Revealing challenges in human mobility predictability

Douglas Do Couto Teixeira, Aline Carneiro Viana, Jussara Marques Almeida, Mário S. Alvim

REVEALING CHALLENGES IN HUMAN MOBILITY PREDICTABILITY

A PREPRINT

Douglas do Couto Teixeira
Department of Computer Science
Federal University of Minas Gerais
Belo Horizonte, Brazil
douglas@dcc.ufmg.br

Aline Carneiro Viana
Inria, Université Paris-Saclay
Palaiseau, France
aline.viana@inria.fr

Jussara M. Almeida
Department of Computer Science
Federal University of Minas Gerais
Belo Horizonte, Brazil
jussara@dcc.ufmg.br

Mário S. Alvim
Department of Computer Science
Federal University of Minas Gerais
Belo Horizonte, Brazil
msalvim@dcc.ufmg.br

February 2, 2021

ABSTRACT

Predicting mobility-related behavior is an important yet challenging task. On one hand, factors such as one's routine or preferences for a few favorite locations may help in predicting their mobility. On the other hand, several contextual factors, such as variations in individual preferences, weather, traffic, or even a person's social contacts, can affect mobility patterns and make its modeling significantly more challenging. A fundamental approach to study mobility-related behavior is to assess how *predictable* such behavior is, deriving theoretical limits on the accuracy that a prediction model can achieve given a specific dataset. This approach focuses on the inherent nature and fundamental patterns of human behavior captured in that dataset, filtering out factors that depend on the specificities of the prediction method adopted. However, the current state-of-the-art method to estimate predictability in human mobility suffers from two major limitations: low interpretability, and hardness to incorporate external factors which are known to help mobility prediction (i.e., contextual information). In this article, we revisit this state-of-the-art method, aiming at tackling these limitations. Specifically, we conduct a thorough analysis of how this widely used method works by looking into two different metrics that are easier to understand and, at the same time, capture reasonably well the effects of the original technique. We evaluate these metrics in the context of two different mobility prediction tasks, notably next-cell and next-place prediction, which have different degrees of difficulty. Additionally, we propose alternative strategies to incorporate different types of contextual information into the existing technique. Our evaluation of these strategies offer quantitative measures of the impact of adding context to the predictability estimate, revealing the challenges associated with doing so in practical scenarios.

***K*eywords** human mobility · predictability · entropy estimators · contextual information

1 Introduction

Predicting human behavior is hard, given the great degree of uncertainty and heterogeneity, as well as the variability of parameters, influencing such behavior. At a more fundamental level, one may desire to estimate how *predictable* a certain type of behavior can be, i.e., to assess its *predictability*, which can be defined as the maximum theoretical accuracy that any prediction model could achieve given a particular dataset capturing the behavior under consideration.

The focus of *predictability analysis* is to detect patterns as they appear in the data. More precisely, this approach abstracts away from any specific prediction strategy and concentrates instead on the inherit nature and fundamental patterns of human behavior as captured by the available data. Because this approach is neither tied to particular prediction techniques nor relies on the fine tuning of a multitude of sensible parameters, it can be used as a common ground to study certain types of behavior.

Predictability also has practical applications. First, it can be used as a tool to assess the confidence in predictions. This is useful, for instance, in the analysis of highly unpredictable individuals, particularly when a misprediction has a high cost. It can also be employed in outlier identification: in a particular dataset, users who exhibit levels of predictability very different from the rest are likely to be outliers, and therefore may deserve special attention. Additionally, predictability can be used as a measure of *openness to new things*. For instance, in a recommendation scenario, users with low predictability are possibly those who are more open to novelty, diversity and serendipity.

In this article, we focus on the predictability of a particular type of individual behavior, namely *human mobility*. Besides the already mentioned applications, studying predictability in this domain can, for instance, help uncover relevant insights for driving the design of more cost-effective traffic management policies, as well as of content distribution and recommendation techniques [1, 2].

The state-of-the-art technique to measure predictability in human mobility was proposed by Song et et al. al. [3] and has been used by several research communities [1, 2, 4, 5, 6]. It exploits the concept of entropy as a measure of how complex (or, inversely, how predictable) a person's mobility patterns are. In a nutshell, it estimates the entropy of the person's mobility trace (again, as expressed in the data), and subsequently uses this value to obtain a predictability estimate in the range $[0, 1]$, with 0 meaning completely unpredictable, and 1 meaning totally predictable. In this way, the crux of predictability is an entropy estimate.

Despite its wide use and applicability, Song et al.'s technique has two major shortcomings. First, *it has low interpretability*, which can be attributed to its inner workings: it is based on a sophisticated compression algorithm, which makes it hard for people to understand what affects the predictability of an individual's mobility trajectories. Second, Song et al.'s technique uses only the person's history of visited locations to compute their predictability, but *it does not capture other aspects that may influence human mobility*, such as contextual information (e.g., weather, time of the day, etc). Indeed prior work [7, 8] has already argued for the impact of context on predictability estimate, without, however, neither quantifying such impact nor discussing how exactly this kind of information can be incorporated to improve predictability.

In this article, *we tackle these limitations by aiming at understanding what affects predictability (i.e., entropy) estimates in human mobility and assessing the impact of introducing contextual information to such estimates*[1]. Towards that goal, we first investigate easier-to-interpret proxy metrics that can help explain the entropy value associated with a mobility trace. Previous work [9, 8] showed that predictability of human mobility, as captured by Song et al.'s technique, is strongly related to stationary patterns. Yet, as we show in this article, stationarity alone does not explain predictability. We propose another metric, called *regularity* which, together with stationarity, helps us better understand the predictability of an individual's mobility. Next, we show how to incorporate contextual information into the limits of predictability, as well as how to quantify its impact on predictability. We do so by using different entropy estimators which, compared to Song et al.'s technique, are simpler and more easily allow for the use of contextual information. We also propose two strategies to encode contextual information directly into Song et al.'s entropy estimator.

Our investigation of predictability in human mobility relies on the analysis of two datasets of different spatial and temporal granularities, as these properties may influence predictability. It also considers two prediction tasks: namely *next-cell* and *next-place prediction* [8]. In both tasks, the spatial area is divided into cells and time is discretized into bins of a given duration. In the next-cell prediction problem the goal is to correctly guess the cell identifier (location) of a person in the next time bin (which could be the same cell the person is currently in), given the person's previously visited cells [3, 8]. In the next-place prediction problem, the goal is to guess the next *distinct* place (cell) the person will visit [8]. Notice that the latter is concerned only with transitions between different places, which eliminates stationarity (i.e., transitions to the same, current location), making the prediction considerably harder.

In a preliminary version of this work [10], we first studied the relationship between regularity and stationarity, with respect to Song et al.'s entropy estimate, for the next-cell prediction task only. We then evaluated the impact of contextual information into alternative entropy estimates, comparing them against Song et al.'s original approach. We here build on this prior effort by presenting a much more comprehensive investigation and offering:

[1]For the sake of readability, unless otherwise noted, throughout the rest of this paper we use the term predictability to refer to the theoretical limit on the predictability of one's mobility patterns proposed by Song et al.

- *A more detailed evaluation of the relationship between regularity, stationarity and entropy estimates for both next-cell and next-place prediction.* We show that these two simple metrics are complementary and, when used jointly, are able to explain most of the variation in Song et al.'s predictability. As such, we here use them as proxies of that technique to analyze how the predictability of one's mobility varies.
- *A novel strategy that seamlessly incorporates contextual information into Song et al.'s predictability technique.*
- *A broad evaluation and discussion of the benefits of contextual information on several entropy estimators, for both next-cell and next-place prediction.* Our results show that introducing context information can indeed improve predictability estimate, when simpler, alternative entropy estimators are employed. However, the benefits of introducing context into the Song et al.'s entropy estimator strongly depends on other factors such as the size, regularity and stationarity of the input sequence. Indeed, we found that in several cases, the original estimator, without context, remains the best approach, producing lower entropy values (higher predictability). These results hint at the observation that using context may not always bring benefits to predictability estimate in practical acenarios.

The rest of this article is organized as follows. We start by discussing related work in Section 2. In Section 3, we revisit the state-of-the-art method proposed by Song et al., offering a thorough analysis and insights on how it works and its robustness in different scenarios, and formally defining our target problem. In Section 4 we present our datasets and the prediction tasks under study. In Section 5 we discuss regularity and stationarity, and show how they can be used to understand and interpret predictability values both in the next-cell and next-place prediction tasks. We then examine, in Section 6, the benefits and the challenges of introducing contextual information in predictability estimation, considering different estimates of predictability. Finally, in Section 7 we summarize our main results and conclusions, and discuss future work.

2 Related Work

The study of human mobility has received considerable attention in the literature [6, 11, 12, 13, 14, 15, 16, 7, 9, 6]. Many previous studies have proposed prediction strategies by employing a plethora of different techniques (e.g., Markov chains [6], logistic regression [8], neural networks [17], and so on) and relied on several types of data sources (Call detail records (CDRs) [6, 3], GPS traces [18, 8], and social media data [19, 20], among others). These studies often evaluate the proposed techniques on specific datasets, comparing them with alternative baseline methods. A different body of work has focused on analyzing the *predictability* of human mobility, as captured by different datasets [7, 9, 6, 21]. As such, these studies focus on the information contained in the data only. By decoupling the analysis from the intricacies of a given prediction technique, these studies offer upper limits on the accuracy of a prediction technique on the given dataset, based solely on the inherent nature of human mobility behavior expressed in it. This is the approach taken by Song et al. in their seminal paper [3], and it is the focus of our present study.

Song et al. [3], proposed a technique, which will be explained in detail in Section 3, that leverages the concept of entropy to estimate the predictability of a sequence of locations visited by a given person. This technique has been extensively used to assess predictability in scenarios such as human mobility [6, 8, 7, 21], taxi demand prediction [1], cellular network traffic [2], radio spectrum state dynamics [4], among others. Previous work also evaluated the robustness of this technique with respect to its assumptions [22] and to its mathematical details [23]. Other studies computed the limits of predictability for several temporal and spatial resolutions [7], but they did not try to explain what (besides temporal and spatial granularity) may impact predictability.

Orthogonally, prior work has computed the limits of predictability for the two prediction tasks that we study here. For example some authors [6, 1, 7, 8] applied the technique proposed by Song et al. to study predictability of human mobility in the context of next-cell prediction, reporting high values for the maximum predictability. Others analyzed predictability for the next-place prediction task [8, 9]. In particular, Cuttone et al. [8] pointed out that the removal of stationary patterns, which occurs when one focuses on next-place prediction, corresponds to lower predictability.

Despite such great attention, the method proposed by Song et al. has low interpretability, as it exploits a sophisticated entropy estimator based on compression and pattern-matching, whose output bears little resemblance to its input. Thus, it is hard to keep track of what the algorithm is really capturing in terms of mobility patterns. Moreover, no prior work has tried to quantify the extent to which context affects predictability in both next-cell and next-place prediction. Indeed, we are aware of only one recent attempt to add context information to estimator used by Song et al., though in a different domain [5], where the authors estimated how much knowing the contents of the tweets of a person's friends helps in predicting the contents of this person's tweets. As we will argue in Section 6, it is quite challenging to extend the Song et al.'s method to incorporate contextual information (e.g., weather, time of the day), which is a desirable feature to improve predictability estimates, as previously suggested [7, 8].

In a preliminary version of this work [10], we investigated the relationship of both stationarity and regularity, two formally defined metrics, with the predictability (i.e., entropy) estimate adopted by Song et al.. We showed that, in the particular case of next-cell prediction, these two metrics, which are much easier to compute and interpret, are able explain reasonably well the entropy estimate. In that work, we also argued that using contextual information with the same entropy estimator used by Song et al. is quite a challenge, and showed that other types of entropy estimators can be used to incorporate context into predictability estimates [10].

Our present work is complementary to and greatly contributes over previous studies (including ours [10]) in several aspects. First, we build on our initial effort [10] by analyzing regularity and stationarity in both next-cell and next-place prediction tasks. Our experimental results show that these two metrics, jointly, can be used to interpret predictability estimates reasonably well in both scenarios. Second, we propose and evaluate a new technique to incorporate context directly into the entropy estimate used by Song et al. . Our experimental results reveal, in a much more fundamental perspective, the challenges associated with introducing context to the method, and offers a much deeper and more thorough discussion of the potential benefits of using contextual information in predictability estimates in practice.

3 Theoretical Foundations of Predictability

In this section we present the theoretical foundations of how to estimate predictability in human mobility. In particular, we revisit Song et al.'s technique in the light of more fundamental theoretical concepts and well-established measures of complexity. To the best of our knowledge, no previous work has summarized the roots of predictability – tracing the equivalences between entropy and compressibility and showing why entropy is a good approximation for the complexity of a sequence of symbols – as we do here.

3.1 Estimating Complexity

The predictability of a sequence of symbols is intimately related to its complexity (randomness): in general, the more complex a sequence is, the harder it is to predict it. Analogously, a sequence's complexity is also related its *compressibility* [24]: more random sequences are less compressible, whereas more predictable sequences are more compressible. The rationale behind compressibility and predictability is that if we are able to compress a sequence, then there exists a decompressor that uses the same algorithm and is able to reconstruct (predict) the original sequence.

Kolmogorov Complexity [25], a general measure of the complexity of a sequence, is also related to compressibility. It is defined as the size of the smallest program, which runs on a universal computer such as a Turing Machine, that generates the sequence and terminates. Notice that the use of different programming languages will result in different values of the Kolmogorov Complexity, but it has been shown [26] that these differences are bound by a fixed, additive constant (hence, the choice of any particular programming language is irrelevant when dealing with large enough sequences).

Intuitively, the Kolmogorov Complexity of a sequence X, denoted by $K(X)$, is the minimum amount of information required by a program to produce X. Consider, for instance, the sequence of digits of the constant π that, although infinite, can be generated by a program of only a few lines. Since π has an infinite number of digits yet can be generated (up to any desired precision) by a small program, the sequence of digits of π is highly compressible (and has small Kolmogorov Complexity). The program that generates the sequence can be seen as the compressed version of π.

If a sequence is completely predictable (deterministic), there exists a program that generates it and terminates, such as in the case of any finite sub-sequence of the decimal digits of π. If, however, the input sequence is produced by a non-deterministic process, there may not be a program that generates it exactly. For instance, a program that is able to generate the sequence of locations that a given person visits would have to take into account all of the complexities involved in the person choosing (or not) to visit a given place. Although the sequence of locations visited by a human may not be non-deterministic in an strict sense, in cases like this it is common to make a simplification and assume that the symbols of the sequence are generated by a stochastic process. Entropy is a good approximation of such random sequences' complexity because it is a lower bound on the sequences' compressibility [26]. Throughout this study, we will rely on this relationship between predictability and compressibility to first use a sequence's compressiblity to estimate its entropy and, then, use entropy to estimate the sequence's predictability.

It is important to note, however, that predictability limits obtained via entropy estimation are not hard bounds, as entropy estimation techniques, including the one used by Song et al., are *approximations* of the true entropy of the sequence. Computing true predictability would require the true Kolmogorov Complexity of the sequence, which is not computable. For practical purposes, however, off-the-shelf entropy estimators (e.g., [27, 28]) can be reliably used to estimate the complexity of sequences of symbols.

3.2 Entropy as a Measure of Complexity

Song et al.'s work relies on the fact that the predictability of a sequence of locations that a person visited is related to the entropy of the sequence [3]. Hence, their work proposes to assess predictability based on three estimates of the entropy of the sequence. The first estimate is the Shannon entropy [29] of a uniform distribution on possible locations (which is known to yield the highest possible entropy value), establishing a lower bound on predictability. The second variation computes the entropy of a refined distribution of locations reflecting the frequency with which users visit places. The third, more precise variation, estimates entropy while incorporating both frequency of visitations and temporal patterns. This more precise estimator, described by Kontoyiannis et al. [30], is related to the Lempel-Ziv compression algorithm [31] and to the Lempel-Ziv measure of the complexity of a sequence [31]. According to this estimator, the entropy S of an input sequence of locations X can be approximated by:

$$S_{real} \approx \frac{n \times \log_2(n)}{\sum_{i \in X} \Lambda_i}, \tag{1}$$

where Λ_i is the length of the shortest time-ordered sub-sequence starting at position i which does not appear from 1 to $i-1$ in the sequence X, and n is the size of the sequence.

For ergodic, stationary processes, this estimator converges to the entropy of the source as the size of the input goes to infinity [26]. This estimator does not require an explicit computation of the underlying probability distribution of the symbols of the source. As such, it is suitable for computing the entropy of mobility traces, for which we may never know the true underlying probability distribution.

Note that different values of entropy yield different limits of predictability: while the first two variations work by changing the underlying probability distribution of the locations, the third one leverages the relation between entropy and compressibility to estimate the entropy of the input sequence. Unless otherwise noted, *all of our references to the Song et al's technique refer to the third, more precise, entropy estimator.*

The predictability of a sequence is directly related to the entropy of the sequence, and can be defined as follows. Given a time-ordered sequence $X = (x_1, x_2, \ldots, x_{n-1})$ of input symbols where x_i is the i-th location visited by a person, the predictability Π_{max} of X is the maximum possible accuracy that any prediction algorithm[2] can achieve when trying to guess the next location $x_n \in X$.

The basic formula to compute the predictability Π_{max} takes as input an entropy value S and the number N of unique locations, and uses Fano's Inequality [26] to get an upper bound on predictability, as follows:

$$S = -H(\Pi_{max}) + (1 - \Pi_{max}) \log(N - 1), \tag{2}$$

where

$$H(\Pi_{max}) = \Pi_{max} \log_2(\Pi_{max}) + (1 - \Pi_{max}) \log_2(1 - \Pi_{max}).$$

A proof that these equations estimate the correct limits of predictability can be found in related work [3, 7, 1]. In particular, Smith et al. [7] provide a detailed, thorough derivation of the formula above.

The limits of predictability are thus directly related to a good estimate of the entropy. For that reason, throughout the rest of this study, we will focus on entropy estimates and not on the limits of predictability *per se*.

4 Methodology and Datasets

Recall that in Section 3, the predictability Π_{max} of a sequence X of locations is computed based solely on the data from which X is extracted. As such, Π_{max} is a fundamental expression of human behavior, as captured by that data. Thus, properties of the data, notably its spatial and temporal resolution, are of key concern to understanding Π_{max} values. In other words, both the prediction task under study (e.g., specific properties of the next symbol in the sequence) and the underlying datasets are factors delimiting the scope of the predictability study.

In this section, we define the particular scenarios defining the scope of our study, namely the datasets that we will use and the prediction tasks on which we will focus.

[2] Strictly speaking, Song et al.'s predictability estimates hold for a specific class of predictors, called *universal predictors.*

4.1 Datasets

Our study is composed and driven by a series of analyses performed on two different mobility datasets, of distinct temporal and spatial resolutions, which allow us to study the impact of spatiotemporal factors on Song et al.'s technique. These datasets are representatives of two categories of datasets often used in mobility studies: GPS datasets and call detail record (CDR) datasets. These datasets are summarized in Table 1 and discussed below.

	GPS dataset	CDR dataset
Number of users	45	2,780
Period covered	18 months	2 weeks
Min number of locations	577	170
Mean number of locations	2,388	174
Max number of locations	5,911	192
Temporal resolution	5 minutes	1 hour
Spatial resolution	200 meters	200 meters

Table 1: Summary of our GPS and CDR datasets.

GPS Dataset: This dataset consists of GPS traces with high temporal and spatial resolution. It as obtained through an Android mobile phone application, called MACACOApp.[3] Users who volunteered to install the app allowed it to collect data such as uplink/downlink traffic, available network connectivity, and visited GPS locations from their mobile devices. These activities are logged with a fixed periodicity of five minutes, making it a high temporal resolution dataset, and the precision in the acquisition of GPS coordinates from mobile devices makes it a high spatial resolution dataset as well. The regular sampling in this data provides a more comprehensive overview of a user's movement patterns. The dataset contains a total 45 users, spanning a period of 18 months, from July 10, 2014 to February 4, 2016. Notice that some users were active during only part of the study, which is reflected by the variability in the number of locations shown in Table 1 for this dataset.

CDR dataset: The second dataset spans a period of two weeks in 2015 and contains call detail records (CDRs) at the rate of one location per hour during that period. Each location in the trace represents the user's recorded location for the hour with the precision of 200 meters and was registered at the nearest phone tower. Notice that, unlike other CDR datasets, this one does not contain the area covered by each tower, but the data provider guarantees that the recorded position is the centroid of the positions of the user during the associated time period, within a 200 meter radius. As some users do not have data for the whole period, we focused on the ones who have at least one location registered each two hours, on average. This filtering criterion is the same adopted by Song et al. After this filtering process, we ended up with 2,780 users. The data was provided by a major cellular operator in China.

Unless otherwise noted, we will use a temporal resolution of one observation every five minutes for the GPS dataset and a spatial resolution of squared cells of side length of 200 meters. For the CDR dataset, there is at least one observation per user every two hours and the size of the side of each square grid is also 200 meters.

Preprocessing. The fundamental task regarding mobility prediction is to guess the next item in a sequence of symbols, but mobility data usually consists of latitude and longitude pairs, so it is necessary to preprocess the data to make it fit the expected format. For our purposes, it is also necessary to record location measurements at fixed time intervals. In order to do that, we discretized the time into bins of a given duration, and divided the geographical area into a grid of non-overlapping, uniformly spaced squares of equal sizes. We then distribute the activity records into the cells of the grid according to the location in which they were registered. Thus, the sequence of locations that a person visited becomes a sequence of integers containing the identifiers of the cells that correspond to those locations at each time bin.

4.2 Prediction Tasks

We here study predictability in human mobility considering two underlying prediction tasks: *next-cell* and *next-place* predictions. In both tasks, the goal is to predict the next location x_n in the input sequence X, but they differ in what type of location x_n can be: the next-cell prediction task aims at determining the value of x_n, whereas the next-place problem is to determine the first *distinct* location to appear in the sequence after x_{n-1}.

These simple definitions have important implications. In next-cell prediction, much of the prediction accuracy comes from stationary periods in a person's mobility, i.e., periods when the person stays in the same location for a long period

[3] `http://macaco.inria.fr/macacoapp/`

of time, therefore making the prediction of his or her next location relatively easy. For next-place prediction, however, stationary periods are not considered, since we are only interested in the next *distinct* location. in other words, this prediction task is concerned only with transitions between different places.

When doing next-place prediction, one can choose to simply ignore the next location if it is equal to the current one. When studying predictability, however, we focus on the data, i.e., we do not perform prediction. Therefore, we have to adapt the *dataset* to this prediction task. We do so by filtering the dataset so as to remove stationariry. Thus, in this study, when we refer to a particular dataset for the next-place prediction task, we are referring to the dataset after we filter out stationarity. Notice that, while filtering out stationary periods, we remove symbols from the sequences, therefore producing smaller sequences compared to the next-cell prediction task. As detailed in Section 6, the lack of stationarity and the smaller size of the sequences makes this prediction task harder than next-cell prediction.

5 Using Proxy Metrics to Understand Predictability

As we mentioned before, Song et al.'s technique has been extensively used to assess the predictability of mobility datasets. However, to our knowledge, no previous work has analyzed *how* predictability varies across different users and datasets. This is our goal in this section. To that end, we focus on the next-cell prediction task, as Song et al. [3], as well as on next-place prediction task. We investigate the entropy estimate across different users on both GPS and CDR datasets. We also present two metrics that capture key aspects of mobility patterns and show that they can effectively be used as proxies to understand predictability estimates. Finally, using these two metrics we analyze when and how the predictability estimate proposed by Song et al. changes with respect to the spatial and temporal granularity of the dataset.

We start by arguing that analyzing the entropy estimate itself, particularly the more precise one based on compressibility, which is our present focus, is quite challenging, as the result of the method is hard to interpret. Thus, we look for simpler and easier to understand proxy metrics, which can be used in its place to understand predictability in human mobility. Specifically, we employ two *simple* metrics that help explain what affects predictability in a sequence of locations visited by a user, as captured by Song et al.'s estimate.

The first metric, called the *stationarity* of a sequence of locations, is related to the number of observations for which the person stays continuously in the same location. Given a time-ordered sequence $X = \{x_1, x_2, \ldots, x_i, x_{i+1}, \ldots, x_n\}$ of observations of a person's location, we say that a *stationary transition* occurs at time i if $x_i = x_{i+1}$. Thus, we can define the stationarity of sequence X as the ratio of stationary transitions over the total number of transitions in X. The total number of transitions in a sequence of length n is equal to $n-1$. For example, sequence $X = \{1, 1, 2, 2, 3, 3, 4, 4\}$ contains a total of seven transitions, four of which are stationary. Therefore, the stationarity of the sequence $st(X)$ is $st(X) = 4/7 = 0.57$. Intuitively, if a person stays at the same location for a long period of time, there will be many consecutive repeated symbols in the sequence. Sequences with many consecutive repeated symbols are easier to compress, therefore the higher the stationarity of a sequence, the lower its entropy.

Yet, *stationarity* alone does not explain predictability. Consider, for instance, two input sequences $X_1 = \{1, 2, 3, 4, 1, 2, 3, 4\}$ and $X_2 = \{1, 2, 1, 2, 1, 2, 1, 2\}$. Both have the same length and the same stationarity, but X_2 has lower entropy than X_1: the entropy of X_2 is equal to 1.33, whereas the entropy of X_1 is 1.71.

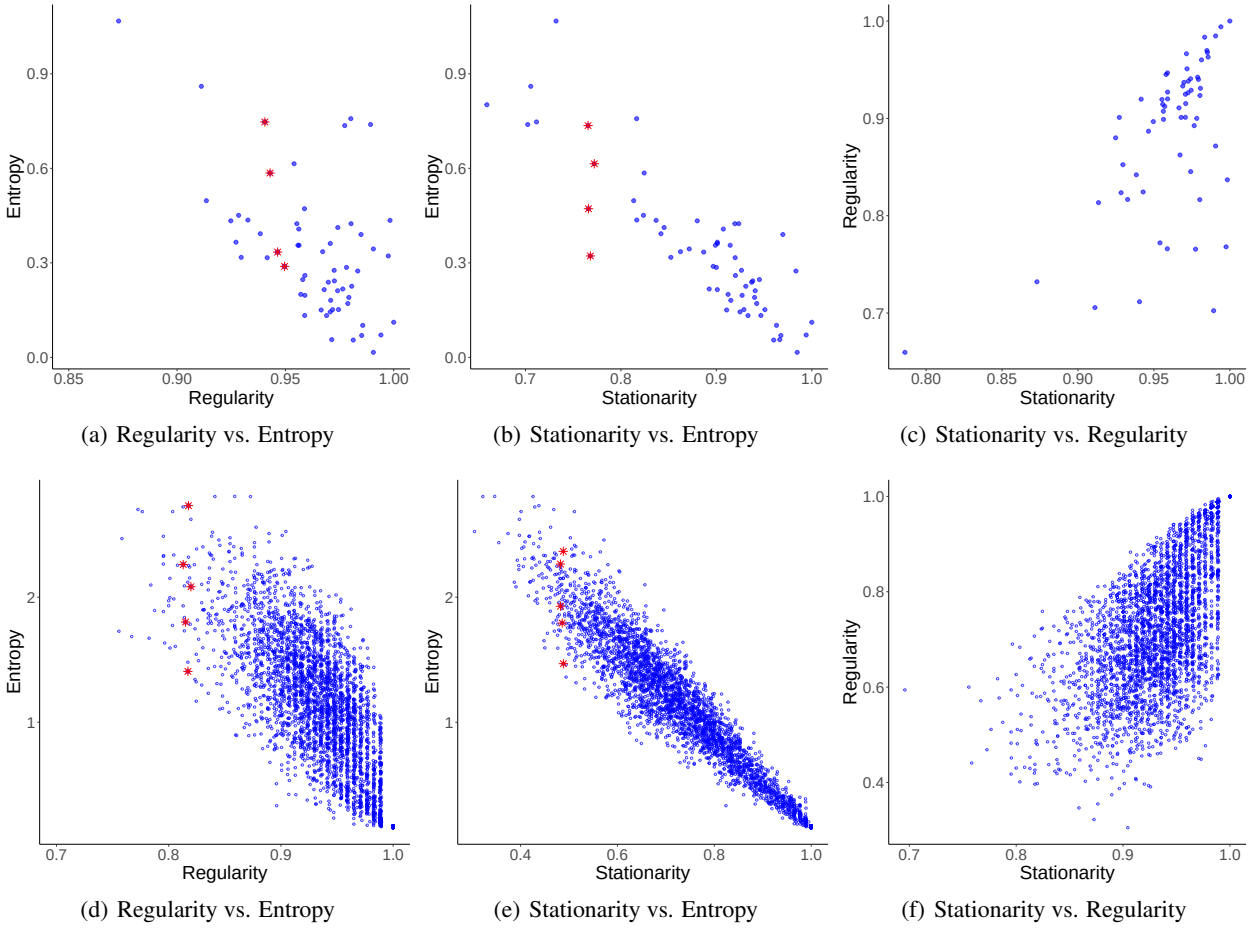
In order to capture that, we introduce another metric, called *regularity*, that also helps explain the entropy of a person's observed location history. The *regularity* of a sequence captures the preferences of a person to return to previously visited locations. It is defined as one minus the ratio between the number of *unique* symbols and the length of the sequence — notice that this is different from other regularity metrics such as *semantic regularity* [32], in the sense that we do not differentiate between the categories of the places a user visits. For instance, the regularity of input sequence $X = \{1, 2, 2, 3, 3, 3, 4, 4, 4, 4\}$ is given by $reg(L) = 1 - 4/10 = 0.6$. If we compute the *regularity* of the two aforementioned example sequences (X_1 and X_2), we obtain $reg(X_1) = 1 - 4/8 = 0.5$ and $reg(X_2) = 1 - 2/8 = 0.75$, which helps explain why X_2 has lower entropy than X_1 (X_2 is more regular than X_1). Intuitively, the more regular a sequence, the fewer distinct symbols it has, and sequences with few distinct symbols are easy to compress. Therefore, the higher the regularity, the lower the entropy of a sequence.

We note that our choice of metrics of regularity and stationarity comes from experimental observations of how Song et al.'s technique works. Intuitively, such metrics capture two key and complementary components of a person's mobility patterns: *the ratio between previously visited places and new places (irregular transitions)*, and *the amount of time spent in each place (stationary transitions)*. Although the importance of stationarity to predictability has been noted before [8], using regularity to help understand predictability and thoroughly evaluating both metrics in two different prediction tasks is a novel contribution of our work. These results are discussed next.

5.1 Regularity and Stationarity in the Next-Cell Prediction Task

In this section, we explain the relationship between the two metrics – regularity and stationarity – and entropy as well as how they can be used as proxies to understand predictability in the next-cell prediction task.

We first illustrate the relationship between entropy and regularity/stationarity by computing the Spearman correlation coefficient ρ [33] between entropy and the two metrics. Figures 1-a) and 1-d) show scatter plots with the relationship between regularity and entropy for the GPS and CDR datasets, respectively. Similar plots for stationarity and entropy are shown in Figures 1-b) and 1-e). According to these figures, we found a correlation between regularity and entropy equal to -0.47, and a correlation between entropy and stationarity equal to -0.83 for the GPS dataset. Corresponding values for the CDR dataset are -0.74 and -0.94, respectively.



(a) Regularity vs. Entropy (b) Stationarity vs. Entropy (c) Stationarity vs. Regularity

(d) Regularity vs. Entropy (e) Stationarity vs. Entropy (f) Stationarity vs. Regularity

Figure 1: Relationship between our two metrics and entropy (GPS dataset: a-c; CDR dataset: d-f). Notice that the plots are at different scales. The red dots represent users who have similar values for one of the metrics, but very different entropy values, indicating that one of the metrics alone is not able to fully explain entropy.

Moreover, these two metrics, regularity and stationarity, are themselves reasonably well correlated (0.46 and 0.63 in the GPS and CDR datasets, respectively), as illustrated in Figures 1-c) and 1-f). Thus, the relationships between these metrics and entropy are far from perfect. Indeed, a manual inspection of our datasets revealed that there are several users who, despite having similar regularity (or stationarity) have very different entropy values —see the red points in Figures 1(a, b, d, e). This suggests that each variable, in isolation, cannot explain entropy. Investigating such cases, we found that large differences in entropy for users with similar regularity could often be explained by great differences in stationarity, and vice versa. For instance, the users in red in Figure 1(b) have stationarity equal to 78%, but different entropy (ranging from 0.32 to 0.86, respectively). These observations suggest a hypothesis the two metrics are, to some extent, complementary, and that the two of them, in conjunction, could explain predictability better than either of them in isolation.

In order to verify this hypothesis, we analyzed the extent to which each metric alone versus the two in conjunction can explain the predictability of a sequence of locations. To that end, we employed a regression analysis by fitting the entropy $H(X)$ of a sequence X as a function of: (i) regularity $reg(X)$ alone, (ii) stationarity $st(X)$ alone, and (iii) as a function of both metrics, for all users in each dataset. For the latter, we experimented with different regression functions and the one that led to the best fitted model is given by:

$$H(X) \approx \alpha \cdot reg(X) + \beta \cdot st(X) + \gamma \cdot reg(X)st(X) + \epsilon, \tag{3}$$

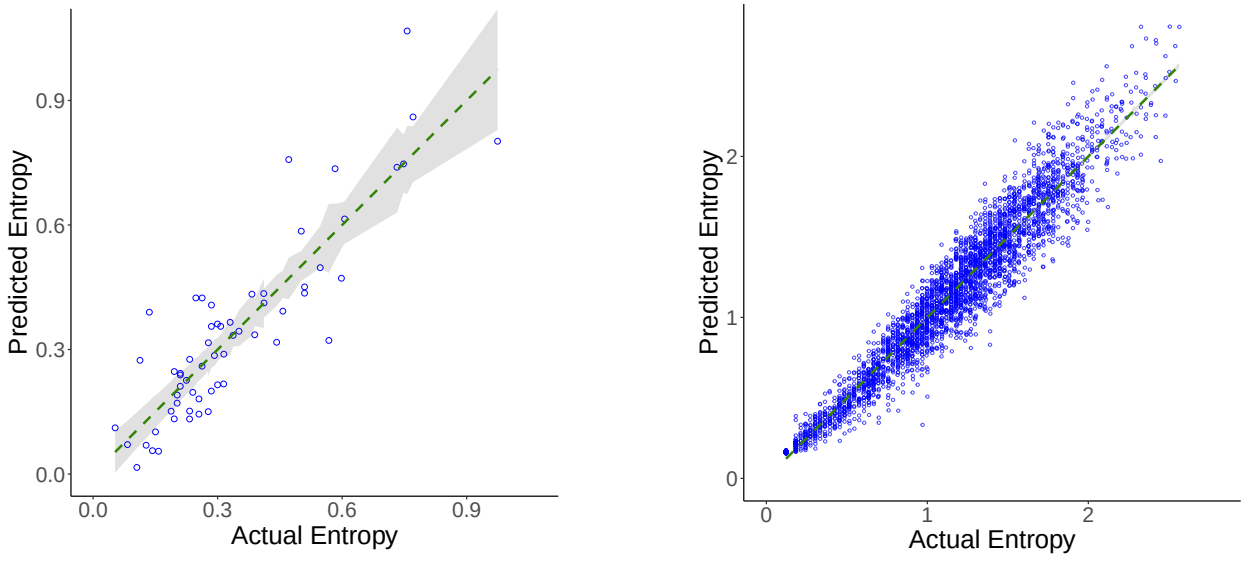
where α, β, and γ are the coefficients of regression and ϵ is the regression error. Furthermore, it was necessary to consider the interaction between the two metrics because there is a confounding effect between them. This function was chosen to illustrate that the two proposed metrics can by themselves explain most of the variation observed in the entropy values and, as such, can be used as proxies for understanding the entropy of a person's location history. Among all regression models we tested with the two variables, it was the one that produced the best fittings.

This model, albeit simple, is able to explain a large fraction of the total variation in the entropy values *in both datasets*. It also shows better entropy fittings when compared to two other models that employ only regularity or stationarity, as shown by the adjusted R^2 of the models listed in Table 2. Additionally, as we further discuss in Section 5.3, we experimented with different spatial resolutions. We found that the model in Equation 3 also performed well for other spatial resolutions we tested. For instance, for the GPS dataset, the R^2 varied from 0.761 (highest spatial resolution, smaller cells) to 0.834 (lowest spatial resolution, larger cells).

	$reg(X)$ only	$st(X)$ only	$reg(X)$ and $st(X)$
GPS dataset	0.318	0.758	0.761
CDR dataset	0.565	0.903	0.936

Table 2: Adjusted R^2 of three different regression models, each of which using a combination of our metrics, for both the GPS and CDR datasets in the next-cell prediction task.

Figure 2 shows scatter plots of the actual entropy (x-axis) versus entropy estimated by Equation 3 (y-axis) for all users in both datasets. Notice that most dots (users) lie close to the diagonal, especially in the larger CDR dataset. Therefore, regularity and stationarity can indeed be used as proxies for the purpose of studying predictability in human mobility.



(a) GPS dataset

(b) CDR dataset

Figure 2: Entropy (in bits per symbol) predicted by the regression model (y-axis) versus actual entropy (x-axis), in bits per symbol, for the next-cell prediction task. The green, dashed line shows the regression model (Equation 3) and the gray area shows the confidence interval. As there are many more data points in the CDR dataset, the confidence interval area is narrower and almost invisible in the plot.

It is important to note that, as shown in Table 2, the model that uses only stationarity performed almost as well as the one that uses both metrics, in both datasets. This fact illustrates the importance of stationarity for the next-cell prediction task, and suggests that most of the predictability (i.e., achievable prediction accuracy) stems from stationary behavior in the next-cell prediction task. In the next section, we will discuss what happens when stationarity is taken out of the equation, i.e., in the next-place prediction task.

5.2 Regularity in the Next-Place Prediction Task

In this section, we evaluate the extent to which regularity helps understand and interpret predictability results in the next-place prediction task. Recall from Section 4 that in the next-place prediction task, there is no stationarity. In other words, given a sequence $X = \{x_1, x_2, \ldots, x_{n-1}\}$, we are interested in estimating the maximum achievable accuracy when trying to predict x_n, where x_n is different from x_{n-1}.

In the last section, we showed that stationarity plays a central role in explaining predictability in the next-cell prediction task. In fact, the role of stationarity is significantly larger than that of regularity. In this section, we evaluate how the role of regularity changes when there is no stationarity involved. Specifically, we would like to answer the following questions: *Does the importance of regularity increase in the next-place prediction task when compared to next-cell prediction? Does this increase make up for the lack of stationarity?*

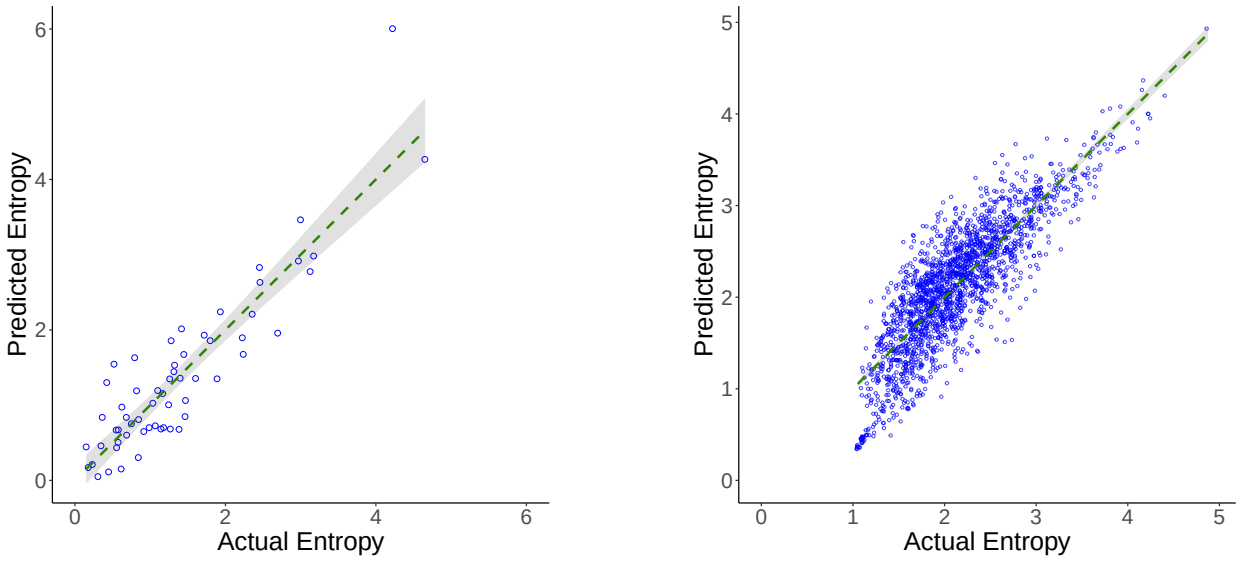
To answer the first of these questions, we examine the Spearman correlation coefficient between regularity and entropy in the next-place prediction task. We found that the correlation is -0.83 and -0.82, for the GPS and CDR dataset, respectively. Contrast these values with the corresponding values in the next-cell prediction task: -0.47 and -0.74 for the GPS and CDR dataset, respectively. Thus, regularity indeed plays a larger role in next-place prediction than it does in next-cell prediction.

To answer the second question, we build a regression model that uses regularity to fit the entropy of next-place prediction. Our regression model is as follows:

$$H(X) \approx \alpha \cdot reg(X) + \epsilon, \tag{4}$$

where α is the coefficient of regression and ϵ is the regression error.

We evaluate this model in both of our datasets and discover that the adjusted R^2 is 0.681 and 0.709 for the GPS and CDR datasets, respectively. Figure 3 shows the entropy fittings for the resulting model.



(a) GPS dataset

(b) CDR dataset

Figure 3: Entropy (in bits per symbol) predicted by the regression model (y-axis) versus actual entropy (x-axis), in bits per symbol, for the next-place prediction task. The green, dashed line shows the regression model (Equation 4) and the gray area shows the confidence interval. As there are many more data points in the CDR dataset, the confidence interval area is narrower and almost invisible in the plot.

From the R^2 of the model as well as from Figure 3, we observe that although regularity plays a larger role in next-place prediction, its increased importance cannot completely make up for the lack of stationarity when trying to explain predictability in the next-place prediction task. As argued in Section 2 and shown in our results, this is a harder task, and further work is necessary to fully understand what drives and affects next-place prediction.

5.3 Spatiotemporal Interplay

Previous studies [7, 8, 34] have shown that the estimate of predictability in mobility is influenced by the temporal and spatial resolutions of the data. Specifically, greater predictability is expected as temporal resolution increases (more observations per time period) or spatial resolution decreases (larger cells). We now revisit this analysis by looking into how both factors affect regularity and stationarity.

Table 3 shows how average regularity and average stationarity vary as temporal resolution varies in our GPS dataset, for both next-cell and next-place prediction tasks. We can only perform this analysis in this dataset, as its higher temporal resolution (compared to the CDR dataset), allows for time bins with various durations. For each given time bin, we randomly pick one observation in the interval for analysis.

As shown in the table, a decrease in temporal resolution makes the average stationarity decrease as well (third column). This occurs because the longer time intervals between measurements make it more likely for those measurements to occur at different locations—there is a higher chance that the user moved in a longer time interval. The decrease in stationarity leads to an increase in entropy and thus, lower predictability (on average). A less obvious observation is that a decrease in temporal resolution reduces average regularity. This is due to the fact that lower temporal resolution means fewer observations being made overall, which leads to shorter sequences. Recall that regularity is related to the ratio between the number of unique symbols and the size of the sequence. Therefore, a reduction in the size of the sequence will cause a decrease in regularity. In general, less regular sequences will have larger entropy, as shown in the last column of Table 3.

	Next-Cell			*Next-Place*	
Temporal Resolution (mins)	**Average Stationarity (%)**	**Average Regularity (%)**	**Average Entropy**	**Average Regularity (%)**	**Average Entropy**
5	96.0	88.5	0.340	81.6	1.359
10	93.8	84.0	0.611	82.0	1.422
20	92.0	80.3	0.821	80.1	1.642
30	90.8	77.6	0.980	79.1	1.748
40	89.8	75.3	1.100	77.9	1.804
50	88.9	73.6	1.191	77.1	1.878
60	87.9	71.8	1.261	76.4	1.940

Table 3: Variation of regularity, stationarity, and entropy according to the temporal resolution of the data. The results are shown for the GPS dataset, and for both (next-cell and next-place) prediction tasks.

We now turn our attention to the relation among spatial resolution, regularity, and stationarity. As with the temporal resolution, it is only possible to perform this analysis on the GPS dataset: as it has high spatial resolution, we can tesselate grids of arbitrary size on the target geographical area. A decrease in spatial resolution means that the cells in the spatial grid are larger, which means that more measurements are going to me made inside the same cell, thus increasing average stationarity, as shown in Table 4. The entropy decreases as stationarity decreases.

A decrease in spatial resolution, in turn, also causes an increase in regularity, as it will be less likely that a person moves outside a larger cell. In Table 4, we show how entropy decreases as regularity increases, both on average.

We summarize the discussion in this section as follows:

- The reduction in entropy (and corresponding increase in predictability) seen as the spatial resolution increases comes at a cost. If the dataset is broken into a grid of larger cells, most of the user's activity tends to be confined within fewer cells, with two opposing effects. On one hand, predictability, or prediction *accuracy*, increases, since it becomes relatively easier to correctly infer the user's next location. On the other hand, prediction *utility* degrades, since the bigger the region the user is predicted to visit, the less informative the corresponding prediction is. In the extreme case of a grid with a single region, prediction is always trivially correct but it is also of little use. Hence, by adopting grids with higher resolution (i.e., smaller cells), it is possible to increase prediction utility, but at the possible cost of hurting accuracy.

	Next-Cell			*Next-Place*	
Spatial Resolution (m)	**Average Stationarity (%)**	**Average Regularity (%)**	**Average Entropy**	**Average Regularity (%)**	**Average Entropy**
200	96.0	88.5	0.340	81.6	1.359
300	96.5	88.9	0.333	84.2	1.262
400	96.9	89.7	0.305	86.1	1.179
500	97.2	92.2	0.262	87.8	1.047
600	97.4	92.5	0.256	88.6	1.036
700	97.7	92.7	0.244	89.6	1.029
800	97.9	93.6	0.228	90.4	0.971
900	98.0	93.7	0.226	90.8	0.953
1000	98.1	94.0	0.227	91.4	0.917

Table 4: Variation of regularity, stationarity, and entropy according to the spatial resolution of the data. The results are shown for the GPS dataset, and for both (next-cell and next-place) prediction tasks.

- As discussed above, users with high regularity and stationarity also exhibit high predictability. We now argue that these two metrics could be exploited not only for understanding predictability but also for practical applications. For instance, they could be used to decide whether or not include a user's data in a data sample to be released to the public, with some care as to whether users with high regularity or stationarity should be included in the sample. For a given temporal and spatial resolution, and assuming a uniform temporal sampling, if a user is highly stationary, revealing her location in a given moment in time may also reveal her location for the next hours, which is privacy compromising. Furthermore, depending on the dataset, it should be possible to release only a few metrics related to the mobility of each user, instead of releasing the complete mobility trace. We argue that regularity and stationarity are examples of such metrics.

- Stationarity and regularity explain most of the variability in the entropy of a sequence of locations, but there seems to be something else at play here. Intuitively, one expects that external factors such as day of the week, hour of the day, weather conditions, and even socio-economic factors play a role in a person's mobility patterns. While these types of information affect people's mobility patterns, the state-of-the-art technique for computing the limits of predictability in human mobility does not take them into account. In the next section, we investigate how to add such types of (contextual) information into the computation of the limits of predictability.

6 Predictability and Context

Recall from Section 3 that predictability is a function of the entropy of the sequence of visited locations. However, given prior arguments that contextual information may indeed improve the predictability of one's mobility [8, 9], we would like to use not only the history of visited locations while computing the entropy, but also contextual information associated with each visit. In this section, we study different strategies to incorporate such side information into entropy (and thus, predictability) estimates, quantifying its impact on those estimates.

We start by investigating how to explore context using entropy estimators that are based on the frequency (probability) with which the locations are visited (Section 6.1). We choose to focus first on those entropy estimators, which are alternatives to the state-of-the-art compression-based approach discussed in the previous section, because extending them to incorporate context is easier and more intuitive. After quantifying the impact of context into these entropy estimators, we then move on to explore the more challenging task of adding context to the compression-based estimator used by Song et al. (Section 6.2).

In both sections, we consider three types of contextual information, namely day of the week, hour of the day, and weather information. The latter, obtained through an an external service[4], is only available for our CDR dataset as we were unable to gather weather information for the period and location covered by the GPS dataset. For the CDR dataset, the weather information corresponds to descriptions of the weather (clouds, rain, snow, etc.) which are mapped into 7 distinct and unique integer identifiers.

[4] `https://openweathermap.org/`

6.1 Adding Context to Predictability Estimates

Given $X = \{x_1, x_2, \ldots, x_n\}$, a time-ordered sequence of locations, and $C = \{c_1, c_2, \ldots, c_n\}$, a sequence of contextual information associated to each of the visits —c_i could be the weather when the person visited location x_i, for instance—, we wish to measure the extent to which knowing sequence C helps estimating the entropy of X. In other words, we wish to know how much X is constrained (or influenced by) C, which can be determined by the *conditional entropy* $H(X \mid C)$ [26], computed as follows:

$$H(X \mid C) = H(X, C) - H(C), \tag{5}$$

where $H(X, C)$ is the *joint entropy* of X and C, given by

$$H(X, C) = -\sum_{x \in X, c \in C} p(x, c) \log_2 p(x, c), \tag{6}$$

and $p(x)$ is the *probability mass function* of variable X given by $p(x_i) = Pr(X = x_i)$. In Equation 5, if X and C are independent, i.e., if C carries no information about X, it follows that $H(X, C) = H(X) + H(C)$, which leads to $H(X \mid C) = H(X, C) - H(C) = H(X)$. Once we have $H(X|C)$ we use it in Equation 2 to compute the predictability of sequence X constrained by the contextual information in C.

Notice from Equations 5 and 6 that the basis for entropy computation is an underlying *probability distribution.* Thus, if one has the *full* probability distribution of a sequence of symbols X, the entropy of that sequence is given by Shannon's formula: $H = -\sum p(x) \log_2 p(x)$. The same is true for the joint entropy of X and C. In real world situations, however, one usually has access to only a *sample* drawn from the underlying probability distribution. As a consequence, entropy values obtained for a sequence are *estimates* of the real entropy of the sequence. Entropy estimators that are based on the probability distribution inferred from a sample usually compensate for the effects of using such sample by adding a bias term to their probability estimates. Different estimators exploring different bias terms exist in the literature [28], but in general their entropy estimates tend to be more conservative (than the exact values) because of the added bias term.

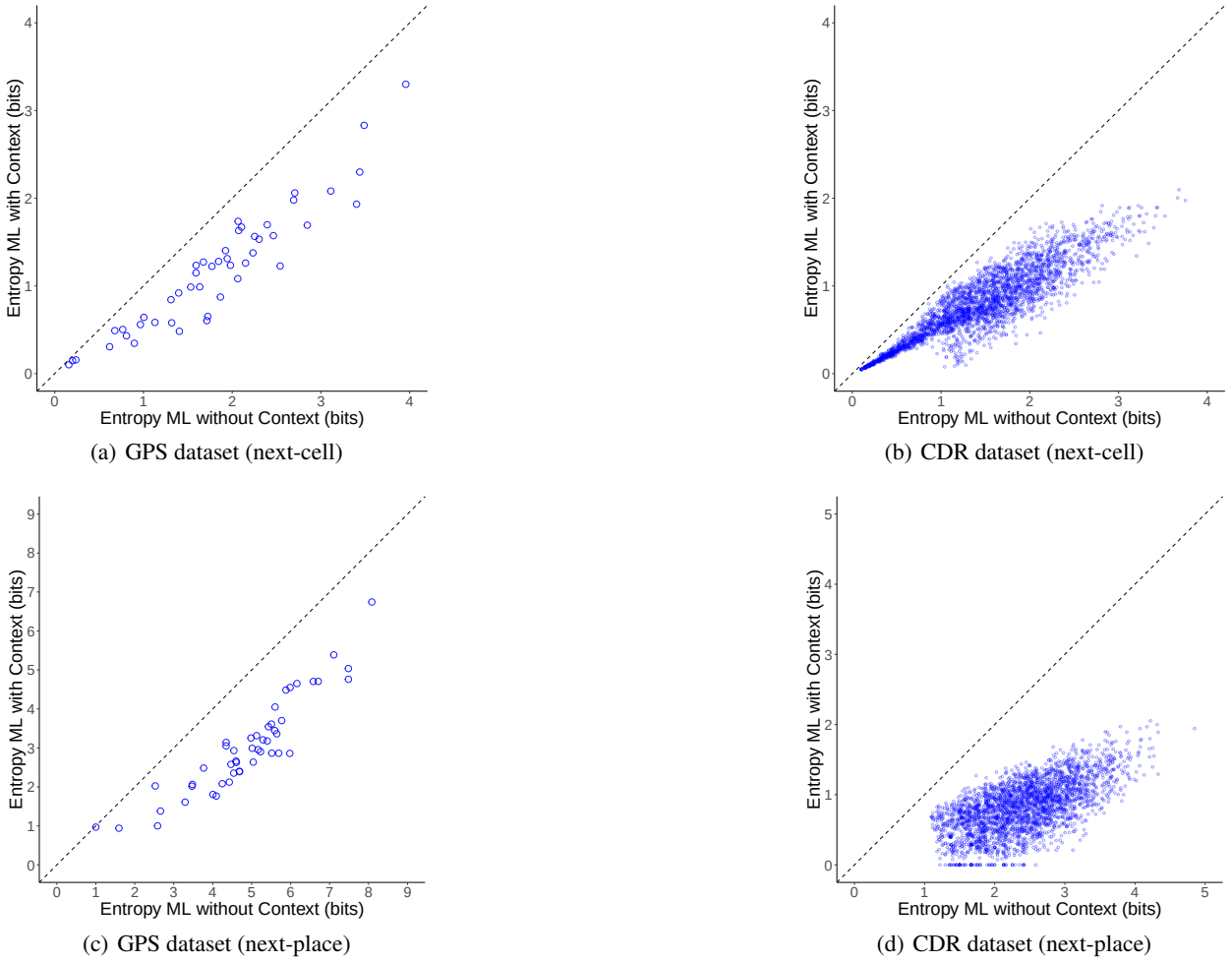
We experimented with various frequency (probability) based entropy estimators, choosing three of them that delivered the best results in our preliminary experiments.[5] The first one is called *Maximum Likelihood* (ML), which estimates entropy using the empirical frequencies of observations, and therefore is equivalent to Shannon entropy [26]. This estimator is also used in Song et al.'s work as a baseline for comparison against the more refined compression-based estimator, explained in Section 3. The second one, called *Miller-Madow* (MM), estimates entropy by applying the Miller-Madow bias correction [35] to Shannon entropy. The third one, called *SG*, estimates entropy using the Dirichlet multinomial pseudo-count model [36] with parameter $a = 1/n$ where n is the length of input sequence X. All of these estimators directly apply Equations 5 and 6 to compute the predictability of sequence X given sequence C.

		GPS dataset			CDR dataset			
		No Context	**Weekday**	**Hour**	**No Context**	**Weekday**	**Hour**	**Weather**
Next-Cell	**Maximum Likelihood**	6.01	1.48	1.18	1.45	1.13	0.79	1.23
	Miller-Madow	10.6	1.61	1.22	7.86	1.20	0.91	1.32
	SG	4.62	1.48	1.19	2.66	1.17	0.84	1.27
Next-Place	**Maximum Likelihood**	4.82	3.80	2.98	2.39	1.74	0.59	1.72
	Miller-Madow	5.55	4.17	3.36	8.80	2.05	1.02	1.98
	SG	5.55	3.85	3.07	2.92	1.87	1.02	1.84

Table 5: Evaluation of three entropy estimators in both datasets and for the two prediction tasks (next-cell and next-place). The reported average entropy values are given in bits per symbol (each location is a symbol in the input sequence). For probability-based entropy estimators, context reduces the entropy of the original sequence.

Table 5 shows the entropy estimates produced by these three estimators with and without context, for our two datasets, three types of contextual information, and two prediction tasks. As shown in the table, *context does enhance predictability estimates*, i.e., entropy values with context are much lower than those without context. The gaps are larger when the hour of the day is used as context, which *suggests stronger ties between hour of the day and location*. For instance, a person may visit different locations every day of the week, but almost always stays at home from midnight to early morning, or at workplace during morning and afternoon hours.

[5] These three estimators, along with others, are available as off-the-shelf tools in the R package called *entropy* [28].



(a) GPS dataset (next-cell)

(b) CDR dataset (next-cell)

(c) GPS dataset (next-place)

(d) CDR dataset (next-place)

Figure 4: Reduction in the entropy values when contextual information (hour of the day in this case) is applied to the Maximum Likelihood estimator, in both datasets and prediction tasks.

We further illustrate these enhancements by showing, in Figure 4, scatter plots of entropy values with and without context (hour of day) for the ML estimator (the best of the three estimators in Table 5). As shown in these figures, the entropy values for all users were reduced when context was used, for both datasets and prediction tasks. These results confirm the intuition that *human mobility is constrained by several factors*. As Table 5 shows, some of these factors can be related to people's routine, e.g., day of the week and hour of the day, but *external* factors such as weather also have the ability to influence people's mobility.

6.2 Context and Song et al.'s Estimator

In Section 6.1, we showed that context reduces the entropy of probability-based entropy estimators. However, in the case of the estimator employed in Song et al.'s work, it is not possible to exploit contextual information by directly applying Equations 5 and 6. Recall from Section 3 that the algorithm used by Song et al. (originally proposed by Kontoyiannis et al. [30]) works by compressing the input sequence of symbols to estimate its entropy, therefore leveraging the relation between entropy and compressibility. In doing so, the algorithm becomes oblivious to the underlying probability distribution of the symbols in the sequence, which poses a barrier to computing the conditional entropy using Equations 5 and 6.

We here investigate two strategies to circumvent the aforementioned barrier and incorporate context into Song et al.'s estimator, thus using a compression strategy instead of a probability one. The first one, referred to as *sequence-splitting* is based on breaking the original sequence of locations into sub-sequences conditioned to specific contexts. The second

one builds a new sequence by combining locations and associated contexts. It is referred to as *sequence-merging*. We discuss both strategies next.

Sequence-Splitting. Our first approach relies on splitting the original sequence X according to the contextual information into consideration and on computing the entropy for visits that occur with the same context [10]. In other words, we basically hard-code context into each sub-sequence and in the end, use the entropy of those sub-sequences to obtain the entropy of the original one.

We will illustrate how this strategy works through an example, shown in Figure 5. Let's assume we want to use weather as contextual information (i.e., sequence C in the figure), discretized into three different types (e.g., sunny, cloudy, rainy, represented by the symbols sun, cloud, and umbrella in the figure). To do that, we split the original sequence X into three sub-sequences, one for each type of weather, each of which contains all of the locations visited when the weather was of the same given type. We then, run the entropy estimation algorithm in each of the three sequences, taking the weighted average of the results, to consider differences in the size of the sequences.

More formally, let $X = \{x_1, x_2, \ldots, x_n\}$ be the original sequence and $C = \{c_1, c_2, \ldots, c_n\}$ be the contextual information sequence. Moreover, let k be the number of *distinct* elements in C, i.e., different contexts in C. We split X into sub-sequences $X_1, X_2, \ldots, X_k$, so that each X_j is the (sub)sequence of locations in the original sequence X which are associated with the same type of context $c_j, j = 1..k$ in sequence C. We then apply Equation 1 to each X_j, taking the weighted average at the end, where the weight is the size of each sequence.

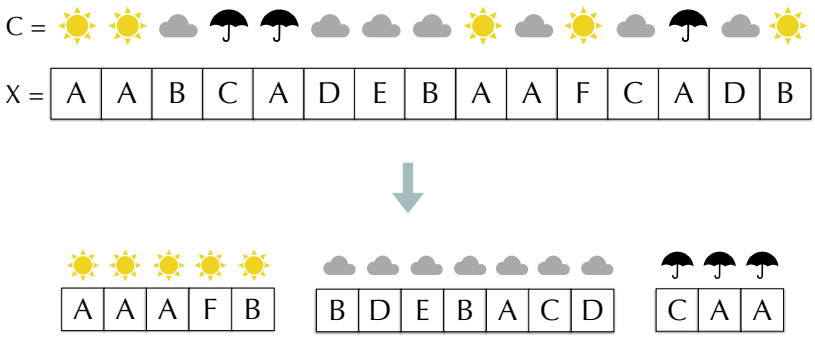


Figure 5: Example of our sequence-splitting strategy. We divide the original sequence into sub-sequences according to each type of context.

Sequence-Merging. Our second strategy relies on the fact that, by combining locations and contexts in the same sequence, we can estimate their joint distribution using a compression-based estimator. In a nutshell, our sequence-merging approach is based on an analogy with Equation 5. We propose to estimate the conditional entropy using a compression-based entropy estimator, such as the one used by Song et al. Defining $H_c(X)$ as the *compression-based entropy* of sequence X, and $H_c(X, C)$ as the *joint compression-based entropy* of X and C, we have that:

$$H_c(X \mid C) = H_c(X, C) - H_c(C), \tag{7}$$

where $H_c(X, C)$ is the *joint (compression-based) entropy* of X and Y.

The computation of $H_c(C)$ is a direct application of Song et al.'s estimator on sequence C, so the challenge lies in computing $H_c(X, C)$. What follows is a procedure to do so.

The key insight to computing $H_c(X, C)$, illustrated in Figure 6, is to merge sequences X and C into a sequence X', where each symbol is now a pair (x, c) with $x \in X$ being a location and $c \in C$ being a context. Recall that, according to Song et al.'s estimator, the entropy is inversely proportional to the number of repeated sub-sequences in the target sequence. As we compute the entropy of a sequence using this estimator, we keep track of every sub-sequence encountered, so that further sub-sequences can be matched against the previously discovered ones. Assuming that X is somehow related to C, i.e., there will be repeated location-context pairs throughout the new sequence X', which may help us obtain lower entropy for sequence X by using sequence C as context.

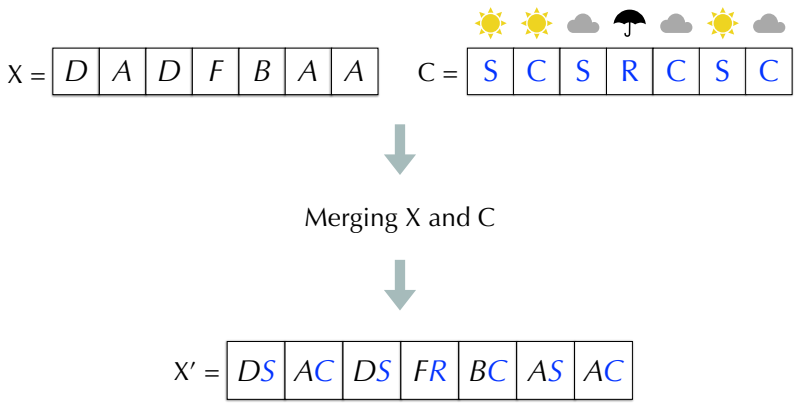


Figure 6: Example of our sequence-merging strategy. Each symbol $x_i \in X$ is combined with the associated context $c_i \in C$ to form sequence X'.

More formally, consider two sequences $X = \{x_1, x_2, \ldots, x_n\}$ and $C = \{c_1, c_2, \ldots, c_n\}$. It is possible that some symbols in C tend to appear together with some symbols in X, e.g., when it rains, one tends to stay at home. When we build sequence $X' = \{(x_1, c_1), (x_2, c_2), \ldots, (x_n, c_n)\}$ by merging sequences X and C, some pairs $(x_i, c_i), 1 \leq i < n$, will appear at several points in sequence X. This effect can also happen with several pairs that appear consecutively, i.e., $(x_i, c_i), \ldots, (x_j, c_j), 1 \leq i < j \leq n/2$. In other words, compressing (estimating the entropy of) X' may require fewer bits than the sum of the bits required to compress X and C isolated.

		GPS dataset			CDR dataset			
		No Context	**Weekday**	**Hour**	**No Context**	**Weekday**	**Hour**	**Weather**
Next-Cell	**Maximum Likelihood**	6.01	1.48	1.18	1.45	1.13	0.79	1.23
	Sequence-Splitting	0.34	0.46	0.90	1.10	1.42	1.62	1.43
	Sequence-Merging	0.34	0.35	0.50	1.10	1.27	1.01	1.04
Next-Place	**Maximum Likelihood**	4.82	3.80	2.98	2.39	1.74	0.59	1.72
	Sequence-Splitting	1.36	1.58	1.96	1.96	1.87	1.20	1.94
	Sequence-Merging	1.36	1.43	1.62	1.96	2.03	1.62	1.74

Table 6: Evaluation of our sequence-splitting and sequence-merging strategies (compared to the best estimator from Section 6.1) in both datasets and for the two prediction tasks (next-cell and next-place). The reported average entropy values are given in bits per symbol (each location is a symbol in the input sequence).

Table 6 shows a comparison of the two proposed strategies, namely sequence-splitting and sequence-merging, to the overall best probability-based estimator from Section 6.1, the Maximum Likelihood (ML) estimator. For each estimator, the table shows average entropy values with and without context, for both datasets, both prediction tasks, and all types of contexts considered. Note that for both sequence-splitting and sequence-merging, the results without context are those estimated by the original Song et al.'s estimator. The results for the ML estimator are the same as in Table 5, shown again here to facilitate comparison.

There are three key observations to make out of the results in Table 5. First, we note that in all cases without context, *Song et al.'s estimator does produce lower entropy values that the ML estimator.* In other words, it is indeed a very good entropy estimator, justifying its broad use to estimate predictability in human mobility [3]. Second, we also note that introducing context into this estimator, by applying either the sequence-splitting or the sequence merging approach, can *yield lower entropy values than the ML estimator with context* in several cases, especially for the GPS dataset.

Yet, quite strikingly and perhaps most importantly, the table also shows that, unlike observed for the ML estimator (and other probability-based entropy estimators), *the introduction of context into the Song et al.'s estimator, according to our sequence-splitting and sequence-merging strategies, often leads to an increase in entropy (lower predictability), compared to the estimated entropy without context.* Out of all scenarios analyzed, adding context only leads to reduced entropy when the sequence-splitting strategy is used on the CDR dataset and for the next-place task. In that case, there are reductions on entropy values, especially if hour of the day is used as contextual information.

The negative results for both sequence-splitting and sequence-merging in all other scenarios may be at first counter-intuitive, and thus, calls for a deeper investigation on the challenges of using context together with the compression-based entropy estimator proposed by Song et al.

Sequence-Splitting challenges. Let's start with the sequence-splitting approach. Recall from Equation 1 that Song et al.'s entropy estimate converges to the real entropy as the sequence grows to infinity, therefore being influenced by the length of the sequence. That is, the larger the sequence the better the entropy estimate. Thus, though useful, this strategy suffers from the downside that, by splitting the original input sequence, we are effectively estimating the entropy of *smaller* sequences. *Song et al.'s estimator has trouble converging to the real entropy for such small sequences*, and we end up with possibly inflated entropy values.

As an example, consider a sequence $X_s = \{A_0, A_1, \ldots, A_{99}\}$ of size 100, where all observations consist of the same symbol (a completely stationary sequence). The entropy of this sequence, according to Song et al.'s estimator is 0.26. Further, suppose that each block of 25 consecutive symbols of sequence X_s is associated with a different context. Following our sequence-splitting approach, we divide this sequence according to each context, which results in four sub-sequences of size 25, each of which has entropy 0.68. Thus, the entropy of the original sequence X_s is $4 \cdot 0.68/4 = 0.68$, which is higher than the entropy of the original sequence. Thus, changes in context during stationary periods can lead to higher estimates of entropy values.

Consider now a more realistic scenario of using hour of the day as contextual information. In that case, the history of locations of each user is split into 24 sequences, one for each hour. *This division results in sequences with considerably smaller sizes, which makes it harder for Song et al.'s estimator to converge to the real entropy.* This may be further aggravated by the splitting of longer stationary periods that span more than one hour into separate sequences, which also contributes to raising the final entropy estimate. This explains why this approach performed poorly (i.e., its entropy with context was higher than the one without context) for the next-cell task, for which longer stationary periods greatly contribute to the real entropy of the original sequence. In the case of the CDR dataset, as the period covered by the data is smaller, and the temporal resolution is lower (fewer observations per time unit), there is reduction in stationarity, as argued in Section 5.3, which aleviates the problem we just described. In the case of next-place prediction in the CDR dataset, as there is no stationarity involved, we observe a reduction in entropy values when context is used in the sequence-splitting approach.

Sequence-Merging challenges. Although fundamentally different from sequence-splitting, the sequence-merging strategy suffers from a somewhat similar problem. Recall that, when building sequence X', each symbol $x_i \in X$ is combined with the corresponding context $c_i \in C$ to form a tuple location-context in the form (x_i, c_i). This combination of symbols produces a new sequence X' which is much more complex than the original sequence X in terms of unique symbols, as the alphabet of X' is the cartesian product of the alphabets of sequences X and C. Song et al.'s estimator is based on the Lempel-Ziv compression algorithm, which is a universal compressor [26]. These compressors learn the distribution of symbols in the input sequence on-the-fly. Thus, for more complex inputs they may take longer to learn the underlying distribution of symbols. *If the input sequence is not long enough, such methods may produce inaccurate entropy estimates.* Thus, given the higher complexity of X', compared to the original sequence X, Song et al.'s estimator may indeed produce quite inflated estimates of entropy, as observed in Table 6.

As an example, consider the two following sequences: $X = \{A, B, A, B, A, B, A, B, A, B, A, B\}$, and $C = \{1, 1, 2, 2, 3, 3, 4, 4, 5, 5, 6, 6, 7, 7\}$. The entropy of X alone is 1.00, and the entropy of C is 2.66, both computed using Song et al.'s estimator. Yet, the entropy $H_c(X \mid C)$ is 1.14, i.e., using C to estimate the entropy of X actually increased the entropy (compared to the entropy without context). It is important to note that, though theoretically context can never increase entropy ("information cannot hurt" property [26]), it is not clear whether the use of context can actually help predictability in practice, when using compression-based entropy estimators such as Song et al.'s. In particular, we cannot compute the conditional entropy directly, but only estimate it through universal compressors, which may need a large sequence to learn the input distribution and start approximating the real entropy closely.

Common challenges. *Another aspect that is important to consider regarding both sequence-splitting and sequence-merging is the variability of symbols in sequence C.* For the former, less variability implies fewer (and longer) sub-sequences, which favors the convergence of the entropy estimator. For the latter, little variability means a smaller alphabet, which makes it easier for the entropy estimator to converge. Out of the three types of context considered, weather has the lower variability: in total there are seven types of weather in our dataset, but four of them appear only in three days out of the two weeks analyzed. Such lower variability may have contributed to sequence-splitting producing improved estimates in the CDR dataset.

Summary. *Introducing context into Song et al's estimator is clearly quite challenging*. Several interdependent factors play a role in the convergence of the estimator to the real entropy, making it hard to know when introducing context will actually be helpful, i.e., producing lower entropy values. *Our discussion in this section suggests that for highly stationary or highly regular location sequences, the estimate of the entropy with context may be higher than that without it.* This may also be the case for a very diverse set of contexts. In practice, for some types of sequences and contexts, one may obtain better (higher) predictability values by ignoring contextual information and focusing only on the history of visited locations.

Taking a step further, we also conjecture that our *sequence-merging approach could be used as a test to determine if a given contextual information can be useful for prediction.* Suppose one wishes to use a given type of context when predicting an individual's locations. Before performing the prediction itself, one may run our technique and check whether context reduces the estimation of entropy. If so, there is enough information in the context to possibly help prediction. Otherwise, the size of the sample (length of sequences X and C) may not be large enough for context to be useful for prediction, therefore one may be better off not using it. Investigating how to translate this general idea into a practical solution is an interesting avenue we intend to pursue in the future.

7 Conclusions and Future Work

In this paper, we analyzed the state-of-the-art technique to compute the limits of predictability in human mobility. We argued that, despite its extensive use in the literature, this technique has two shortcomings, namely low interpretability and lack of contextual information, that we began to address in previous work and continued in this study.

To tackle the first issue, we used two metrics (regularity and stationarity) that explain most of the variability in the entropy of mobility traces in both next-cell and next-place prediction. We built a regression model that uses these two metrics as a proxy to explain entropy values. Our simple model was able to explain 76% of the variability in the entropy for the GPS dataset and 94% for the CDR dataset in the next-cell prediction task, and 68% and 70%, respectively, in the next-place prediction task. These good model fittings suggest that both regularity and stationarity – which are much simpler and intuitive – can indeed be used to understand how predictability the mobility of a person actually is.

To address the second issue, we quantified the benefits of introducing different types of contextual information into predictability estimate, for different entropy estimators. Our results revealed that context can indeed greatly reduce entropy (thus improving predictability) of probability-based estimators. Yet, improving the entropy estimate produced by the compression-based method used by Song et al. proved to be quite a challenge. We proposed two strategies to incorporate context to that entropy estimator: sequence-splitting and sequence-merging. Though simpler, sequence-splitting suffers from the fact that, by dividing the original sequence into smaller ones, conditioned by the same context, the final entropy estimate with context may indeed be higher than the original one (without context). Our sequence-merging strategy considers the whole sequence, but suffers from a different, yet related, problem: the fact that, due to an increased alphabet, it takes longer to converge and sometimes produces entropy values with context that are higher than those without it. Our discussion of these results are grounded by a more fundamental understanding of how this state-of-the-art entropy estimator works, revealing that, unlike argued by previous studies, exploring context into predictability estimate may not always be beneficial in practical scenarios. In that direction, we also argue that the sequence-merging approach could be used to identify types of contextual information that may increase the accuracy of mobility prediction models.

As future work, we intend to investigate efficient alternative algorithms to compute the entropy of a sequence using Song et al.'s estimator and explore it as a measure of confidence in prediction results. For example, predictions produced for a person with higher predictability can be considered more reliable than predictions performed for a very unpredictable individual. As such, a person's predictability estimate could be exploited by several services (e.g., recommendation, content distribution) that makes use of location prediction to drive their decisions. Another direction for future exploration involves investigating the effectiveness of specific prediction algorithms in light of predictability estimates, aiming to understand when a method can or can not reach the predictability estimate associated with a given mobility dataset. Such knowledge can drive the design of more effective prediction methods.

References

[1] K. Zhao, D. Khryashchev, J. Freire, C. Silva, and H. Vo. Predicting taxi demand at high spatial resolution: Approaching the limit of predictability. In *Proc. IEEE International Conference on Big Data*, 2016.

[2] X. Zhou, Z. Zhao, R. Li, Yifan Zhou, and H. Zhang. The predictability of cellular networks traffic. In *2012 International Symposium on Communications and Information Technologies (ISCIT)*, pages 973–978, 2012.

[3] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327(5968), 2010.

[4] G. Ding, J. Wang, Q. Wu, Y. Yao, R. Li, H. Zhang, and Y. Zou. On the limits of predictability in real-world radio spectrum state dynamics: from entropy theory to 5g spectrum sharing. *IEEE Communications Magazine*, 53(7), 2015.

[5] James P. Bagrow, Xipei Liu, and Lewis Mitchell. Information flow reveals prediction limits in online social activity. *Nature Human Behaviour*, 3(2):122–128, 2019.

[6] Xin Lu, Erik Wetter, Nita Bharti, Andrew J. Tatem, and Linus Bengtsson. Approaching the limit of predictability in human mobility. *Scientific Reports*, 2013.

[7] Gavin Smith, Romain Wieser, James Goulding, and Duncan Barrack. A refined limit on the predictability of human mobility. In *2014 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 88–94. IEEE, 2014.

[8] Andrea Cuttone, Sune Lehmann, and Marta C. González. Understanding predictability and exploration in human mobility. *EPJ Data Science*, 2018.

[9] Edin Lind Ikanovic and Anders Mollgaard. An alternative approach to the limits of predictability in human mobility. *EPJ Data Science*, 6(1):12, Jun 2017.

[10] Douglas do Couto Teixeira, Aline Carneiro Viana, Mário S. Alvim, and Jussara M. Almeida. Deciphering predictability limits in human mobility. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '19, pages 52–61, New York, NY, USA, 2019. ACM.

[11] Aarti Munjal, Tracy Camp, and William C. Navidi. Smooth: A simple way to model human mobility. In *Proc. 14th ACM Int. Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, 2011.

[12] Lucas Maia Silveira, Jussara M. Almeida, Humberto Torres Marques-Neto, Carlos Sarraute, and Artur Ziviani. Mobhet: Predicting human mobility using heterogeneous data sources. *Computer Communications*, 95, 2016.

[13] Yu Zheng, Licia Capra, Ouri Wolfson, and Hai Yang. Urban computing: Concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology*, 5, 2014.

[14] Wei Dong, Nick Duffield, Zihui Ge, Seungjoon Lee, and Jeffrey Pang. Modeling cellular user mobility using a leap graph. In *Proc. 14th International Conference on Passive and Active Measurement*, 2013.

[15] Andrea Hess, Karin Anna Hummel, Wilfried N. Gansterer, and Günter Haring. Data-driven human mobility modeling: A survey and engineering guidance for mobile networking. *ACM Computing Surveys*, 48(3), 2016.

[16] Joanne Treurniet. A taxonomy and survey of microscopic mobility models from the mobile networking domain. *ACM Computing Surveys*, 2014.

[17] Gordon Moon and Jihun Hamm. A large-scale study in predictability of daily activities and places. In *Proceedings of the 8th EAI International Conference on Mobile Computing, Applications and Services*, MobiCASE'16, pages 86–97, 2016.

[18] Juan C. Herrera, Daniel B. Work, Ryan Herring, Xuegang Ban, Quinn Jacobson, and Alexandre M. Bayen. Evaluation of traffic data obtained via gps-enabled mobile phones: The mobile century field experiment. *Transportation Research Part C: Emerging Technologies*, 18(4), 2010.

[19] Samiul Hasan, Xianyuan Zhan, and Satish V. Ukkusuri. Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. In *International Workshop on Urban Computing*, 2013.

[20] Mariano G. Beiró, André Panisson, Michele Tizzoni, and Ciro Cattuto. Predicting human mobility through the assimilation of social media traces into mobility models. *EPJ Data Science*, 5(1), 2016.

[21] Douglas Teixeira, Mário Alvim, and Jussara Almeida. On the predictability of a user's next check-in using data from different social networks. In *Proceedings of the 2Nd ACM SIGSPATIAL Workshop on Prediction of Human Mobility*, PredictGIS 2018, pages 8–14, 2019.

[22] Vaibhav Kulkarni, Abhijit Mahalunkar, Benoit Garbinato, and John D. Kelleher. Examining the limits of predictability of human mobility. *Entropy*, 2019.

[23] Paiheng Xu, Likang Yin, Zhongtao Yue, and Tao Zhou. On predictability of time series. *Physica A: Statistical Mechanics and its Applications*, 523:345 – 351, 2019.

[24] Meir Feder, Neri Merhav, and Michael Gutman. Universal prediction of individual sequences. *IEEE transactions on Information Theory*, 38(4):1258–1270, 1992.

[25] Ming Li and Paul M. B. Vitányi. Handbook of theoretical computer science (vol. a). chapter Kolmogorov Complexity and Its Applications, pages 187–254. MIT Press, Cambridge, MA, USA, 1990.

[26] Thomas M Cover and Joy A Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012.

[27] Jean Hausser and Korbinian Strimmer. Entropy inference and the james-stein estimator, with application to nonlinear gene association networks. *J. Mach. Learn. Res.*, 10:1469–1484, December 2009.

[28] Jean Hausser, Korbinian Strimmer, and Maintainer Korbinian Strimmer. Package 'entropy'.

[29] Claude E Shannon and Warren Weaver. *The mathematical theory of communication*. University of Illinois press, 1998.

[30] I. Kontoyiannis, P. H. Algoet, Yu. M. Suhov, and A. J. Wyner. Nonparametric entropy estimation for stationary processes and random fields, with applications to english text. *IEEE Transactions on Information Theory*, 2006.

[31] A. Lempel and J. Ziv. On the complexity of finite sequences. *IEEE Trans. Inf. Theor.*, 22(1):75–81, September 2006.

[32] Vinicius Monteiro de Lira, Salvatore Rinzivillo, Chiara Renso, Valeria Cesario Times, and Patricia Cabral Tedesco. Investigating semantic regularity of human mobility lifestyle. In *Proceedings of the 18th International Database Engineering & Applications Symposium*, IDEAS '14, pages 314–317, 2014.

[33] Charles Spearman. The proof and measurement of association between two things. 1961.

[34] Guangshuo Chen, Aline Carneiro Viana, Marco Fiore, and Carlos Sarraute. Complete Trajectory Reconstruction from Sparse Mobile Phone Data. *EPJ Data Science*, October 2019.

[35] AG Carlton. On the bias of information estimates. *Psychological Bulletin*, 71(2):108, 1969.

[36] Alan Agresti and David B. Hitchcock. Bayesian inference for categorical data analysis. *Statistical Methods & Applications*, 14(3):297–330, 2005.