Deciphering Predictability Limits in Human Mobility

Douglas Teixeira, Aline Carneiro Viana, Mário Alvim, Jussara Almeida

▶ To cite this version:

Douglas Teixeira, Aline Carneiro Viana, Mário Alvim, Jussara Almeida. Deciphering Predictability Limits in Human Mobility. SIGSPATIAL 2019 - 7th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Nov 2019, Chicago, United States. 10.1145/3347146.3359093 . hal-02286128

HAL Id: hal-02286128

https://hal.inria.fr/hal-02286128

Submitted on 23 Sep 2019

DECIPHERING PREDICTABILITY LIMITS IN HUMAN MOBILITY

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

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

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

September 23, 2019

ABSTRACT

Human mobility has been studied from different perspectives. One approach addresses *predictability*, deriving theoretical limits on the accuracy that any prediction model can achieve in a given dataset. This approach focuses on the inherent nature and fundamental patterns of human behavior captured in the dataset, filtering out factors that depend on the specificities of the prediction method adopted. In this paper, we revisit the state-of-the-art method for estimating the predictability of a person's mobility, which, despite being widely adopted, suffers from low interpretability and disregards external factors that have been suggested to improve predictability estimation, notably the use of contextual information (e.g., weather, day of the week, and time of the day). We also conduct a thorough analysis of how this widely used method works, by looking into two different measures (one proposed by us) which are easier to understand and, as shown, capture reasonably well the effects of the original technique. Additionally, we investigate strategies to incorporate different types of contextual information into predictability estimates, and show that the benefits vary depending on the underlying prediction task. Finally, we propose and evaluate alternative estimates of predictability which, while being much easier to interpret, provide comparable results to the state-of-the-art.

***K*eywords** human mobility · predictability · entropy

1 Introduction

Measuring the *predictability* of any phenomenon is a very useful, but hard task, and especially so in the case of human behavior. Such complexity is due to the uncertain and heterogeneous behavior of humans, as well as to the variability of parameters influencing such behavior. Predictability is concerned with the maximum theoretical accuracy that an ideal prediction model could achieve in a scenario expressed by a given dataset. As such, unlike particular comparisons of alternative prediction models on different datasets, it does not depend on a specific prediction strategy but rather on human behavior, as captured by the available data. Besides, it does not rely on the tuning of a multitude of sensible parameters, providing instead a parameter-free view of how predictable human mobility can be (as expressed in the data).

Studying predictability has thus the advantage of laying out a common ground (i.e., an upper-bound) to which particular prediction techniques can be compared. It is also useful in other practical scenarios. For instance, for a particular dataset, users who exhibit levels of predictability very different from the rest are likely to be outliers and may deserve special attention. In a recommendation scenario, for example, users with low predictability are possibly those who are more open to novelty, diversity and serendipity. We emphasize that by studying *predictability*, as opposed to particular prediction techniques, we focus on the inherent nature and fundamental patterns of human behavior, filtering out factors that depend on the particular prediction method adopted.

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

The state-of-the-art technique to measure predictability in human mobility was proposed by Song et al. [3] and has been used by several research communities [1, 2, 4, 5, 6]. In a nutshell, it exploits the concept of entropy to estimate how complex (or, inversely, how predictable) a person's mobility patterns are.

Despite its wide use and applicability, this technique has two shortcomings, which we wish to address in this paper. 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. Previous work [7, 8] show that predictability of mobility, as captured by Song et al.'s technique, is strongly related to stationary patterns. Yet, as we show later, stationarity alone does not explain predictability. We here propose another measure, called *regularity* that, together with *stationarity*, helps us understand *when* and *how* the predictability of a person's mobility patterns varies. [1]

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 can influence human mobility. For instance, previous work [9, 8] has suggested that contextual information (e.g., weather, time of the day, etc) might be useful for estimating predictability in human mobility. These prior studies did not mention, however, *how* to use such types of information to improve predictability. As we argue in Section 5.1, this may be quite hard, especially for Song et al.'s method. Furthermore, to the best of our knowledge, no previous work has quantified the impact of such information on predictability estimation. Our work aims at filling this gap by showing how to use contextual information to compute the limits of predictability, and how to quantify the impact of this kind of information on predictability.

Moreover, as we will show later, the predictability of one's mobility depends on the underlying prediction task. Different tasks can be defined depending on the particular goal, and the importance of introducing contextual information to estimating predictability may vary depending on the specific task. We here consider two prediction tasks, namely *next-cell prediction* and *next-place prediction*, previously defined by Cuttone et al. [8].

In both of these 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, given the person's previously visited cells [3, 8], which could be the same cell the person is currently in. In the next place prediction problem, the goal is to guess the next *distinct* place the person will visit [8]. Notice that this prediction task is concerned only with transitions between different places, which eliminates stationarity (i.e., self-transitions), making the prediction considerably harder. To our knowledge, although prior studies have analyzed predictability under these two prediction tasks [3, 8, 6, 7], none has investigated the effectiveness of introducing contextual information to the predictability estimates.

Our study covers the aforementioned two prediction tasks and relies on the analysis of two datasets of different spatial and temporal granularities, as these properties may influence predictability. In sum, we make the following contributions:

- *Detailed investigation of the state-of-the-art estimator of mobility predictability in next-cell prediction problem.* We propose a new measure, *regularity*, which together with *stationarity* [8], helps us understand what makes a person's mobility trajectory more or less predictable, as captured by Song et al.'s technique. We show that these two simple measures are complementary and 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 one's mobility predictability varies.
- *Evaluation of the benefits of contextual information on predictability for both next-cell and next place prediction.* We are the first to quantify the impact of different types of contextual information on predictability in human mobility, for different prediction tasks and datasets. Our results show that, for the next place prediction problem, the use of contextual information plays a larger role than one's history of visited locations in estimating their predictability.

[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.

- *Proposal of alternative estimates of mobility predictability for both next-cell and next place prediction.* Given the difficulties in using Song et al.'s technique with contextual information and because of its low interpretability, we propose the use of different entropy estimators that are simpler and easily allow for the use of contextual information. We show that these estimators, while being more interpretable, provide comparable results in terms of predictability.

The rest of this paper is organized as follows. Section 2 discusses related work, while Section 3 formally defines our target problem and the scope of our study. We revisit the state-of-the-art method proposed by Song et al. in Section 4, offering a thorough analysis and insights on how it works and its robustness to different scenarios. We then examine the importance of introducing contextual information in predictability estimation, considering different estimates of predictability, in Section 5. A recap of our results and our final remarks are presented in Section 6.

2 Related Work

The study of human mobility has received considerable attention in the literature [6, 10, 11, 12, 13, 14, 15, 9, 7, 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 [16], and so on) and using diverse types of data sources (Call detail records (CDRs) [6, 3], GPS traces [17, 8], and social media data [18, 19], 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 [9, 7, 6]. 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 any 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 4.3, 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, 9, 20], 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 [21] and to its mathematical details [22]. Other studies computed the limits of predictability for several temporal and spatial resolutions [9], but they did not try to explain what (besides temporal and spatial granularity) may impact predictability. Previous work has also computed the limits of predictability for the two tasks that we study here [7, 8], but they did not not show how context affects predictability in the proposed tasks.

Despite such great attention, the method proposed by Song et al. has low interpretability, as it is based on a compression algorithm whose output bears little resemblance to its input, most of the time. Thus, it is hard to keep track of what the algorithms is really capturing in terms of mobility patterns. Moreover, as we will argue in Section 5, it is quite challenging to extend it to incorporate contextual information (e.g., weather, time of the day), which is a desirable feature to improve predictability estimates, as suggested by previous studies [9, 8]. Indeed, we are aware of only one very recent attempt to do so, though in a different domain.

In it, Bagrow et al. [5] estimated how much knowing the contents of the tweets of a person's friends helps in predicting the contents of this person's tweets. In the mobility scenario, this problem would be the equivalent to measuring how much knowing the location history of a person's friends helps in predicting that person's next location. While well suited to their domain, Bagrow et al.'s technique does not generalize to our situation. In their case, both the supporting data sequences (the tweets of a person's friends) and the target sequence (the person's tweets) are taken from the same alphabet. In our case, the supporting data sequences (contextual information) have a different alphabet than the target sequence (a person's visited locations), therefore making it hard to use their technique in our scenario. We are investigating ways of changing such technique to suite our needs, but that remains as future work.

Subsequent papers [6, 1, 9, 8], apply Song's technique to the next-cell prediction problem, reporting high values for the maximum predictability. Recently, other studies analyzed the predictability of the next place prediction problem [8, 7]. Cuttone et al. [8] point out that the removal of stationary patterns correspond to lower predictability. They also suggest that other types of information could be used to enhance predictability, but they do not mention how to do so.

Our work is complementary to and greatly contributes over all previous studies in several aspects. First, we show that stationarity patterns alone do not explain predictability. We propose another measure, called *regularity* that, together with *stationarity*, helps us understand when and how the predictability of a person's mobility patterns varies. Second, we show that it is possible, under certain circumstances, to add contextual information to predictability estimates. Third, we show that, depending on the prediction task, contextual information may play an even more important role than the history of visited locations in determining a person's predictability. Fourth, we propose the use of simpler entropy estimators that allow for the use of contextual information in a natural manner, unlike Song's technique.

3 Problem Definition

In this section, we formally define our target problem—predictability in human mobility— and present the particular scenarios defining the scope of our study.

3.1 Predictability in Human Mobility

We start with an alphabet $\mathcal{S}$, which in our case consists of locations in a target area, and a time-ordered input sequence of symbols $X = (x_1, x_2, \ldots, x_{n-1})$ where $x_i \in \mathcal{S}$ is the i-th location visited the user. We also assume there exists a prediction algorithm $\mathcal{A}$ to estimate a probability distribution given by $P(x_n = s \mid x_1, x_2, \ldots, x_{n-1})$ for each $s \in \mathcal{S}$. Our task, then, is to derive the predictability Π_{max} (maximum possible accuracy) that any algorithm $\mathcal{A}$ could achieve in X when trying to guess X_n, i.e., $P(x_n = s \mid x_1, x_2, \ldots, x_{n-1})$.

Note that the formulation of Π_{max} depends on a particular prediction task under study (e.g., specific properties of the next symbol in the sequence). Moreover, Π_{max} is computed based solely on the data from which the input sequence 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 underlying prediction task and the datasets are factors delimiting the scope of the predictability study. Next, we present how we capture these factors in our investigation.

3.2 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 in X. 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}.

The original prediction task 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.

3.3 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.

GPS Dataset: The first dataset is a high temporal and spatial resolution dataset consisting of GPS traces. This dataset was obtained through an Android mobile phone application, called MACACOApp[2]. 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.

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

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

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 300 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 200 meters.

4 Estimating Predictability: Background and Analyses

In this section, we offer a broad discussion of how to estimate predictability in human mobility. We review the literature trying to connect Song's technique to more fundamental theoretical concepts and well-established measures of complexity. To the best of our knowledge, no previous work summarized the roots of predictability going back to Kolmogorov Complexity, tracing the equivalences between entropy and compressibility and showing why entropy is a good approximation to the complexity of a sequence of symbols, as we do here.

We also present two measures that capture key aspects of mobility patterns and show that they can effectively be used as proxies to understand the predictability estimate. Finally, using these two measures, we analyze when and how the predictability estimate proposed by Song et al. change with respect to the spatial and temporal granularity of the dataset.

4.1 Estimating Complexity

The predictability of a sequence of symbols is intimately related to its complexity (randomness). In this sense, more complex sequences are harder to predict. The complexity of a sequence is also related to how *compressible* it is [23]. Random sequences are less compressible, and highly predictable sequences are highly compressible. The rationale 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 [24] is a general measure of the complexity of a sequence, and 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 [25] that these differences are bound by a fixed, additive constant.

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 small program. Thus, because π has an infinite number of digits but can be generated by a small program, the sequence of digits of π is highly compressible (has constant Kolmogorov Complexity).

If a sequence is completely predictable (deterministic), there exists a program that generates it and terminates, such as in the case of the constant π. If, however, the input sequence is produced by a non-deterministic process, there may not be a program that completely generates it. 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. In cases such as this, it is common to make a simplification to the problem and assume that the symbols of the sequence are generated by a stochastic process. For sequences generated by this type of process, the entropy is a good approximation of the sequence's complexity because the entropy of a sequence is a lower bound on its compressibility [25].

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 the true predictability would require the true Kolmogorov Complexity of the sequence, which is not computable.

4.2 Entropy as a Measure of Complexity

Song et al. [3] proposed a technique to estimate the predictability of a sequence of locations based on the entropy of the sequence. Their work established limits on the predictability of a sequence of locations that a person visited based on three estimates of the entropy of the sequence. The first estimate is the Shannon entropy [26] 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 the entropy using a distribution that accounts for both the frequency of visitations as well as temporal patterns. The third estimator, described by Kontoyiannis et al. [27], is related to the Lempel-Ziv compression algorithm [28] and to the Lempel-Ziv measure of the complexity of a sequence [28]. According to this estimator, the entropy H of an input sequence of locations L can be approximated by:

$$H \approx \left(\frac{1}{n} \sum_{i \in L} \Lambda_i \right)^{-1} \log_2(n), \tag{1}$$

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

For ergodic, stationary processes, this estimator is said to converge to the entropy rate of the source as the size of the input goes to infinity [25]. This estimator does not require 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 basic formula to compute the maximum predictability takes as input an entropy value S and the number N of unique locations, and uses Fano's Inequality [25] to get an upper bound on predictability, as follows:

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

where $H(\Pi_{max})$ is given by

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

A proof that these equations estimate the correct limits of predictability can be found in related work [3, 9, 1]. In particular, Smith et al. [9] 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*.

4.3 Understanding Predictability

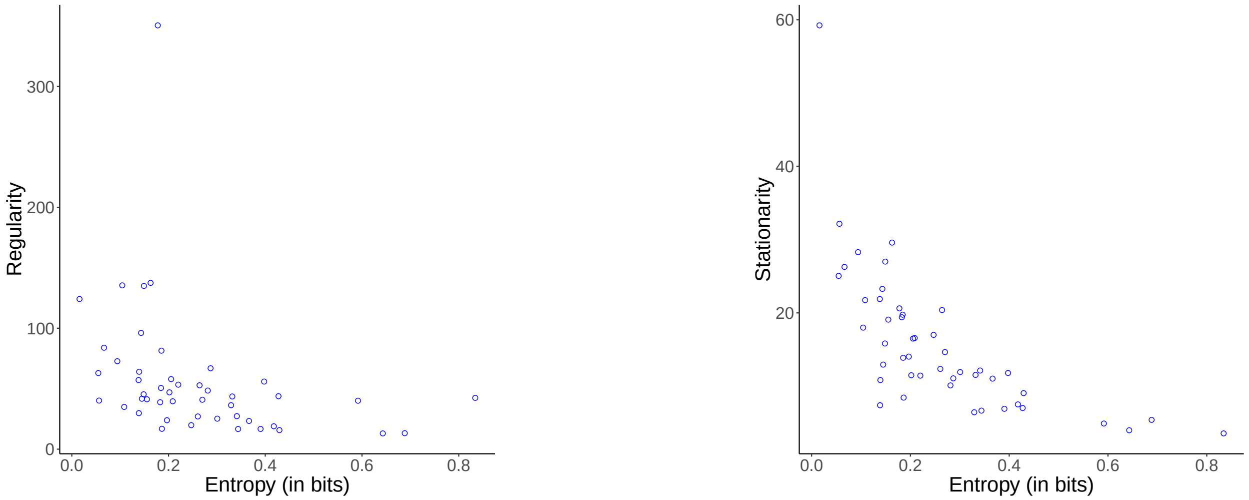
As we mentioned before, Song et al.'s technique has been extensively used to assess the predictability of mobility datasets. However, to our knowledge, no previous work has analyzed how it works in terms of *why* and *how* the predictability varies across different users and datasets. This is our goal in this section and in the next one. To that end, we focus on the same underlying prediction task as Song et al. [3], i.e., next-cell prediction, investigating the entropy estimate across different users on both GPS and CDR datasets.

We start by arguing that analyzing the entropy estimate itself, particularly the more precise one based on compressibility, is quite challenging as the result of the method is hard to interpret. Thus, we look for simpler and easier to understand proxy measures, which can be used in its place to understand predictability in human mobility. Specifically, we employ two *simple* measures 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 measure, called the *stationarity* of a sequence of locations, is the mean number of observations for which the person stays continuously in the same location. For instance, the stationarity of sequence $L = (1, 2, 3, 4, 1, 2, 3, 4)$ is given by $st(L) = (1 \times 8)/8 = 1$, meaning that, on average, the person stays in the same location for one observation.

Yet, *stationarity* alone does not explain predictability. Consider, for instance, two input sequences $L_1 = (1, 2, 3, 4, 1, 2, 3, 4)$ and $L_2 = (1, 2, 1, 2, 1, 2, 1, 2)$. Both have the same size and the same stationarity, but L_2 has lower entropy than L_1.

Thus, we introduce another measure, called *regularity*, that helps explain the entropy of a person's observed location history. The *regularity* of a sequence is the ratio between the length of the sequence and the number of *unique* symbols in it. For instance, the regularity of input sequence $L = (1, 1, 1, 1, 2, 2, 3, 3, 4, 4)$ is given by $reg(L) = 10/4$. If we compute the *regularity* of the two aforementioned example sequences (L_1 and L_2), we obtain $reg(L_1) = 8/4 = 2$ and $reg(L_2) = 8/2 = 4$, which helps explain why L_2 has lower entropy than L_1 (L_2 is more regular than L_1).



(a) Regularity

(b) Stationarity

Figure 1: Entropy (in bits per symbol) as a function of regularity and stationarity for the GPS dataset.

We note that the choice of regularity and stationarity was based on experimental observations of how Song et al.'s technique works. Intuitively, they capture two key and complementary components in a person's mobility patterns: the ratio between previously visited places and new places, and the amount of time spent in each place. Although the importance of stationarity to predictability has been noted before [8], using regularity to help understand predictability and thoroughly evaluating both measures is a contribution of our work. As discussed next, both measures complement each other.

We further illustrate the relationship between entropy and regularity/stationarity, by showing in Figure 1 scatter plots of these measures computed for all users in the GPS dataset (each dot is a user). The measures shown on each plot were computed considering each user's complete history of visited locations in the dataset. Similar results were also obtained for the CDR dataset (omitted). As the figure shows, entropy drops as either regularity or stationarity increases, although the relationships are not linear. Indeed we found a Spearman correlation ρ between entropy and regularity equal to -0.84 on both datasets, and a Spearman correlation between entropy and stationarity equal to -0.79 and -0.83 in the GPS and CDR datasets, respectively. Thus, both measures are strongly correlated with entropy.

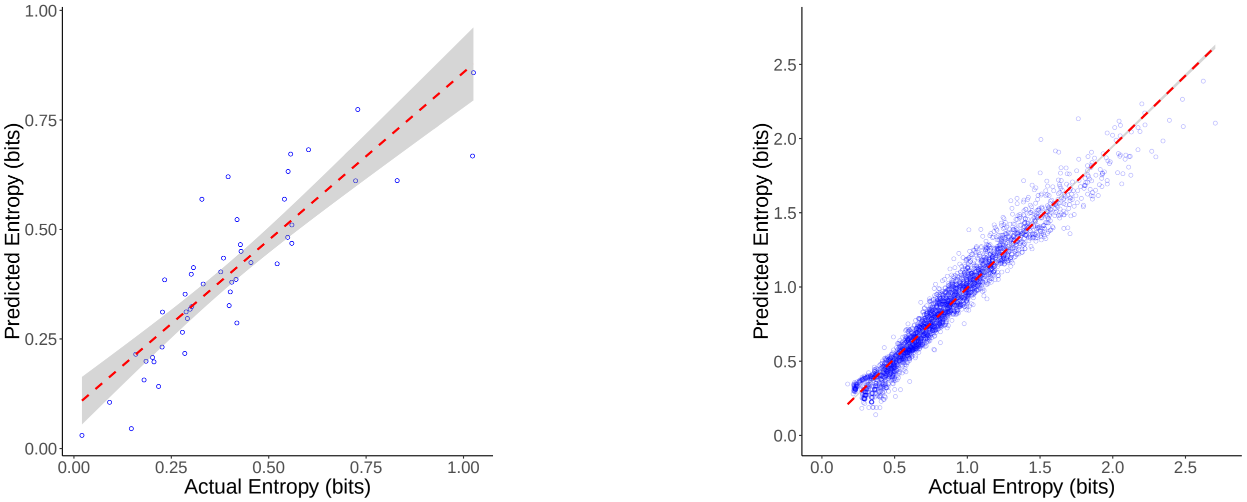
Moreover, though these two measures are themselves reasonably well correlated (0.63 and 0.48 in the GPS and CDR datasets, respectively), such relationship is far from perfect. Indeed, in both charts of Figure 1, there are several users who, despite having similar regularity (or stationarity) have very different entropy values, indicating that each variable, in isolation, cannot explain entropy. We did look into several such cases and found that large differences in entropy for users with similar regularity could often be explained by great differences in stationarity, and vice versa. These observations suggest the two measures are, to some extent, complementary.

Next, we analyzed the extent to which *only* these two measures, when used jointly, can reasonably explain the predictability of a sequence of locations. To that end, we employed a regression analysis by fitting the entropy $H(L)$ of a sequence L as a function of regularity $reg(L)$ and stationarity $st(L)$. We experimented with different regression functions and the one that led to the best fitted model is:

$$H(L) = \alpha \log(reg(L)) + \beta \log(st(L)) + \gamma(\log(reg(L)) \times \log(st(L))) + \epsilon,$$

where α, β, and γ are the coefficients of regression and ϵ is the regression error.

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 fitting. Given the non-linear relationship between H and each measure, shown in Figure 1, we made a transformation in the data, taking the logarithm of reg and st. Furthermore, it was necessary to consider the interaction between the metrics because there is a confounding effect between them. This model, albeit simple, is able to explain a large fraction of the total variation in the entropy values ($R^2 = 0.76$ for the GPS dataset and $R^2 = 0.94$ for the CDR dataset). The model also performed well for other spatial resolutions. For instance, for the GPS dataset, the R^2 varied from 0.76 (higest spatial resolution, smaller cells) to 0.83 (lowest spatial resolution, larger cells).



(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). The red, dashed line shows the regression model 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.

Such good fittings become clear in Figure 2, which shows predicted versus actual entropy values for both datasets. Note 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.

4.4 Spatiotemporal interplay

Previous studies [9, 8, 29] 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 1 shows how average regularity and average stationarity vary as temporal resolution varies in our GPS dataset. 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 the regularity is the ratio between the size of the sequence and the number of unique symbols. Therefore, a reduction in the numerator (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 1.

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 2. The entropy will decrease 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 2, we show how entropy decreases as regularity increases, both on average.

Temporal Resolution (mins)	Average Regularity	Average Stationarity	Average Entropy
5	80.72	41.28	0.40
10	57.89	30.79	0.50
20	38.24	19.49	0.66
30	30.75	14.68	0.80
40	25.75	12.48	0.90
50	22.79	10.76	0.98
60	20.27	9.68	1.05

Table 1: Average regularity, stationarity and entropy (in bits per symbol) as functions of the temporal resolution of the data (GPS dataset).

Spatial Resolution (meters)	Average Regularity	Average Stationarity	Average Entropy
300	129.76	41.65	0.404
400	140.37	71.16	0.389
500	165.66	90.04	0.379
600	168.60	76.66	0.351
700	184.58	94.82	0.285
800	194.21	96.32	0.286
900	202.66	96.29	0.284
1000	210.40	95.69	0.264

Table 2: Average regularity, stationarity and entropy (in bits per symbol) as functions of the spatial resolution of the data (GPS dataset).

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.

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 their whole mobility trace. We argue that regularity and stationarity are examples of such metrics.

Open remark: 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.

5 Context and Predictability

In this section, we investigate the role of contextual information in the predictability of human mobility. We do that in two ways. First, we show that context helps in predictability, in the sense that lower entropy values are obtained when context is used. Second, we explain how to use context with other entropy estimators, showing that it could be incorporated into predictability measures.

Recall from Section 4.3 that the maximum predictability is a function of the entropy of the sequence. Thus, we experiment with different entropy estimators and show that by using contextual information we are able to obtain higher values for the maximum predictability of a dataset. What we would like to do is, given $X = (x_1, x_2, \ldots, x_{n-1})$, a sequence of locations visited by a person, and $Y = (y_1, y_2, \ldots, y_{n-1})$, a sequence of contextual information associated with each of the visits (y_i could be the weather when the person visited location y_i, for instance), we wish to measure how much knowing Y helps in estimating the entropy of X, which can be computed as follows:

$$H(X \mid Y) = H(X) - I(X, Y), \tag{3}$$

where $I(X, Y)$ is the Mutual Information [25] between X and Y, that is, $I(X, Y)$ tells us how much information Y carries about X, and is given by:

$$I(X, Y) = \sum_{x \in X} \sum_{y \in Y} P(x, y) \log \frac{P(x, y)}{P(x)P(y)}. \tag{4}$$

5.1 Context and the Lempel-Ziv Estimator

The difficulty in using contextual information with Song's technique stems from the inner-workings of the algorithm it uses to estimate the entropy of the sequence. Recall that, in order to estimate the conditional entropy $H(X \mid Y)$ between two sequences X and Y, we have to know the underlying conditional probabilities of the symbols in the two sequences. The algorithm used by Song et al. works by compressing the input sequence of symbols to estimate its entropy, therefore leveraging the relation between entropy and compressibility. However, in doing so the algorithm becomes oblivious to the underlying probability distribution of the symbols of the sequence, which poses a barrier to computing the conditional entropy, as it depends on the probability distribution of the symbols. As this is the state-of-the-art entropy estimator for computing the entropy in mobility studies and, as shown in Table 3, performs well when no context is used, we decided to work with its limitations and include it in the comparison with other estimators.

One approach to try to compute the conditional entropy of two sequences $X = (x_1, x_2, \ldots, x_{n-1})$ and $Y = (y_1, y_2, \ldots, y_{n-1})$ using a compression algorithm would be to create a sequence XY of the form $XY = ((x_1, y_1), (x_2, y_2), \ldots, (x_{n-1}, y_{n-1}))$ and compress this sequence. However, by doing so we would be estimating the (joint) entropy of *both* X and Y, and not the entropy of X *given* Y. The other entropy estimators, described in Section 5.2 work with the probability distributions of the symbols in the sequences and therefore do not suffer from this limitation.

In order to circumvent this limitation, we have to create a procedure to hard-code contextual information into our estimates of the entropy when using the Lempel-Ziv estimator. We will illustrate this procedure with an example. Suppose we want to use the hour of the day as contextual information. To do that, we create 24 time series, one for each hour of the day, each of which containing all of the locations visited in that hour in the past. We then run the entropy estimation algorithm in each of the 24 time series, averaging the results accordingly. All of our results for the Lempel-Ziv estimator were obtained using this procedure.

5.2 Context and Other Entropy Estimators

Given the theoretical framework presented in the last section, we evaluated several entropy estimators to measure the impact of three types of contextual information (day of the week, hour of the day, and weather) in their entropy estimates. The first estimator used is called *Maximum Likelihood* (ML), which estimates the entropy using the empirical frequencies of observations, and therefore is equivalent to Shannon entropy. The second estimator, called *Miller-Madow* (MM), estimates the entropy by applying the Miller-Madow bias correction to Shannon entropy. The third, called *SG*, estimates the entropy using the Dirichlet multinomial pseudo-count model with parameter $a = 1/length(X)$. The latter gives the best results among three other estimators that also use a Dirichlet process.

As explained in the previous sections, we compute the entropy of each estimator by considering X and Y, the sequence of visited locations, and contexts associated to each visit, respectively. Then, we compute the mutual information using the same estimator used to compute the entropy, and finally we use Equation 3 to compute the entropy given the mutual information between the variables. Each piece of contextual information was discretized so as to obtain an identifier for each type of information (rain, snow, clouds, etc., in the case of weather), which was then used in the computation of the mutual information.

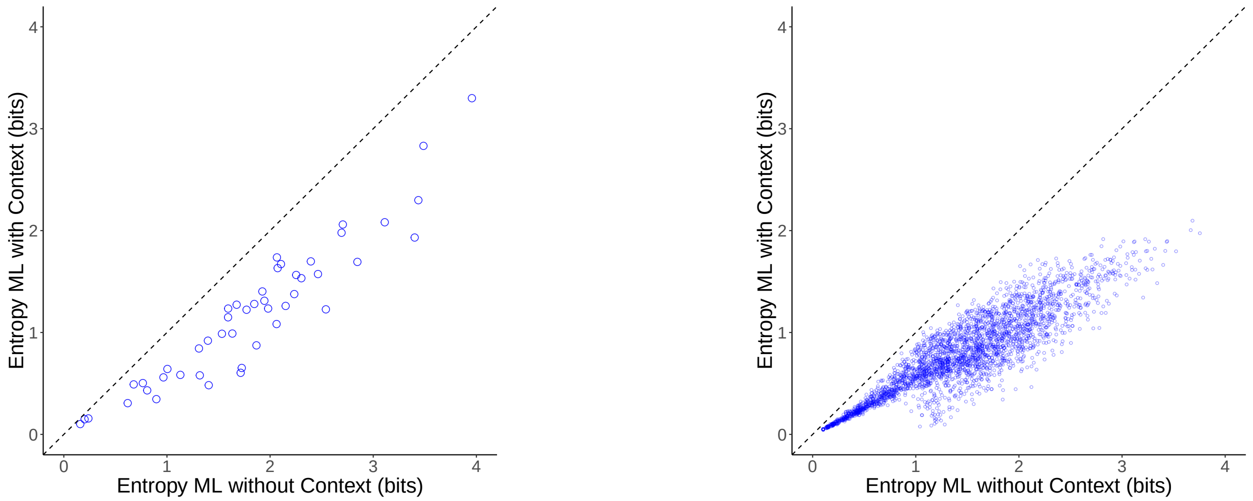
		GPS dataset			CDR dataset			
		No Context	**Weekday**	**Hour**	**No Context**	**Weekday**	**Hour**	**Weather**
Next-Cell	**Maximum Likelihood**	6.01	1.48	1.18	1.45	1.13	0.79	1.23
	Miller-Madow	10.6	1.61	1.22	7.86	1.20	0.91	1.32
	SG	4.62	1.48	1.19	2.66	1.17	0.84	1.27
	Lempel-Ziv	0.40	0.39	0.72	1.01	1.34	1.38	1.30
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
	Lempel-Ziv	3.11	2.80	2.14	1.94	1.50	0.64	1.21

Table 3: Evaluation of four entropy estimators in both datasets and for two prediction tasks (next-cell and next place). The reported entropy values are given in bits per symbol (each location is a symbol in the input sequence).

The basis for entropy computation is an underlying probability distribution. Thus, if one has the *full* probability distribution of a sequence of symbols, the entropy is given by Shannon's formula: $H = -\sum p(j) \log_2 p(j)$. In real world situations, however, one usually has access to only a *sample* of the true underlying probability distribution. Therefore, entropy estimates have to deal with the fact that there may be fluctuations in the probabilities due to the sampling process. With the exception of the Maximum Likelihood and the Lempel-Ziv estimator, the other two estimators that we used try to compensate for these fluctuations by adding a bias term to the entropy computation. Because of this bias term, their entropy estimates are more conservative.

5.3 Context and Next-Cell Prediction

In this section, we show that contextual information is useful for predictability and evaluate the entropy estimators previously discussed. To illustrate the usefulness of contextual information, Figure 3 shows that by using temporal information (hour of the day) with the *Maximum Likelihood* estimator, we are able to reduce the entropy by 45% in the next-cell prediction problem. These results confirm previous claims [8, 9] that external information is helpful to assess the predictability of a dataset.



(a) GPS dataset

(b) CDR dataset

Figure 3: Reduction in the entropy values when contextual information (hour of the day in this case) is used, in the next-cell prediction problem.

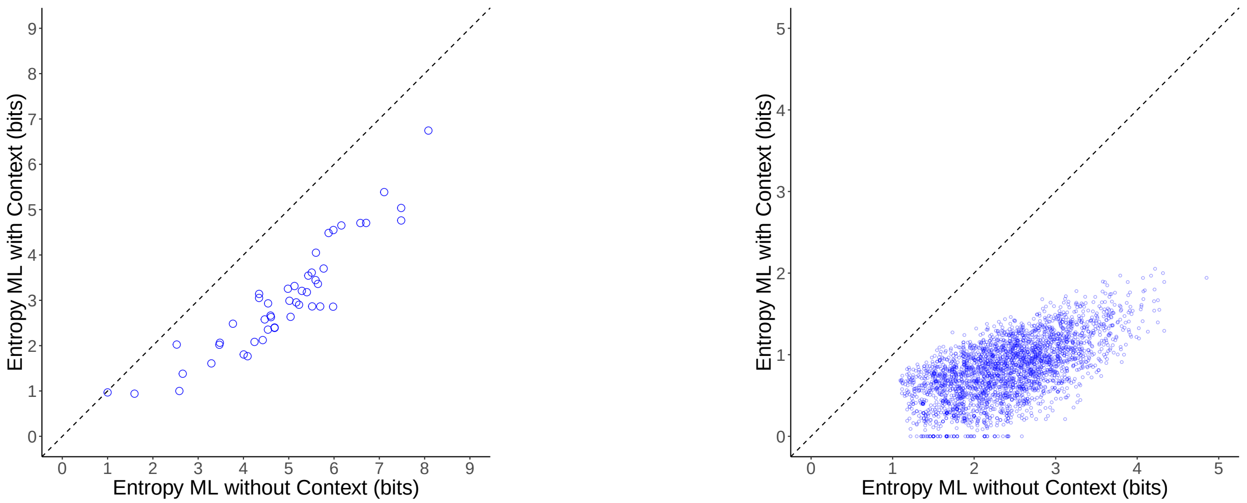
We also evaluated the other entropy estimators previously mentioned with different types of contextual information in the next-cell prediction problem. As shown in Table 3, the use of contextual information reduces the entropy of all estimators, showing that this type of information is indeed useful for predictability.

As we show in Table 3, the Lempel-Ziv estimator performs better (yields lower entropy values) than the other estimators in almost all scenarios. But, why does it perform worse than the other estimators with temporal information? As mentioned before, we split the history of locations of each user into 24 time series, one for each hour. This division makes each of the new sequences considerably smaller, which makes it hard for the Lempel-Ziv estimator to converge to the real entropy of the sequence. If we look at the results for the GPS dataset using the hour of day as contextual information, we see that the Lempel-Ziv estimator performed better than the other estimators. This is precisely the scenario (among the ones that use the hour of the day as contextual information) in which we have longer time series.

From Table 3, we see that, except for the case of weekday information, the Lempel-Ziv entropy estimate (without any contextual information) performs better than all of the other entropy estimates. And even in the case of weekday information, the Lempel-Ziv estimate without context gave comparable results. Thus, even though contextual information is useful to reduce the entropy estimates in the next cell problem, our results suggest that history plays a more important role in this prediction task. In other words, even though context affects people's decisions to visit (or not) a given place, in the long term, routine (sequences of places often visited) seems to play a bigger role. For instance, people usually go to work and go home at a certain time, regardless of the weather. Thus, using weather information to try to predict a home/work situation would probably add noise to the prediction.

5.4 Context and Next Place Prediction

In this section, we measure the impact of contextual information on the entropy estimators discussed in Section 5.2 in the next place prediction problem. First, as shown in Figure 4, the use of contextual information is useful for predictability in this prediction task. This figure shows entropy values for the *Maximum Likelihood* estimator with and without contextual information. It also shows that by using contextual information we were able to obtain lower entropy values, which indicates the such information brings up mobility patterns that were hidden when no context was used.



(a) GPS dataset

(b) CDR dataset

Figure 4: Reduction in the entropy values when contextual information (hour of the day in this case) is used, in the next place predictionn problem.

Another important result shown in Table 3 regards the gap between the entropy value with and without contextual information. This gap is wider for the next place prediction problem, which suggests that contextual information plays a larger role in the next place prediction problem. This problem, as mentioned before, is a harder task than next-cell prediction, given that in the next place problem there is no stationarity involved (the next location has to be different than the current one).

We have argued that highly stationary sequences tend to have lower entropy, and we have also argued that context tends to reduce entropy. But what would happen if we removed the stationarity of our input sequences but kept contextual information? This is precisely what we do in the next place prediction problem. As we show in Table 3, the Lempel-Ziv estimator also tends to produce lower entropy in this situation, except in the case of temporal information, for which the ML estimator yields lower entropy values than the Lempel-Ziv estimator. In this case, as argued before, the time series are too short for the Lempel-Ziv estimator to converge.

Another way to interpret these results is that, for small sequences, it is better to use other entropy estimators. This is useful, for instance, in computing the predictability for short-term mobility. It also suggests that a combination of estimators may work well in certain situations, e.g., when a new user enters the dataset. As there is not enough history for the Lempel-Ziv to find long-term correlations in the data, it might be better to estimate the predictability using the *Maximum Likelihood* estimator. Later on, when the user's history reaches a certain size, Lempel-Ziv estimator could be used. The size for which the Lempel-Ziv starts to converge varies according to the dataset.

6 Conclusions and Future Work

In this paper, we have analyzed the state-of-the-art technique to compute the limits of predictability in human mobility. We have claimed that, despite its extensive use in the literature, this technique has two shortcomings that were addressed in this study.

First, because it is based on a sophisticated compression algorithm, it has low interpretability. We have used two measures (regularity and stationarity) that help explain most of the variability in the entropy of mobility traces. We have built a regression model that uses these two measures as a proxy to explain the 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.

Second, we showed that the technique to estimate the predictability in human mobility does not allow for the use of contextual information in a natural manner. However, this type of information is used by prediction models, which leads to a mismatch between the predictability estimates and the prediction strategies in human mobility. We proposed to address this issue by using different entropy estimators that can use contextual information. We showed that, in most scenarios, when we hard code contextual information into the Lempel-Ziv estimator we still get lower entropy values than the other estimators. The only cases in which the other estimators resulted in lower entropy was in the case of small sequences. We argued that these estimators could be useful for predictability in short-term mobility. Finally, given the widespread use of the limits of predictability and the diversity of interpretations it has received, we found it appropriate to discuss its theoretical foundations and practical aspects.

As future work, we would like to evaluate the benefits of contextual information using the measures proposed in the first part of the paper, namely regularity and stationarity, as well as the effect of other metrics such as semantic regularity [30] on predictability. We would also like to investigate the predictability of specific algorithms to determine possible cases in which certain methods could or could not reach the predictability estimates of mobility datasets.

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-Lászlό 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] 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.

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

[9] 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.

[10] 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.

[11] 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.

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

[13] 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.

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

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

[22] 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.

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

[24] 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.

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

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

[27] 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.

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

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

[30] 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.