cote D'Ivore, with a prediction goal of estimating the last location of the day at the prefecture (county) level. Under these conditions the models perform extremely well, reaching an accuracy of over 90%. In [19] the

authors apply the Markov models to WiFi traces at Darthmouth campus and find that the best performing model is order 2 and has a median accuracy of about 65-72%. Krumme et al. [20] applied Markov models on shopping locations inferred from credit card transactions, and found that models trained on the aggregate data from a large number of people perform better than individual models. Bapierre et al. [21] applied a variable-order Markov chain to the Reality Mining [6] and Geolife [22] datasets.

Another frequently used type of prediction framework is based on naive Bayes models, where the probability of next location is factorized as independent probabilities for a number of context variables. Gao et al. [23] applied this approach to the Nokia Data Challenge dataset [24] using time and location features, and obtained an accuracy of approximately 50%. Do et al. [25] applied the same technique but used a larger number of features including also SMS, calls and Bluetooth proximity, and obtained an accuracy of approximately 60%. In a subsequent paper [26] the same authors explore a kernel density estimation approach for improving performance.

A number of more complex methods have also been explored in the literature, including non-linear time series [27], Principal Component Analysis [28], Gaussian Mixtures [29] and Dynamic Bayesian Networks [30].

While recent work on predictability has resulted in richer methods and incorporated interesting new features such as social contacts, they have not identified the intrinsic characteristics of human mobility that form the basis for the limitations in predicting the next visited location. In this paper we focus on that aspect, showing that over 53 weeks, individuals visit on average 200 unique locations, of which 70% of them are visited only once. Despite most of the trips taking place among 30% of their (recurrent) locations; we can only predict the occurrence of an exploration with at best 41% of accuracy. Separating visited locations into explorations and recurrent visits may advance the current methods in this area.

3 Materials and methods

3.1 Data description

In this study we analyze a dataset from the Copenhagen Network Study [31]. The project has collected mobile sensing data from smartphones for more than 800 students at the Technical University of Denmark (DTU). The data sources include GPS location, Bluetooth, SMS, phone contacts, WiFi, and Facebook friendships. Data collection was approved by the Danish Data Protection Agency, and written informed consent has been obtained for all study participants.

For this study we focus on the location data, which is collected by the smartphones with a frequency of one sample every 15 minutes. Each location sample contains a timestamp, a latitude and longitude, and an accuracy value. Location is determined by the best available provider, either GPS or WiFi, with a median accuracy of ≈ 20 meters; more than 90% of the data samples are reported to have an accuracy better than 40 meters. For individual participants, there may be periods missing data. These periods can occur for a number of reasons, for example due to a drained battery, the phone being switched off, the location probe being disabled, or due to software issues. Since we are interested in reconstructing mobility histories without large gaps, we select the longest period that has at least one sample in 90% of the 15-minutes time-bins for each participant. Moreover we consider only participants that have at least 3 months of such contiguous data. We are left with

454 users, with data collection periods of data ranging from three months to one year (Additional file 1 Figure 1).

The data is mainly concentrated in Denmark where the study takes place, but because students use their phones during travel, the dataset spans several other countries as well (Additional file 1 Figure 2).

In this work we are interested in the location prediction task. This task can be broadly stated as follows: given your location history, how well can we predict your future location? The specific details of how this question is implemented have a core impact on the prediction accuracy. Below we investigate how various factors, e.g. spatial and temporal data resolution play a role in determining the reported accuracy for the same underlying dataset.

Because the prediction task can be stated in many different ways, we start the discussion by analyzing different formulations of the problem. In terms of spatial prediction it is possible to discretize space in grid cells, Voronoi cells or define places using a clustering method. In terms of temporal prediction we could decide to predict a location in the next time-bin - within a certain time horizon - or as the next visited place. Below we examine two of the most common problem formulations: *next-cell* and *next-place*. In the *next-cell* formulation we discretize space into grid cells, and we predict the cell in the next time-bin. In the *next-place* formulation we detect visits to places and we predict the next visited place. The following sections provide details on the two alternative formulations, and show how each formulation affects the prediction task.
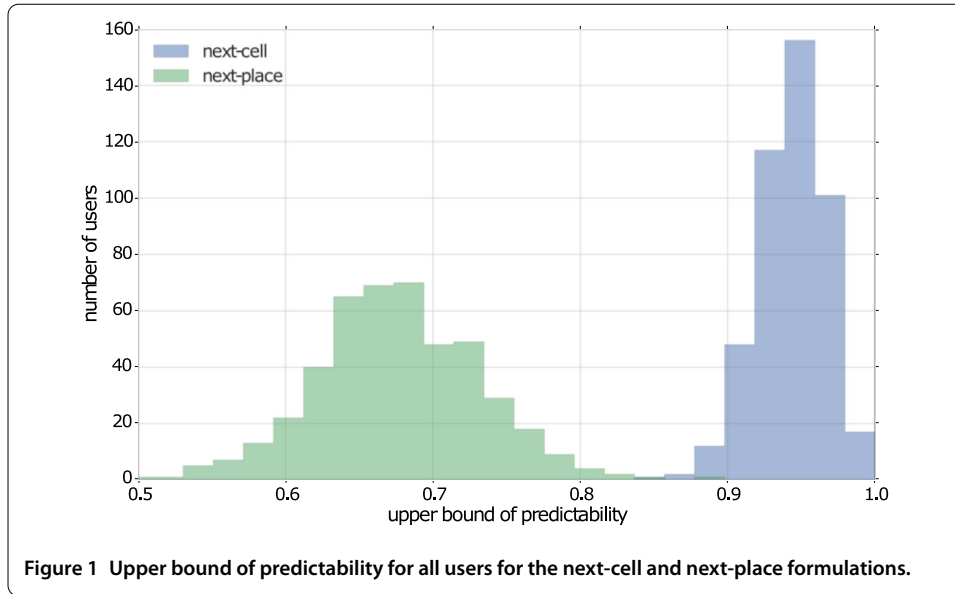
3.2 Next-cell prediction

In the first problem formulation, we convert geographical coordinates into discrete symbols by placing a uniform grid on the map and retrieving the grid cell id associated with the coordinates. We define a cell id as a tuple ($\lfloor \frac{lon}{\Delta lon} \rfloor$; $\lfloor \frac{lat}{\Delta lat} \rfloor$), where *lat* and *lon* are the geographical coordinates, and Δlat and Δlon define the cell-size. Specifically, we start by considering a grid of 50 meters × 50 meters. At each Δt = 15 minute timestep, we convert the current (*lon*, *lat*) into a cell id, producing a sequence of symbols through which we can represent a user's location history.

In this formulation, the prediction problem can be restated as follows: given your past cell sequence up to time t, which cell will you visit at time $t + \Delta t$? Before attempting any prediction, we follow the process suggested in [13] and calculate the theoretical upper limit for the predictability of the sequence of cells. Figure 1 shows how the maximum predictability for the grid cell formulation is peaked at around 0.95.

We now consider different baseline strategies for next grid cell prediction. For each of the strategies, we perform prediction in an online manner, by training the algorithm on the data up to timestep t, and predicting cell at timestep $t + \Delta t$. We measure the accuracy as number of correct predictions over the number of total predictions.

We first consider the *toploc* strategy, where at each timestep we predict the most frequent symbol in the history so far (Algorithm 1). Given the highly stationary nature of most human mobility trajectories, we expect this simple strategy to achieve relatively high accuracy. Figure 2(A) shows the distribution of accuracies for all the users. The accuracy of the *toploc* is indeed reasonable, peaking at around 0.4.

Next, we consider the Markov chain model (Algorithm 2). In this model, the prediction of next state depends only on the current state. The transition probabilities between

Figure 1 Upper bound of predictability for all users for the next-cell and next-place formulations.

Algorithm 1 Toploc prediction algorithm
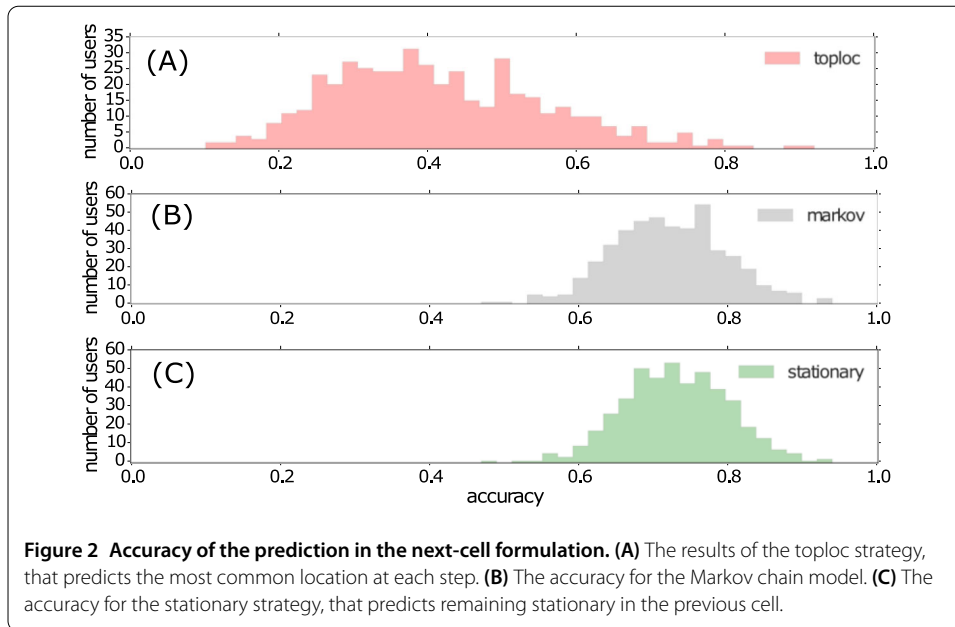
1: **function** TOPLOC(*locations*)
2:     $C \leftarrow \{\}$
3:     **for** l in *locations* **do**
4:         **if** l not in C **then**
5:             $C[l] \leftarrow 1$
6:         **else**
7:             $C[l] \leftarrow C[l] + 1$
8:         **end if**
9:     **end for**
10:     **return** $i : C[i] = \text{MAX}(C[:])$
11: **end function**

---

locations are estimated based on past transitions in the location history. For making a prediction, we then consider the transition that has the highest probability among all possible transitions from the current cell. If the current state has never been seen before, then we have no information about the transition probability to other states. In this case we fall back to the toploc strategy and predict the most frequent state. Again we fit the model in an online manner, updating at each step the transition probabilities and then making a prediction for the next timestep. Figure 2(B) shows the distribution of accuracies for all the users. The accuracy of the *Markov* model is significantly higher than toploc, peaking at around 0.7.

Considering the highly stationary nature of typical trajectories, a reasonable hypothesis is that a significant part of the Markov prediction power (in this formulation) comes from self-transitions, that is, the model predicting the user to remain in the same state as in the previous time-bin. To test this hypothesis, we consider the *stationary* strategy: at each step we predict that the user will remain in the current cell (Algorithm 3). Figure 2(C) shows that the distribution of accuracies for *stationary* closely matches the one for *Markov*. Furthermore the two are very strongly correlated (Additional file 1 Figure 3), with Pearson's

Figure 2 Accuracy of the prediction in the next-cell formulation. (A) The results of the toploc strategy, that predicts the most common location at each step. **(B)** The accuracy for the Markov chain model. **(C)** The accuracy for the stationary strategy, that predicts remaining stationary in the previous cell.

Algorithm 2 Markov prediction algorithm

1: **function** MARKOV(*locations*)
2:     $m \leftarrow |\text{UNIQUE}(locations)|$
3:     $T \leftarrow 0_{m,m}$
4:     $n \leftarrow |locations|$
5:     **for** $i = 0$ **to** $n - 2$ **do**
6:         $a \leftarrow locations[i]$
7:         $b \leftarrow locations[i + 1]$
8:         $T[a][b] \leftarrow T[a][b] + 1$
9:     **end for**
10:     $current \leftarrow locations[n - 1]$
11:     **if** $T[current][:] = 0$ **then**
12:         **return** TOPLOC(*locations*)
13:     **else**
14:         **return** $i : T[current][i] = \text{MAX}(T[current][:])$
15:     **end if**
16: **end function**

Algorithm 3 Stationary prediction algorithm

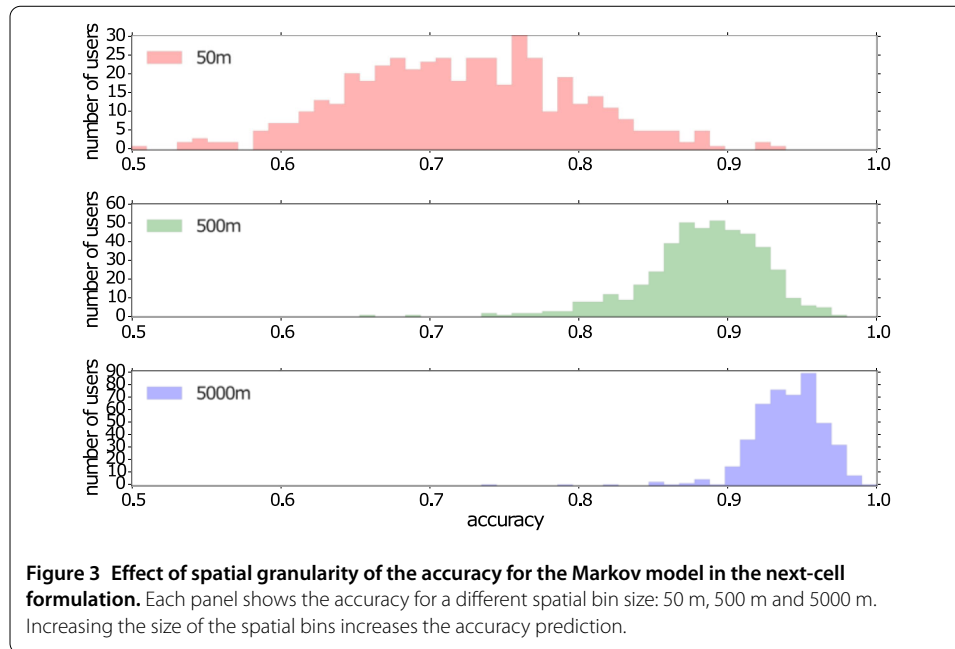1: **function** STATIONARY(*locations*)
2:     $n \leftarrow |locations|$
3:     **return** $locations[n - 1]$
4: **end function**

$r = 0.993$, $p < 0.001$. This finding confirms that, in the current formulation which predicts a user's location in the next time bin, the Markov model's predictive power is based on self-transitions.

Figure 3 Effect of spatial granularity of the accuracy for the Markov model in the next-cell formulation. Each panel shows the accuracy for a different spatial bin size: 50 m, 500 m and 5000 m. Increasing the size of the spatial bins increases the accuracy prediction.

We now investigate another issue related to the *next cell* problem formulation. Intuitively, we expect that the size of our spatial units will influence the accuracy of prediction. Predicting a user's location with the precision of few meters is intuitively much more difficult that predicting with precision of several kilometers, cf. [18]. In order to examine the effect of spatial resolution, we also consider results for cell sizes of 500 meters and 5000 meters, and apply the *Markov* model. Figure 3 compares the accuracy for different spatial resolutions. As expected the accuracy improves strongly as the spatial size increases.
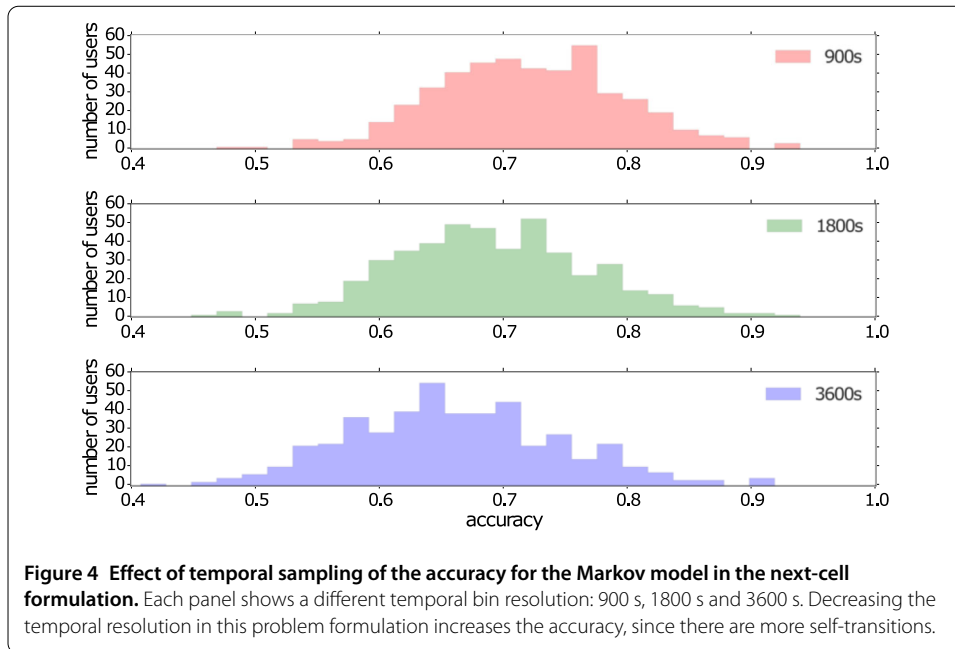
Finally we investigate the effect of temporal resolution within this problem formulation. Our findings above suggest that using a very fine-grained temporal resolution will increase the number of self-transitions, thus driving up the accuracy of the prediction that is mainly able to capture stationarity. We achieve this by discretizing the location at 50 meters cell size, but varying the temporal time binning to 15 minutes, 30 minutes and 60 minutes, and then running the *Markov* model for each scenario. Figure 4 compares the accuracy for different temporal resolutions. As expected, accuracy is decreased as the time-bins grow larger due fewer self-transitions. We also study the accuracy versus the training size, and we find that the models converge after 50-100 samples (Additional file 1 Figure 6).

3.3 Next-place prediction

We now consider an alternative problem formulation. Instead of predicting the cell in the next timestep, we want to predict only transitions between places, eliminating self-transitions. In order to do so, we convert the raw GPS locations into a sequence of stops at places. A large amount of literature has been dedicated to the problem of place detection, such as methods based on WiFi fingerprints [32], grid clustering [33], and kernel density estimation [34]. In this paper we use a density-based clustering approach such as [35–37].

When detecting places, each user is analyzed individually. We define a 'stop' as sequence of locations where the user has been approximately stationary, corresponding to a period of time when the user is visiting a place. In order to extract stops, the location samples

Figure 4 Effect of temporal sampling of the accuracy for the Markov model in the next-cell formulation. Each panel shows a different temporal bin resolution: 900 s, 1800 s and 3600 s. Decreasing the temporal resolution in this problem formulation increases the accuracy, since there are more self-transitions.

$t_1, \ldots, t_n$ are analyzed sequentially in time, and are grouped into a stop whenever they are within a distance threshold. That is, the distance between position at time t and $t + \Delta t$ is less than a threshold δ = 50 meters, roughly corresponding to the GPS accuracy. If a new location sample is farther away than δ, then a new stop is created. This produces a sequence of stops, each one with a centroid calculated as the median of the locations coordinates, and a duration equal to the time between the last location and the first location sample. In order to filter out the short stops that arise, e.g. during commutes, we consider only stops with duration greater than 15 minutes.

We now group stops into places, where a 'place' is a group of spatially proximate stops representing a self-contained area such as a building. Example of places are a 'home location', a 'work location' or a restaurant. In order to group the stops, we apply the DBSCAN [38] clustering to the stops in the geographical coordinate space, using the haversine distance. We set as parameter the grouping distance ϵ = 50 meters, and *min_pts* = 2. This distance threshold is set to produce places of the approximate size of a large building (but our results are robust to reasonable variation of this parameter). The result of the DBSCAN clustering is an assignment of a cluster label to each stop, where the label represents the place that the stop belongs to. Finally, in order avoid artifacts due to missing samples or noise, we merge multiple consecutive stops at the same place into one. The process described above converts the raw location history into a sequence of stops at places (Additional file 1 Figures 4 and 5).

The prediction task can now be re-formulated as follows: given a sequence of stops up to step n, can we predict your next place at step $n + 1$? Notice that a key difference from the cell grid formulation is that in this case there are (by definition) no self-transitions; we are interested in the place changes only.

As before, we start by investigating the upper predictability limit bound. Figure 1 shows how the maximum predictability for the places formulation is peaked at 0.68, significantly lower than what we observe in the *next-cell* formulation.

We now apply the two prediction strategies *toploc* and *Markov* to this new formulation. The two models remain conceptually the same as above, but instead of predicting the grid cell at each time step, they now predict the next place (note that in this formulation we cannot use the *stationary* strategy, since - by construction - we are interested in transitions to new places). In this case we also fit each user separately, and we perform the prediction in an online manner. Figure 5(A) and (B) show the accuracy for both models. It is evident that the accuracy for these models (around 0.3 for toploc and 0.4 for Markov) is dramatically lower than in the *next-cell* formulation, indicating that this problem formulation is a more challenging task.

Another group [39] has independently analyzed the Copenhagen Networks Study data considering similar questions related to predictability; their conclusions are similar to ours. Our work, however, significantly differs in the sense that we study additional factors that influence predictability, in particular the role of contextual features, and the exploration of new locations.
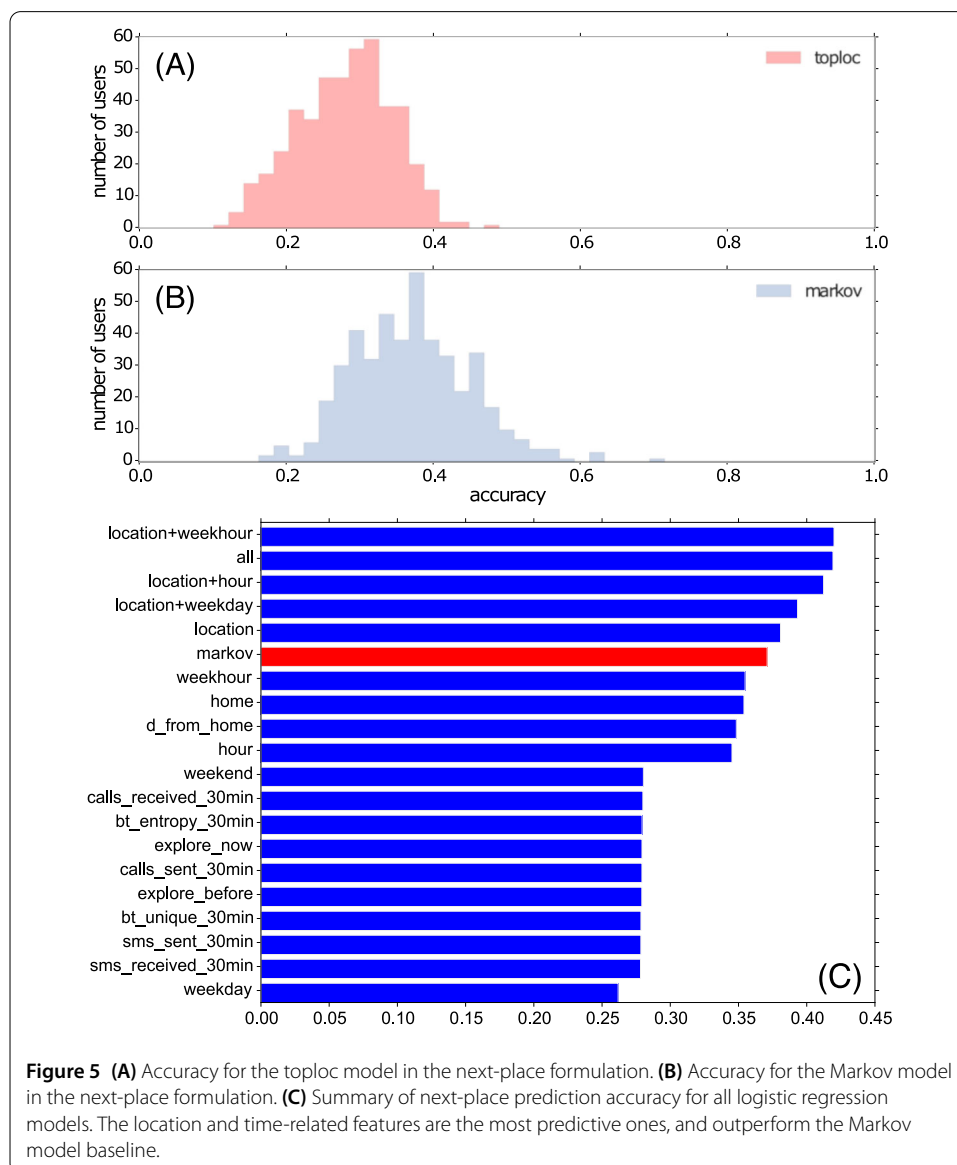
Figure 5 **(A)** Accuracy for the toploc model in the next-place formulation. **(B)** Accuracy for the Markov model in the next-place formulation. **(C)** Summary of next-place prediction accuracy for all logistic regression models. The location and time-related features are the most predictive ones, and outperform the Markov model baseline.

3.4 Importance of contextual features

We have investigated how the details of the problem formulation impact the reported accuracy for location prediction tasks. We now focus on next-place prediction and study the influence of different contextual features on performance in the prediction task. We formalize the problem as follows: At each step, we aim to compute the most probable next location given the current location, as well as other context variables, such as time of the day or call activity. In other words we wish to compute $P(\hat{L}|c_1, c_2, c_3, \ldots, c_n)$, where $\hat{L}$ is the next location, and c_1, c_2, c_n are the variables representing different contexts.

For this purpose we use a logistic regression model, and study how contextual features impact our ability to predict an individual's next place. The goal of this model is not to suggest a new state-of-the-art method, but rather to evaluate the importance of various contextual features. Specifically, we consider the current location, the time metadata (hour of the day, day of the week, hour of the week, weekend), a 'home' binary indication, distance from home, call and SMS activity, and Bluetooth proximity. Table 1 provides a summary of the features.

The logistic regression model defines a linear function of the predictor variables, and then estimates the probability of the target variable $\hat{L}$ using the logit transformation:

$$\phi(c_1, c_2, \ldots, c_n) = \beta_0 + \beta_1 c_1 + \beta_2 c_2 + \cdots + \beta_n c_n,$$

$$P(\hat{L}|c_1, c_2, \ldots, c_n) = \frac{1}{1 + e^{-\phi}}.$$

As there is no closed form solution for setting the weights, the fit is done by numerical optimization with objective of maximizing classification accuracy.

We model each user separately since our goal is to understand next-place prediction at the individual level. We perform online predictions where we fit the data up to step n, and then predict the next location at step $n + 1$ using a multinomial logistic regression model; classes are simply the set of all possible places visited so far. We fit one model for each individual feature separately, then models combining location and time metadata (hour of the day, hour of the week and day of the week) and then a model with all the features. Models are fit using L2 regularization in order to reduce the number of features. Figure 5(C) displays the accuracy for each of the models, averaged by user.

Table 1 Description of the features used for the logistic regression models

Feature name	Description
location	location ID
hour	hour of the day (0-23)
weekhour	hour of the week (0-167)
weekday	day of the week (0-6)
weekend	sat/sun (1) or Mon-Fri (0)
explore_before	1 if the previous stop is an exploration, 0 otherwise
explore_now	1 if the current stop is an exploration, 0 otherwise
home	1 if the current stop is at the home location (most visited place), 0 otherwise
d_from_home	distance from the current stop to the home location
sms_received_30min	number of SMS received in the 30 min before the current stop timestamp
sms_sent_30min	number of SMS sent in the 30 min before the current stop timestamp
calls_received_30min	number of phone calls received in the 30 min before the current stop timestamp
calls_sent_30min	number of phone calls sent in the 30 min before the current stop timestamp
bt_entropy_30min	entropy of Bluetooth devices scanned in the 30 min before the current stop timestamp
bt_unique_30min	number of unique Bluetooth devices scanned in the 30 min before the current stop timestamp

The Markov chain model baseline is highlighted in red. Using the current location and time features, the logistic regression model outperforms the Markov chain based model. Even using the current location only (which is conceptually very similar to a Markov chain model), the logistic regression shows stronger performance, likely due to the explicit optimization of the model. It is interesting that other context variable such as call and SMS data have little predictive power in this model formulation. The most complex model that considers all features is practically identical in performance to the model using only current location and hour of the week.

The logistic regression model does improve the accuracy over the Markov model, but the absolute value of accuracy still remains low (below 45%). Although the current model could be further fine-tuned, our goal here is not to produce a state-of-the-art prediction model, but instead to investigate the limitations and sources of predictability in human mobility. We also verify the model convergence by studying the accuracy versus the training size, and we find that also in this case the accuracy converges after 50-100 samples (Additional file 1 Figure 6). The logistic regression model shows that (in the current problem formulation) the contextual variables add very little additional predictive power. The next sections will therefore be focused on understanding the factors limiting the prediction accuracy.
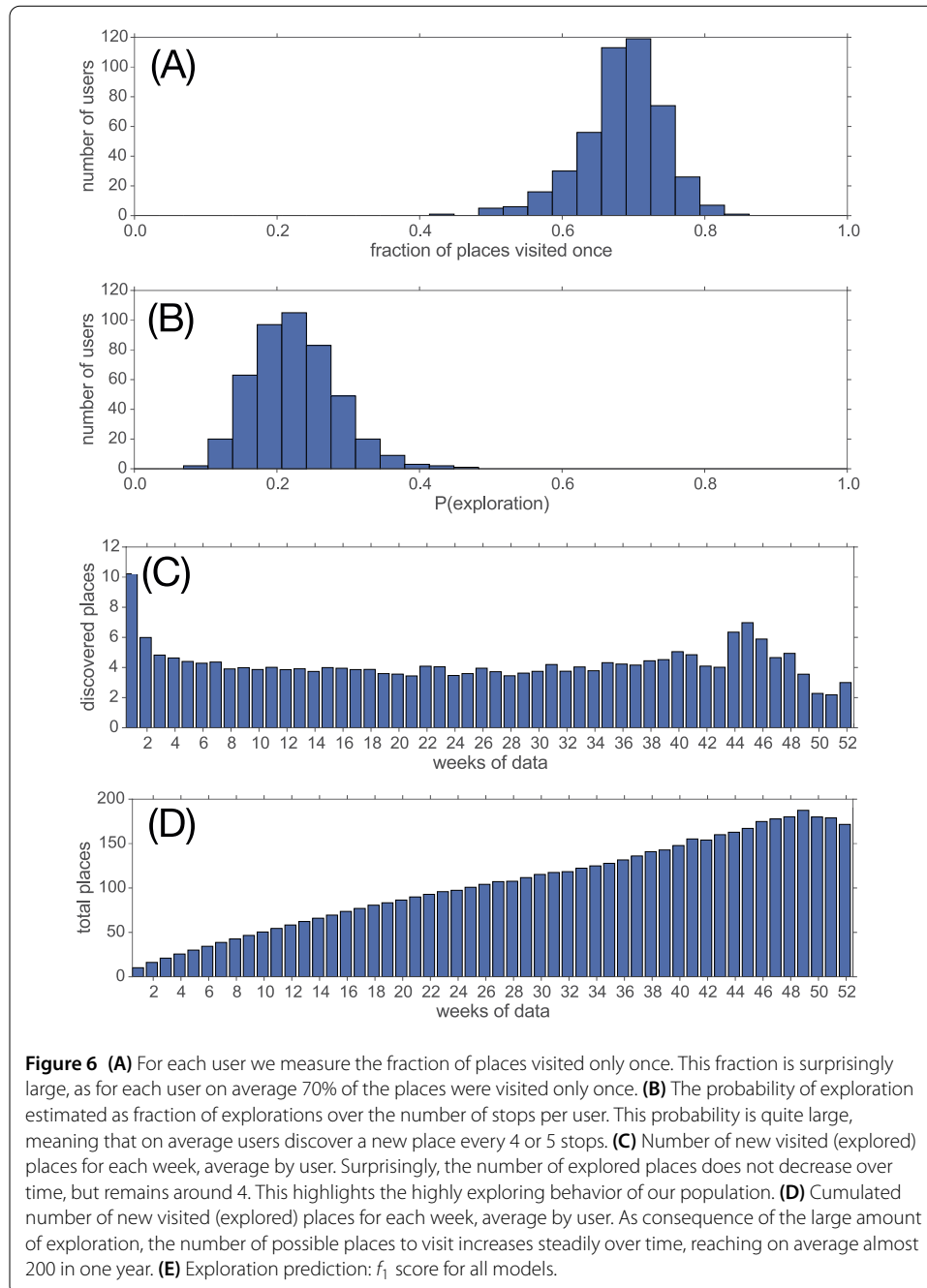
3.5 Understanding the set of location states

It is well known that the majority of individuals tend to spend most of the times at very few places such as home and work, and only sporadically visit other places. This phenomenon has been described using concepts such as preferential return [40], heavy-tailed stay times and return rate based on the number of visits [41]. For the location prediction tasks, the consequence is that the target classes are unbalanced, which implies that most records belong to very few classes and most classes are represented by only few records. To illustrate this problem, we consider the extreme case of places visited only once. Figure 6(A) shows that surprisingly this fraction is quite large (0.7). This fact is, in large part, the central reason behind the difficulty of the prediction task.

As we shall see below, another central challenge is not just that our population visits a large number of different places, but also that many new places are discovered over time. We consider a stop at a place as an 'exploration' if this place has not been seen in the location history so far for a given user. In other words, this place is being visited for the first time by the user. To express this formally, we consider the sequence of stops $s_1, s_2, s_3, \ldots, s_n$ for each user. We consider a stop s_i as *return* ($Y = 0$) if s_i has been seen before in the location history, that is there exists a stop $s_j = s_i$ for $1 \leq j < i$. Otherwise we consider stop s_i an *exploration* ($Y = 1$), that is the place s_i is visited for the first time at step i. For example given a location sequence A B A C B C, the target variable exploration would have values 1 1 0 1 0 0.

We can now estimate the probability of exploration as fraction of explorations over the number of stops. To our surprise, this probability is large: between 0.2 and 0.25 (Figure 6(B)). This implies that most users discover a new place every 4 or 5 stops.

The fact that a large fraction of places have never been seen before poses a challenge for the prediction task, since - by construction - any model that attempts to predict a place from an alphabet of previously visited places will be unable to predict new, unseen symbols. Moreover, another consequence of this exploration is that the pool of possible places

Figure 6 (A) For each user we measure the fraction of places visited only once. This fraction is surprisingly large, as for each user on average 70% of the places were visited only once. **(B)** The probability of exploration estimated as fraction of explorations over the number of stops per user. This probability is quite large, meaning that on average users discover a new place every 4 or 5 stops. **(C)** Number of new visited (explored) places for each week, average by user. Surprisingly, the number of explored places does not decrease over time, but remains around 4. This highlights the highly exploring behavior of our population. **(D)** Cumulated number of new visited (explored) places for each week, average by user. As consequence of the large amount of exploration, the number of possible places to visit increases steadily over time, reaching on average almost 200 in one year. **(E)** Exploration prediction: f_1 score for all models.

constantly grows over time and, given the longitudinal nature of our dataset, ends up being very large. Figure 6(C) and (D) show how the average number of new place explored per week remains approximately constant around 4, and consequently the total number of places keeps growing to hundreds of places. This is a problem for the prediction task, as the number of possible places that the classifier needs to choose from increases constantly.

In fact, if we measure the relation between the number of unique places per user and the performance of the best performing logistic regression model using the Pearson correlation coefficient, we find a negative correlation ($r = -0.478$, $p < 0.001$). Conversely, we find no significant correlation for accuracy with period length or number of stops.

Figure 6 Continued
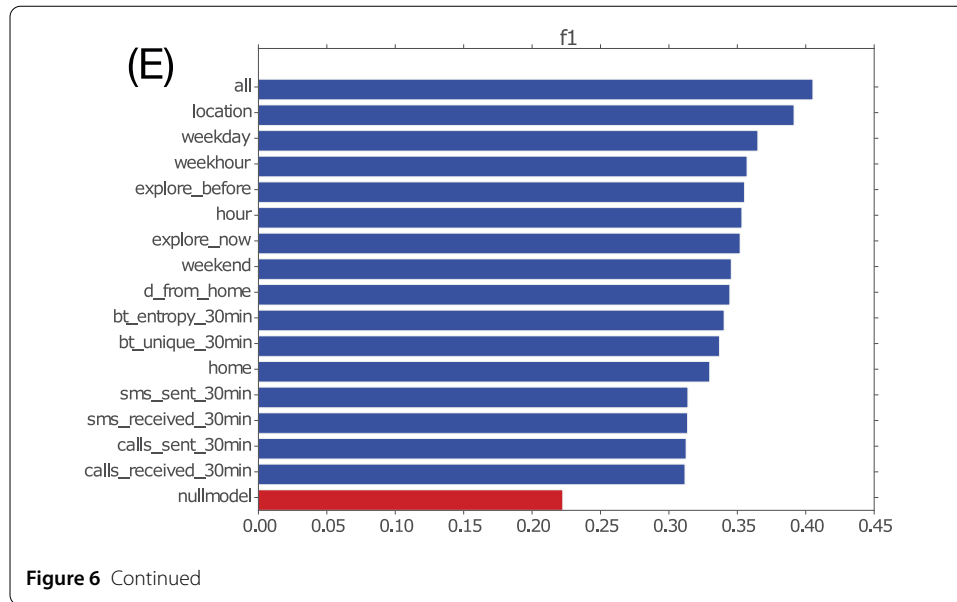
These facts suggest that the exploration phenomenon is a key reason for the relatively low accuracy of mobility prediction tasks at high spatial resolution. Given the importance of exploration, we now introduce a new task in mobility prediction: *exploration prediction*.

3.6 Exploration prediction

The exploration prediction task can be phrased as follows: given a user's location history up to step n, will the stop at step $n + 1$ be an exploration or a return?

The first question is: how do we express a useful baseline model for the exploration prediction task? Surprisingly, most literature on human mobility prediction has focused on next location prediction, but has overlooked the exploration prediction problem. Thus, to the best of our knowledge no suitable solution has been proposed for this task. We therefore suggest, as a reasonable baseline, random guessing with probability equal to our prior knowledge of the fraction of explorations: $P(exploration) \approx 0.2$.

For our main model we use - as before - the logistic regression model with the features used for the next place prediction model. We add two additional features: *explore_now* and *explore_before*, which capture if the current stop or the previous stops were explorations, respectively. The intuition for these is that multiple explorations may occur in a row, and therefore the current exploration may increase the likelihood for an exploration at the next stop.

As before, we fit each individual separately, and we perform an online prediction, that fits based on the data up to step n, and predicts exploration at step $n + 1$. We fit one logistic regression model for each of the single features, and a more complex model with all the features at once.

Measuring the performance of these models requires a few considerations. In this case, the classification problem is imbalanced, that is the number of positive cases (exploration) is much smaller than negative cases (return). This implies that accuracy is therefore not a good metric, since a classifier predicting always the most probable class (return) would have good performance, but would not be useful. Instead we use the f_1 score, which is the harmonic mean of precision (the fraction of correctly predicted explorations over all

predicted explorations) and recall (the fraction of correctly predicted explorations over all true explorations). Figure 6(E) shows the results of the exploration prediction.

As expected, the model with the most complete set of features outperforms the others. Among the single feature models, perhaps not surprisingly, the current location feature has the best performance. This finding can be explained by the role of some places as 'gateways' for exploration such as public transport hubs (e.g. a central train station). The individual features that perform also well are the time-related ones, in agreement with the intuition that exploration tends to happen to at specific times of the day or week. The *explore_now* and *explore_before* also perform well, suggesting an element of burstiness in the exploratory behavior. If we consider our best performing model, we find that it has average precision of 0.3 and recall of 0.65. Overall the performance of this model is far from perfect, showing that the exploration prediction problem is a challenging one.

4 Discussion

We first showed that when interpreting reported results of predictive performance, there are a number of factors to take into consideration. Central among these is the problem formulation itself. This is simply because predicting the next time-bin is a very different (and much easier) task than predicting the next transition. We showed that the most challenging problem is the next-place prediction, which is arguably the most useful task for practical applications such as travel recommendations. Another issue to be taken into account is the spatial resolution of the prediction, here we show how more coarse spatial precision results in an easier task. Similarly the time resolution also has an effect on the predictive power. We suggest that the factors described in this paper should be taken in consideration as context when comparing results from prediction models.

Other than the factors discussed above, we believe that one further reason for performance differences could be the demographics of the dataset. The data set under study arises from students that have no single workplace but tend to change multiple classes per week, even multiple times per day. Moreover a younger population may have a more irregular schedule and more exploratory behavior. Certainly more work is needed to understand the connection between demographics and predictability. For future research, we suggest considering demographic factors when trying to characterize human mobility, as it has been done, for example, by linking changes in mobility patterns with unemployment status [42].

We also discussed the issue of exploration, and we showed that new places are frequently discovered (typically one in every four or five stops is a new location). Based on this observation, we argued that the mechanism of exploration is an important part of human mobility and plays a role in next-place prediction. Any model that attempts to predict a next place from a set of visited place will fail whenever an exploration occurs. This problem has rarely been addressed in the mobility prediction literature, which almost always assumes that the next place can be determined from the past history. Providing a full solution for next explored place prediction is beyond the scope of this work, and here we simply aim to stress the fact that the prediction of explorations is very different from the predictions to returns to known places. Some previous work on next-place prediction using social information [29, 30] or nearby Points Of Interest [43] may be the starting point for investigating this problem.

In this sense, we raise the question on whether the simple location history is enough for successful next-place prediction. As we have discussed, there are strong regularities both in the sequence of visits, and in the daily and weekly temporal patterns of visitation. However, there are a lot of 'exceptions to the rules', where schedules change, plans are canceled, and people run late. We speculate that other channels such as email, social media, calendar, class schedule may be needed for achieving a satisfying accuracy in the prediction task.

Additional material

Additional file 1: Supplementary Information. (pdf)

Author details
[1]DTU Compute, Technical University of Denmark, Richard Petersens Plads Building 324, Kgs. Lyngby, Denmark. [2]The Niels Bohr Institute, University of Copenhagen, Blegdamsvej 17, Copenhagen, 2100, Denmark. [3]Department of Civil and Environmental Engineering and Engineering Systems, Massachusetts Institute of Technology, 77 Massachusetts Avenue, 1-290, Cambridge, 02139, United States. [4]Department of City and Regional Planning, UC Berkeley, 406 Wurster Hall, Berkeley, 94720, United States. [5]Lawrence Berkeley National Laboratory, Cyclotron Road, Berkeley, 94720, United States.

Publisher's Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References
1. Lane ND, Mohammod M, Lin M, Yang X, Lu H, Ali S, Doryab A, Berke E, Choudhury T, Campbell A (2011) Bewell: a smartphone application to monitor, model and promote wellbeing. In: 5th international ICST conference on pervasive computing technologies for healthcare, pp 23-26
2. Quercia D, Lathia N, Calabrese F, Di Lorenzo G, Crowcroft J (2010) Recommending social events from mobile phone location data. In: Data mining (ICDM), 2010 IEEE 10th international conference on. IEEE Comput. Soc., Los Alamitos, pp 971-976
3. Aalto L, Göthlin N, Korhonen J, Ojala T (2004) Bluetooth and wap push based location-aware mobile advertising system. In: Proceedings of the 2nd international conference on mobile systems, applications, and services. ACM, New York, pp 49-58

4. Lu X, Bengtsson L, Holme P (2012) Predictability of population displacement after the 2010 Haiti earthquake. Proc Natl Acad Sci 109(29):11576-11581
5. Çolak S, Lima A, González MC (2016) Understanding congested travel in urban areas. Nat Commun 7
6. Eagle N, Pentland A (2006) Reality mining: sensing complex social systems. Pers Ubiquitous Comput 10(4):255-268
7. Liao L, Patterson DJ, Fox D, Kautz H (2007) Learning and inferring transportation routines. Artif Intell 171(5):311-331
8. Zheng Y, Zhou X (2011) Computing with spatial trajectories. Springer, Berlin
9. Arentze T, Timmermans HA (2000) A learning based transportation oriented simulation system. Eirass Eindhoven
10. Balmer M, Meister K, Rieser M, Nagel K, Axhausen KW, Axhausen KW, Axhausen KW (2008) Agent-based Simulation of Travel Demand: Structure and Computational Performance of MATSim-t, ETH, Eidgenössische Technische Hochschule Zürich, IVT Institut für Verkehrsplanung und Transportsysteme
11. Goodchild MF (2007) Citizens as sensors: the world of volunteered geography. GeoJournal 69(4):211-221
12. Batty M (2013) The new science of cities. MIT Press, Cambridge
13. Song C, Qu Z, Blumm N, Barabási A-L (2010) Limits of predictability in human mobility. Science 327(5968):1018-1021
14. Simini F, González MC, Maritan A, Barabási A-L (2012) A universal model for mobility and migration patterns. Nature 484(7392):96-100
15. Schneider CM, Rudloff C, Bauer D, González MC (2013) Daily travel behavior: lessons from a week-long survey for the extraction of human mobility motifs related information. In: Proceedings of the 2nd ACM SIGKDD international workshop on urban computing. ACM, New York, p 3
16. Lin M, Hsu W-J, Lee ZQ (2012) Predictability of individuals' mobility with high-resolution positioning data. In: Proceedings of the 2012 ACM conference on ubiquitous computing. ACM, New York, pp 381-390
17. Smith G, Wieser R, Goulding J, Barrack D (2014) A refined limit on the predictability of human mobility. In: Pervasive computing and communications (PerCom), 2014 IEEE international conference on. IEEE, Hungary, pp 88-94
18. Lu X, Wetter E, Bharti N, Tatem AJ, Bengtsson L (2013) Approaching the limit of predictability in human mobility. Scientific Reports 3
19. Song L, Kotz D, Jain R, He X (2006) Evaluating next-cell predictors with extensive wi-fi mobility data. IEEE Trans Mob Comput 5(12):1633-1649
20. Krumme C, Llorente A, Cebrian M, Pentland A (2013) The predictability of consumer visitation patterns. Sci Rep 3:1645
21. Bapierre H, Groh G, Theiner S (2011) A variable order markov model approach for mobility prediction. Pervasive Computing, 8-16
22. Zheng Y, Xie X, Ma W-Y (2010) Geolife: A collaborative social networking service among user, location and trajectory. IEEE Data Eng Bull 33(2):32-39
23. Gao H, Tang J, Liu H (2012) Mobile location prediction in spatio-temporal context. In: Nokia mobile data challenge workshop, vol 41. p 44
24. Laurila JK, Gatica-Perez D, Aad I, Bornet O, Do T-M-T, Dousse O, Eberle J, Miettinen M et al (2012) The mobile data challenge: big data for mobile computing research. In: Pervasive computing
25. Do TMT, Gatica-Perez D (2012) Contextual conditional models for smartphone-based human mobility prediction. In: Proceedings of the 2012 ACM conference on ubiquitous computing. ACM, New York, pp 163-172
26. Do TMT, Dousse O, Miettinen M, Gatica-Perez D (2015) A probabilistic kernel method for human mobility prediction with smartphones. Pervasive Mob Comput 20:13-28
27. Scellato S, Musolesi M, Mascolo C, Latora V, Campbell AT (2011) Nextplace: A spatio-temporal prediction framework for pervasive systems. In: Pervasive computing. Springer, Berlin, pp 152-169
28. Sadilek A, Krumm J (2012) Far out: predicting long-term human mobility. In: Proceeding AAAI'12 proceedings of the twenty-sixth AAAI conference on artificial intelligence
29. Cho E, Myers SA, 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. ACM, New York, pp 1082-1090
30. Sadilek A, Kautz H, Bigham JP (2012) Finding your friends and following them to where you are. In: Proceedings of the fifth ACM international conference on Web search and Data mining. ACM, New York, pp 723-732
31. Stopczynski A, Sekara V, Sapiezynski P, Cuttone A, Madsen MM, Larsen JE, Lehmann S (2014) Measuring large-scale social networks with high resolution. PLoS ONE 9(4):95978
32. Kang JH, Welbourne W, Stewart B, Borriello G (2004) Extracting places from traces of locations. In: Proceedings of the 2nd ACM international workshop on wireless mobile applications and services on WLAN hotspots. ACM, New York, pp 110-118
33. Zheng VW, Zheng Y, Xie X, Yang Q (2010) Collaborative location and activity recommendations with gps history data. In: Proceedings of the 19th international conference on World Wide Web. ACM, New York, pp 1029-1038
34. Thierry B, Chaix B, Kestens Y (2013) Detecting activity locations from raw gps data: a novel kernel-based algorithm. Int J Health Geogr 12(1):1
35. Zhou C, Frankowski D, Ludford P, Shekhar S, Terveen L (2007) Discovering personally meaningful places: an interactive clustering approach. ACM Trans Inf Syst 25(3):12
36. Zheng Y, Zhang L, Xie X, Ma W-Y (2009) Mining interesting locations and travel sequences from gps trajectories. In: Proceedings of the 18th international conference on World Wide Web. ACM, New York, pp 791-800
37. Montoliu R, Gatica-Perez D (2010) Discovering human places of interest from multimodal mobile phone data. In: Proceedings of the 9th international conference on mobile and ubiquitous multimedia. ACM, New York, p 12
38. Ester M, Kriegel H-P, Sander J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. AAAI Press, Menlo Park, pp 226-231
39. Ikanovic EL, Mollgaard A (2017) An alternative approach to the limits of predictability in human mobility. EPJ Data Sci 6(1):12
40. Song C, Koren T, Wang P, Barabási A-L (2010) Modelling the scaling properties of human mobility. Nat Phys 6(10):818-823
41. Gonzalez MC, Hidalgo CA, Barabasi A-L (2008) Understanding individual human mobility patterns. Nature 453(7196):779-782

42. Toole JL, Lin Y-R, Muehlegger E, Shoag D, González MC, Lazer D (2015) Tracking employment shocks using mobile phone data. J R Soc Interface 12(107):20150185
43. Yu C, Liu Y, Yao D, Yang LT, Jin H, Chen H, Ding Q (2015) Modeling user activity patterns for next-place prediction. IEEE Syst J PP(99):1-12