HAL open science

Understanding routine impact on the predictability estimation of human mobility

Douglas Teixeira, Jussara M Almeida, Aline Carneiro Viana

To cite this version:

Douglas Teixeira, Jussara M Almeida, Aline Carneiro Viana. Understanding routine impact on the predictability estimation of human mobility. 2021. hal-03128624

HAL Id: hal-03128624

https://inria.hal.science/hal-03128624v1

Preprint submitted on 2 Feb 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

RESEARCH

Understanding routine impact on the predictability estimation of human mobility

Douglas Teixeira[3,1,2*], Jussara M. Almeida[3] and Aline Carneiro Viana[1]

*Correspondence: douglas@dcc.ufmg.br
[3] Federal University of Minas Gerais, 31270-901, Belo Horizonte, Brazil
Full list of author information is available at the end of the article

Abstract

Given the difficulties in predicting human behavior, one may wish to establish bounds on our ability to accurately perform such predictions. In the case of mobility-related behavior, there exists a fundamental technique to estimate the predictability of an individual's mobility, as expressed in a given dataset. Although useful in several scenarios, this technique focused on human mobility as a monolithic entity, which poses challenges to understanding different types of behavior that may be hard to predict. In this paper, we propose to view human mobility as consisting of two components, routine and novelty, with distinct properties. This alternative view of one's mobility allows us to identify unpredictable behavior in each of these components. Additionally, we argue that unpredictable behavior in the novelty component is hard to predict, and we here focus on analyzing what affects the predictability of one's routine. To that end, we propose a technique that allows us to (i) quantify the effect of novelty on predictability, and (ii) gauge how much one's routine deviates from a reference routine that is completely predictable, therefore estimating the amount of unpredictable behavior in one's routine. Finally, we rely on previously proposed metrics, as well as a newly proposed one, to understand what affects the predictability of a person's routine. Our experiments show that our metrics are able to capture most of the variability in one's routine (adjusted R^2 of up to 84.9% and 96.0% on a GPS and CDR datasets, respectively), and that routine behavior can be largely explained by three types of patterns: (i) stationary patterns, in which a person stays in her current location for a given time period, (ii) regular visits, in which people visit a few preferred locations with occasional visits to other places, and (iii) diversity of trajectories, in which people change the order in which they visit certain locations.

Keywords: human mobility; predictability; entropy; mobility metrics

1 Introduction

Human mobility prediction has broad and important applications in areas such as urban planning, traffic engineering, epidemiology, recommender systems, and advertisement, to name a few [1–3]. Many previous studies proposed mobility prediction strategies that use a myriad of techniques (e.g., Markov chains [4], logistic regression [5], neural networks [6], and so on), and used different types of data sources (call detail records from mobile traffic [4, 7], GPS traces [5, 8], and social media data [3, 9], among others). However, human mobility is hard to predict, as there are many factors, such as the person's mood, traffic conditions, and current weather, that play a role in mobility-related decisions.

Given the difficulties involved in predicting mobility-related behavior, one poses the question of *to which extent such behavior can be predicted.* Answering this ques-

tion was the focus of a seminal paper by Song *et al.* [7], in which the authors proposed a fundamental technique that explores the concept of *entropy* to estimate how *predictable* a person's mobility is, as expressed in a given dataset. Specifically, given a dataset containing a sequence of locations visited by an individual and representing a sample of the individual's mobility, they proposed a technique that measures the maximum theoretical accuracy that an ideal prediction model could achieve on that dataset. This maximum is referred to as the *predictability* of the person's mobility. Unlike particular comparisons of alternative prediction models on different datasets, Song *et al.*'s approach is much more fundamental: *it does not focus on any specific prediction technique* but rather on human behavior, as captured by the available data. It is thus an invaluable tool in human mobility studies as it levels the field of mobility prediction, *providing a value of prediction accuracy that prediction strategies should aim for.*

However, Song *et al.*'s work and subsequent studies derived from it [4, 5, 10–12] focused on a person's mobility as a single monolithic entity. This paper proposes a different approach: we argue that one's mobility can be broken into different *components.* Separately studying such components can reveal important insights into an individual's mobility patterns. Specifically, we offer an alternative view of human mobility as having two fundamental components, namely *novelty* and *routine*, with distinct properties. The *novelty* component consists of locations that the person visited for the first time, and we refer to all other visits (which may include the same location multiple times) as the *routine* component. Note that this definition is different from our usual definition of routine (places frequently visited), as it considers every visit except the first one as being part of the routine component.

The division of human mobility into these components highlights important properties about them. As we will discuss in Section 3.1, the novelty component is remarkably unpredictable, mainly because the vast majority of mobility prediction models [9, 13–15] rely on the history of visited locations, as captured in the input dataset, to predict future visits. Therefore, those models have a hard time deciding whether a person will go to a previously unseen location, and an even harder time trying to guess what location that will be.

In contrast, the routine component is the part of a person's mobility where there is more potential for improving prediction accuracy as every location in this component has been visited at least twice (that is, there is visitation history to be exploited by prediction models). However, despite such greater potential, predicting visits in the routine component is still by itself a challenge, as there can be a high degree of unpredictability even if we focus only on previously visited locations. For instance, the mere change in the order in which people visit specific locations, even those they visit more frequently, poses difficulties for prediction models.

Having defined the routine and novelty components, we set up the goal of isolating their effects on the predictability of one's mobility. Specifically, in Section 3, we show how to isolate the effect of the novelty component on predictability, thus allowing us to quantify the effect of routine one a person's mobility predictability. Then in Section 5, we zoom in on the routine component to try to understand what makes a person's routine easier or harder to predict. To do that, we rely on our previously

proposed metrics [12], namely regularity and stationarity, to try to understand what affects a person's routine. We also propose a novel metric, called *diversity of trajectories* which, together with regularity and stationarity, paint a clearer picture of what affects the predictability of the routine component of one's mobility. We evaluate these three metrics by building regression models that use them as proxies to understand the predictability of an individual's routine. Our study relies on the analysis of two datasets of different spatial and temporal granularities, as these properties have been shown to influence predictability [10, 12].

Our contributions can be summarized as follows:

- A new way to look at human mobility that breaks it into two components, *novelty and routine*, which have different properties with respect to how predictable one's mobility is.
- A strategy to filter out the effects of the novelty component and sequence size on the estimate of predictability of one's mobility, thus allowing us to focus on the component with greater potential for prediction purposes, i.e., the routine component. This strategy allows us to estimate how the predictability of one's mobility deviates from a reference case consisting of a completely predictable routine component (but similar novelty), thus offering an estimate of the effect of the routine component on predictability estimates.
- A novel metric, i.e., diversity of trajectories, that, together with previously proposed metrics, can be used as proxy in the interpretation of the predictability of one's mobility. This new metric enhances the understanding of predictability by bringing a finer view of the factors that impact predictability.
- A rigorous evaluation of the impact of the diversity of trajectories and the previously proposed regularity and stationarity on the predictability of one's routine, offering valuable insights into understanding and predicting mobility patterns.

The rest of this paper is organized as follows. In Section 2, we discuss relevant background to understand predictability as well as related work on the topic of predictability. In Section 3, we describe our approach to separate human mobility into novelty and routine, which allows us to assess the effect of novelty on predictability to measure how much one's routine deviates from a reference (completely predictable) routine. In Section 4 we present and briefly analyze the datasets used in this study. We then discuss, in Section 5, what affects the predictability of a person's routine, relying on previously proposed metrics as well as on the new diversity of trajectories metric, and making use of regression-based analyses to show that these metrics can indeed explain reasonably well the entropy (and thus the predictability) of a person's routine-related mobility. Finally, Section 6 summarizes our main results and discusses potential directions for future work.

2 Background and Related Work

In this section, we provide relevant background to understand predictability as well as discuss previous studies on the topic. Specifically, we start by defining, in Section 2.1, the individual human mobility prediction problem as well as two variants of this problem (next-cell and next-place prediction), as predictability and prediction are intimately related, *i.e.*,, predictability is a measure of the maximum prediction

accuracy achievable on a given dataset. In Section 2.2, we discuss several aspects of predictability, including its relation to mobility prediction and its dependence on entropy [7]. Finally, in Section 2.3 we review some prior efforts to analyze predictability.

2.1 Individual Human Mobility Prediction

The prediction of human mobility has been studied under a number of different perspectives. For instance, there are studies that focus on aggregate mobility, that is, at a population level, and aim at predicting the direction in which groups of people will flow. Examples of studies in this area include the work of Brockman *et al.* [16], in which the authors investigated human travelling by analyzing the circulation of bank notes in the United States, and the study of migration flows between regions according to the radiation model [17].

On the other hand, in individual mobility prediction, the goal is to provide a fine-grained approach to human mobility by focusing in forecasting the whereabouts of particular individuals. Examples of prior efforts in individual mobility prediction are the efforts of Gonzalez *et al.* [18] and Mucceli *et al.* [19], in which the authors showed that human trajectories exhibit spatio-temporal regularities, and the mobility of individuals can be characterized by frequent visits to a few preferred locations interspersed with occasional visits to other locations, previously visited or not. In this article, we focus on *individual human mobility*.

A common feature of the vast majority of individual mobility prediction strategies is that they rely on a person's history of visited locations to perform predictions. That is, they extract patterns from an input dataset containing such history and, by assuming that these patterns will hold in the future, use them to predict future locations. Thus, we can define the *individual human mobility prediction problem* as follows:

Definition 2.1 ***Individual human mobility prediction:*** Given an input dataset consisting of a time-ordered sequence $X = (x_1, x_2, \ldots, x_{n-1})$ of observations of a person's location, where each symbol $x_i \in X$ identifies the location the person was at when observation i was made, we wish to predict x_n, the next symbol (location) in the sequence.

The aforementioned mobility prediction problem admits different formulations, depending on specific constraints, corresponding to different prediction tasks. In this study, we will focus on two particular prediction tasks, namely *next cell* and *next place* predictions [5, 20]. Again, given a time-ordered sequence $X = (x_1, x_2, \ldots, x_{n-1})$ of observations of a person's location, these prediction tasks are defined as follows.

Definition 2.2 ***Next-cell prediction:*** Predict x_n, the next location in sequence X. Notice that here, location x_n can be equal to x_{n-1} in case the person stays at her current location for several consecutive observations (stationary period).

Definition 2.3 ***Next-place prediction:*** Predict the next location $x_n \in X$, where $x_n \neq x_{n-1}$. Notice that x_n must be different from x_{n-1} by definition, since we want

to know the next (distinct) location the person will visit. In other words, in this prediction task we ignore stationary periods.

There are two main options for carrying out these prediction tasks in a given dataset. The first option is to work with the full dataset and adjust the predictions accordingly. For instance, in the next-place prediction task, one would ignore every next location that is equal to the previous one while performing predictions. The second option is to filter the dataset so as to eliminate stationary periods when performing next-place prediction. For instance, consider an example sequence $X = (A, B, A, A, A, D, B, B, B, C, F)$. For the next-cell prediction task, X would remain unchanged, whereas for the next place prediction task, X would become $X' = (A, B, A, D, B, C, F)$. Throughout the rest of this paper, whenever we refer to a dataset for the next-cell prediction task, we are using the unchanged dataset, and when we refer to a next-place dataset, we are considering the original dataset after every consecutive repetition of the same location is removed.

It is worth to emphasize that the removal of stationary periods has an impact on prediction and on predictability. While in *next-cell prediction*, stationary periods will contribute to improve prediction accuracy (thus raising predictability), in the *next-place prediction*, their absence will make predictions more challenging (thus lowering predictability). Recall that predictability is a measure of the maximum prediction accuracy achievable on a given dataset, thus, by increasing the potential for prediction accuracy in a particular task and dataset, we are also increasing predictability.

2.2 Predictability in Human Mobility

When tackling individual human mobility prediction, one might ask the extent to which the mobility of one particular individual, captured in a dataset, can be predicted at all. This question relates much more to the inherent behavior of this person, that is, to the *predictability* of her mobility patterns, expressed in an underlying dataset, rather than to the effectiveness of particular prediction methods. Song *et al.*'s seminal paper [7] present a technique to estimate the predictability of an individual given a trace of her mobiltiy. This is the state-of-the-art technique to estimate one's mobility predictability and is the foundation of our present effort. Thus, throughout this paper, whenever we mention predictability, we are referring to Song *et al.*'s predictability technique.

Predictability is a number between zero and one, where zero indicates that a dataset containing a sample of an individual's mobility is unpredictable and one that data is completely predictable. A value of $x\%$ means that *an ideal predictor is expected to accurately guess the next symbol in the dataset representing the person's mobility trace $x\%$ of the time.* As we will discuss later, there are restrictions both on the type of process (e.g., stationary ergodic processes) by which the input sequence is produced as well as the type of predictor (e.g., universal predictors) for which this holds. For now, it is sufficient to notice that this definition *accommodates both the next cell and next place prediction tasks.* Notice also that predictability is a theoretical upper bound, which is obtained by decoupling the analysis of the data from the intricacies of a given prediction technique, and is therefore based solely on the inherent nature of human mobility behavior expressed in the data.

2.2.1 The Relationship Between Predictability, Entropy, and Compressibility

The technique proposed by Song *et al.* [7] estimates the predictability of a person's sequence of visited locations as a function of the *entropy* [21] of this sequence, thereby assuming a connection between entropy and complexity, which in turn is related to predictability.[1] Specifically, the predictability of a sequence of symbols (locations visited by someone, in the present case) is related to the *complexity* of the sequence (less complex sequences are more predictable), and complexity is related to entropy. Entropy, which can be defined as the average uncertainty in the outcomes of a random variable [21], is a good approximation for the complexity of the sequence because the more uncertainty (higher entropy) in a sequence of events, the more complex the sequence.

Additionally, the entropy of a sequence of symbols can be seen as a lower bound on its compressibility [21–24]. The intuition here is that *if a sequence of symbols is highly compressible, it means that there is little uncertainty in the order its symbols appear.* For instance, sequences with many repeated symbols are highly compressible and, intuitively, if a sequence has many repeated symbols it is relatively easy to predict its next symbol at a given point. Similarly, in the case of mobility, if a person visits many repeated locations, the sequence (mobility trace) will have many repeated symbols, which makes prediction easier.

As a result of this equivalence between entropy and compressibility, one can use the entropy of a sequence as a measure of how predictable the sequence is: the lower the entropy the less complex and more predictable the sequence, and vice-versa. Thus, *the problem of estimating the predictability of a sequence reduces to the problem of estimating the entropy of the sequence.*

Song *et al.* leveraged these theoretical connections, and proposed to use three estimates for the entropy of a person's mobility trace: one that assumes the person visits every location the same number of times, another that takes into account differences in the frequencies with which locations are visited, and a third, more precise one, based on compression, that takes into account both frequency and temporal patterns (that is, dependencies among visits). This third, more robust entropy estimator had indeed been originally proposed by Kontoyiannis *et al.* [25]. According to their definition, the entropy S_{real} of an input sequence of locations X of size n can be approximated by:

$$S_{real} \approx \frac{n \log_2(n)}{\sum\limits_{i \leq n} \Lambda_i}, \tag{1}$$

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 sequence X.

The intuition behind this formula is that, given a sequence of size n, its entropy is inversely proportional to the number and size of repeated substrings in the sequence. Thus, for example, a sequence with a lot of repeated sub-sequences has a lot of

[1] We note that the theory behind predictability is valid for a sequence of symbols in general, which, in the case of human mobility, are identifiers of the locations that someone visited

redundancy, and therefore has low entropy, i.e., it is more predictable. Throughout the rest of this article, whenever we mention the approach proposed by Song *et al.*, we are indeed referring to the method that exploits the entropy estimator proposed by Kontoyiannis *et al.*, expressed in Equation 1.

Delving deeper into the literature, we found that there are some caveats with respect to the type of sequence on which we can expect Kontoyiannis *et al.*'s estimator to work reliably. It assumes that the input sequence X is produced by a *stationary ergodic process*. In the case of human mobility, this implies that statistical properties of a person's mobility patterns do not change over time and that these statistical properties can be inferred from a single, sufficiently long random sample of the person's mobility trace. In other words, *the input sequence has to be representative of the person's actual mobility, and there cannot be long-term changes in the patterns.* This seems a reasonable assumption for traces covering daily patterns, though it might not be adequate if a person's mobility patterns undergo significant changes in a given period, *e.g.*,, the person moves to a different city, as discussed in previous work [26].

The assumption that individual human mobility is a stationary ergodic process also has implications on the type of predictor for which Song *et al.*'s technique is expected to work. In particular, Song *et al.*'s predictability estimate holds as an upper-bound only for *universal predictors*.

A universal predictor is one that does not depend on the knowledge of the underlying process generating the input sequence and, as the sequence grows to infinity, it still performs essentially as well as if the process were known in advance [23, 24, 27]. In more practical terms, universal predictors are able to generalize to different datasets, provided that the underlying processes producing these different datasets belong to the same class (e.g., stationary ergodic processes). Markov-based models are examples of universal predictors.

In contrast, *non-universal predictors* must be trained and therefore, are tailored to a specific dataset, and thus may not generalize to other datasets. One example includes a predictor based on neural networks that is specialized to a particular dataset.

2.2.2 Defining Predictability

Using Fano's Inequality [21], Song *et al.* derived a formula to compute the *predictability* of a given sequence, based on the entropy of the sequence. Such formula is based on the intuition that, if a user with entropy S moves between N *distinct* locations, her predictability will be Π, where $\Pi \leq \Pi_{max}(S, N)$ and Π_{max} is an estimate on predictability, such as:

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

and $H(\Pi_{max})$ 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 [7, 10, 28]. In particular, Smith et al. [10] provided a detailed, thorough derivation of the formula above.

Now, with the necessary background in place, we can more formally define the problem of estimating predictability in human mobility.

Definition 2.4 ***Predictability:*** Given a time-ordered sequence of locations $X = (x_1, x_2, \ldots, x_{n-1})$ that a person visited in the past, and assuming that X is a stationary ergodic process, the predictability task is to estimate Π_{max}, the maximum possible accuracy that a universal predictor U could achieve when trying to predict x_n in X.

2.3 Related Work

Song *et al.*'s technique, which is the state-of-the-art predictability technique, has been extensively used to assess predictability in human mobility as well as in other scenarios. In the domain of human mobility, Xin Lu *et al.* [4] investigate whether the prediction accuracy obtained via Song *et al.*'s technique is achievable. They propose and evaluate several markov models to predict people's next location and show that their models achieve Song *et al.*'s estimated predictability for their dataset. Later work proposed models that could even surpass the predictability in some circumstances [26]. Being based on neural networks, those models are not universal predictors. As argued in the previous section, the predictability estimate is not guaranteed to hold as an upper-bound in such cases.

Smith *et al.* [10] evaluate Song *et al.*'s technique in a GPS dataset, showing that users's predictability are sensitive to the temporal and spatial resolution of the data. Ikanovic *et al.* [20] use Song *et al.*'s technique to estimate predictability in different prediction tasks, showing that predictability varies according to the particular prediction task under consideration. Cuttone *et al.* [5] also show that prediction accuracy varies depending on other factors in the data, such as contextual information (day of the week, hour of the day, the weather, etc.) and suggest that context could impact predictability.

Song *et al.*'s predictability technique has also been used in other domains. For instance, Li *et al.* [29]. build on Song *et al.*'s technique to assess spatiotemporal predictability in location-based social networks. Bagrow *et al.* [30] use Song *et al.*'s technique to measure the predictability of the contents of a person's *tweets* based on the content of her friends' *tweets*. Zhao *et al.* [28] use Song *et al.*'s technique to measure the predictability of taxi demand per city block in New York City, and other work also use it scenarios such as travel time estimates [31], cellular network traffic [32], and radio spectrum state dynamics [33].

Previous work also focused on providing more refined limits of predictability. These refined limits usually rely on some sort of external information or assumption about the sequence to adjust the probabilities of the symbols in the input sequence. For instance, Smith *et al.* [10] showed that the limits of predictability can be refined if we exclude from the possible next locations those that are far away from the user's current position. Teixeira *et al.* [12] quantified the impact of context on predictability estimates, showing that context does not always increase predictability.

Previous work also contrasted the limits of predictability for the next-cell and next-place prediction tasks [5, 20], showing that the predictability for the next-place prediction problem is lower than that of the next-cell prediction problem, given the same input sequence. As argued by the authors, this indicates that the next-place prediction is a harder problem. The reason for that is the lack of stationarity (an important feature of human mobility) in the next-place prediction problem, as already argued in Section 2.

Unlike [12] and other predictability studies [4–7, 10] where human mobility is considered as a whole, here we propose a strategy to separate a person's mobility into two components: *novelty and routine*. By doing so, we aim to simplify the understanding of the predictability of a person's mobility, to assess the effects of novelty on predictability estimates, and consequently, to be able to identify routine-related behavior that is hard to predict. In order to analyze routine-related mobility, we rely on previously proposed metrics, namely regularity and stationarity, as well as a newly proposed one, called diversity. These metrics allow us to understand what affects the predictability of the routine component of a person's mobility.

In the next sections, we discuss each of these contributions. We start with our new perspective to study human mobility according to two distinct and complementary components. We then investigate the implications of such components on predictability estimates.

3 Components of Human Mobility

As mentioned, previous studies looked at individual human mobility as one monolithic entity consisting of a collection of locations that a person visited during a certain period. In this paper, we propose to break one's mobility into two key components – *novelty and routine* – as follows.

Given an input sequence $X = (x_1, x_2, \ldots, x_n)$ of locations visited by an individual, the *novelty component* of X consists of all visits to previously unseen locations, whereas its *routine component* includes all other visits, that is, visits to locations that appeared at least once before in X. Figure 1 shows an example input sequence X representing a person's history of visited locations (each letter represents a location). The figure distinguishes the *routine* and *novelty* components by presenting the latter in gray.

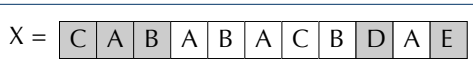


Figure 1 Novelty (in gray) and routine (in white) components of input sequence X.

The difficulty in predicting a person's mobility comes, mainly, from one of two sources: (1) unpredictable behavior due to visits to novel (previously unseen) locations, and (2) unpredictable behavior in the sequence of visits to previously visited locations, due to spatio-temporal changes. In this article, we argue that in order to better understand how predictable an individual's mobility patterns are, we must isolate these two sources of unpredictability and study them separately. By doing so, we can estimate the effect of novelty on predictability, and then zoom in on what affects the predictability of the routine component alone.

We argue that novelty visits contribute to reducing the predictability of a person's mobility. The vast majority of mobility prediction models exploit the history of visited locations, as captured in the input sequence X, to predict future visits (e.g., [4, 5, 14, 20]). Thus, the absence of such history in the novelty component (by definition) challenges prediction. Predicting novelty visits requires different approaches, that may exploit other types of (external) information such as mobility patterns of closely related individuals such as friends and family [15, 34], which are outside our present scope.

The routine component, on the other hand, has a greater potential for prediction accuracy as previous visitation history is available. However, as mentioned above, changes in the sequence of visitations, triggered by a plethora of factors (weather, special events, one's own will, etc.) can introduce a great deal of unpredictability to this component as well.

In this paper, *we study the predictability of one's mobility focusing on the routine component*. We do so while still using the state-of-the-art predictability technique. However, that technique views one's mobility as a whole, i.e., processes the complete input sequence X. By doing so it hardens the understanding of what part of the (un)predictability of a person's mobility, expressed in X, is due to visits in the novelty component and what part is due to changes in the sequence of routine-related visits.

Thus, as a key step towards understanding predictability, we here propose a technique that filters out the effect of other factors that impact predictability, allowing us focus on routine-related mobility captured in the input sequence. Specifically, our approach consists of building a comparable reference sequence, here called simply *baseline* sequence, which differs from the original sequence only in the routine component. Specifically, the routine component of the baseline sequence consists of the same symbol repeated multiple times, thus having maximum predictability (for fixed routine size). By measuring the gap between the (real) predictability of the original sequence to the predictability of this baseline sequence, we are able to estimate the effect of the routine component on predictability estimates.

In the following, we first discuss the impact of one such effect, notably the visits in the novelty component (Section 3.1). We then present our proposed approach to capture the effect of routine-related mobility on the predictability of one's mobility (Section 3.2).

3.1 Assessing the Effect of Novelty on Predictability Estimates

Despite the challenges associated with predicting visits in the novelty component, we here claim that it is possible to estimate the impact of this component on the predictability of an individual's mobility. In this section, we explain how to do so.

Recall from Equation 1 that the entropy of a given sequence X of size n is inversely proportional to the sizes of the *distinct* subsequences in X. For a given size n the larger the sizes of the subsequences, the fewer subsequences, and vice-versa. Thus, the entropy is proportional to the number of distinct subsequences of the original sequence. Symbols in the novelty component have a direct impact on entropy estimates because every time a previously unseen symbol appears in the sequence, it will generate a previously unseen subsequence, which in turn will contribute to increase the entropy estimate of the sequence as a whole.

Specifically, from Equation 1 (reproduced below to facilitate the explanation):

$$S_{real} \approx \frac{n \log_2(n)}{\sum\limits_{i \leq n} \Lambda_i},$$

we notice that, for a sequence of size n, its entropy will be inversely proportional to $\sum_{i \leq n} \Lambda_i$, i.e., the sum of the lengths of all subsequences. In the extreme case of a sequence whose symbols are all unique, every symbol in the sequence will produce a new (previously unseen) subsequence (of length one). In that case, each Λ_i will be equal to 1, and thus the denominator in Equation 1 will be equal to n. In general, for a sequence of size n with $m \leq n$ distinct symbols, these symbols, taken together, will contribute to the denominator of Equation 1 with a value of m.

Consider, as an example, the input sequence $X = (H, W, H, W, S, H, W, H, W, R)$. The entropy estimate, as explained, has to account for every symbol that appears in the sequence for the first time. Table 1 illustrates the effect of these symbols on the entropy by showing the computation of each Λ_i – the size of the shortest subsequence L_i starting at position i that does not appear in positions 1 to $i-1$ in sequence X. To facilitate following the example, the table shows, for increasing values of i from 1 to $n = 10$, the subsequence L_i as well as its corresponding Λ_i. Note that for $i = 3$, we have $\Lambda_3 = 3$, which is the size of HWS, the shortest subsequence starting at position 3 that does not appear before in the input sequence, since S does not appear in the earlier positions of X. In contrast, for $i = 5$, we have $\Lambda_5 = 1$, since the fifth location visited, S, is novel, it has not appeared before in the sequence. The same happens for all visits to new locations: $\Lambda_i = 1$ for $i =$ 1,2,5, and 10.

i	$X_{[1:i]}$	L_i	Λ_i	new symbol?	new subsequence?
1	HWHWSHWHWR	H	1	✓	✓
2	HWHWSHWHWR	W	1	✓	✓
3	HWHWSHWHWR	HWS	3	✓	✓
4	HWHWSHWHWR	WS	2	✗	✗
5	HWHWSHWHWR	S	1	✗	✗
6	HWHWSHWHWR	HWHW	5	✓	✓
7	HWHWSHWHWR	WHW	4	✗	✗
8	HWHWSHWHWR	HW	3	✗	✗
9	HWHWSHWHWR	W	2	✗	✗
10	HWHWSHWHWR	R	1	✗	✗

Table 1 An example illustrating the innerworkings of Equation 1 on an input sequence $X = (H, W, H, W, S, H, W, H, W, R)$. The notation $X_{[1:i]}$ denotes the symbols in X from 1 to $i-1$, L_i denotes the shortest subsequences that starts at position i and does not appear from 1 to $i-1$ in the original sequence, and Λ_i is given by $\mid L_i \mid$. We note that every time a new (previously unseen) symbols appears, a new subsequence is generated, as shown in the last two columns of the table.

In more general terms, we note that every time a new (previously unseen) symbol appears in the sequence, a new (previously unseen) subsequence also appears, each new symbol contributes the value of 1 to its correspondent Λ_i. Furthermore, as shown in the Appendix A, changing the order or positions of the symbols that consitute the novelty component does not affect their contribution to the entropy estimate. Thus, we can isolate the symbols in the novelty component, as described in Section 3, in order to focus on understanding the routine of one's mobility.

Given that we are viewing human mobility in terms of two components, and that we have identified the impact of the symbols in the novelty component on the denominator of Equation 1, we can rewrite that equation as follows:

$$S_{real} \approx \frac{n \log_2(n)}{\sum\limits_{i \leq n-m} \Lambda_i^{routine} + \sum\limits_{i \leq m} \Lambda_i^{novelty}} = \frac{n \log_2(n)}{\sum\limits_{i \leq n-m} \Lambda_i^{routine} + m}, \tag{3}$$

where n is the size of the sequence, m is the number of symbols in its novelty component, $\Lambda_i^{novelty} = m$ is the contribution of the symbols in the novelty component to the denominator of Equation 1, and $\Lambda_i^{routine}$ is the effect of routine on the denominator of Equation 1.

In the following section, we explore these insights to propose a technique that allows us to *estimate* the effect of the routine component on the predictability of an input sequence X. Our technique relies on the fact that we are able to isolate the effect of the novelty component on the entropy (Equation 3), thus facilitating our study of the predictability of the routine component. In isolating the effect of novelty of predictability, we highlight the role of routine and thus are able to focus on what affects the predictability of this component.

3.2 Assessing the Effect of Routine on Predictability Estimates

In order to estimate the predictability of a person's routine, captured in an input sequence X, using the technique proposed by Song *et al.*, we must be able to filter out from the computation, the effects of other unrelated factors present in X. One such factor is the novelty component, which, as argued in the previous section, contributes to reduce predictability. Another factor is the size of the input sequence, given by parameter n, which, as shown in Equation 1, also affects the predictability estimate of X.

Having identified these two factors, we proceed to describe our approach to estimate the effect of the routine component on the predictability of an input sequence X. In a nutshell, our proposed approach works as follows. Given the input sequence X, with size n, our technique consists in creating another sequence, named *baseline* sequence, based on the original, in such a way that this new sequence:

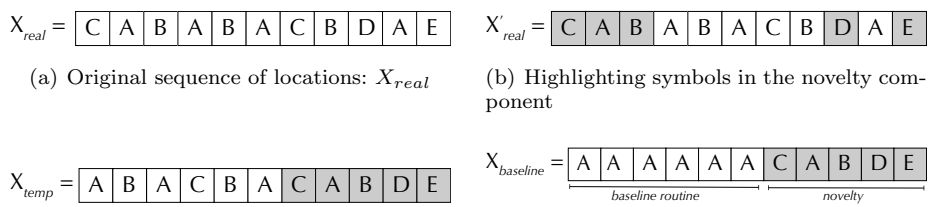
(i) has the same size n as the original sequence;
(ii) has the same number of visits in the novelty component;
(iii) its routine component is completely predictable, i.e., it consists of a single location visited as many times as determined by the size of the routine component.

We note that steps (i) and (ii) are required so as *to filter out the effects due to the size of the input sequence* (notably the size of its routine component) and *to isolate the effects of the novelty component on the predictability estimate.*

By doing so, we guarantee that the two sequences, the original one and the baseline one, created as described, are comparable in terms of the impact of the novelty and the size of the sequence on the predictability estimate. As such, *any difference between the estimates of the predictability of both sequences must highlight the effect of the routine in the original sequence.* In other words, our approach allows us to

assess how much a person's routine deviates from a completely predictable *baseline routine*.

Figure 2 exemplifies how the baseline sequence is built. For the sake of clarity, we refer to the original (input) sequence of visited locations as X_{real} and to the *baseline* sequence as $X_{baseline}$. Consider the sequence X_{real} in Figure 2(a), and assume it consists of locations (each identified by a letter). The first step to build $X_{baseline}$ is to identify visits that constitute the novelty component, which are highlighted in gray in Figure 2(b).



(a) Original sequence of locations: X_{real}

(b) Highlighting symbols in the novelty component

(c) Isolating symbols in the novelty component: X_{temp}

(d) Baseline sequence of locations: $X_{baseline}$

Figure 2 Example of construction of a baseline sequence of locations.

In order to isolate the novelty component, we first move to the back of the sequence all symbols that are part of it. Recall that, as argued in Section 3.1 and shown in the Appendix A, changing the positions of the symbols that compose the novelty component does not impact their contribution to the entropy estimate. Thus, by moving them to the back of the sequence we do not alter its effect on the predictability of the sequence. The result is a temporary sequence X_{temp} shown in Figure 2(c), where visits that constitute the novelty component are isolated. We then consider the following question: *If the routine component of the original sequence were completely predictable, what would be the predictability of the whole sequence?*

In order to answer this question, we change sequence X_{temp} by creating a routine component that is completely predictable, i.e., it consists of only a single symbol repeated multiple times. The resulting sequence constitutes the *baseline sequence* $X_{baseline}$, illustrated in Figure 2(d). Notice that, both $X_{baseline}$ and X_{real} have the same size and the same number of symbols in the novelty component, therefore the effects of size and novelty on predictability are the same for both sequences.

Our goal at this point is to: (i) estimate the entropy $S_{baseline}$ of sequence $X_{baseline}$, and (ii) compare $S_{baseline}$ with S_{real}, the entropy of the original sequence X_{real} so as to measure how much the routine component of S_{real} deviates from the baseline routine. We take this *relative* measure as an estimate of the effect of the routine on the predictability of the original sequence X_{real}. The greater the gap between S_{real} and $S_{baseline}$, the less predictable the routine component of X_{real} is, and the greater its effect on the predictability of the complete sequence.

To tackle the problem of estimating the entropy of the baseline sequence, we will revisit Equation 3. In Section 3.1, we established that the value of $\sum \Lambda_i^{novelty}$ is m, where m is the number of symbols in the novelty component of the sequence. We

will now explain how to compute $\sum \Lambda_i^{routine}$ for our baseline sequence, which has the distinct property that all of its symbols are the same.

Let's start with the example shown in Figure 2(d), where the routine component of $X_{baseline}$ is $AAAAAA$, i.e., has size 6. Table 2 shows the computation of each $\Lambda_i^{routine}$, with i varying from 1 to 6.

i	$X_{[1:i]}$	L_i	Λ_i
1	AAAAAA	A	1
2	AAAAAA	AA	2
3	AAAAAA	AAA	3
4	AAAAAA	AAA	3
5	AAAAAA	AA	2
6	AAAAAA	A	1

Table 2 An example illustrating the innerworkings of Equation 1 on an example input sequence $X = (A, A, A, A, A, A)$. The notation $X_{[1:i]}$ denotes the symbols in X from 1 to $i-1$, L_i denotes the shortest subsequences that starts at position i and does not appear from 1 to $i-1$ in the original sequence, and Λ_i is given by $\mid L_i \mid$.

Notice that, in line 4, even though the string AAA appears before, Λ_i is still 3, as we have reached the end of the sequence, and therefore cannot add more characters to L_i. In practice, this example follows how the Lempel-Ziv compression algorithm encodes substrings, and Λ_i is simply the size of the next substring that would be encoded by the Lempel-Ziv compression algorithm for each i.

From Table 2, we notice that the sum of all $\Lambda_i^{routine}$ can be written as $1 + 2 + 3 + 3 + 2 + 1 = 12$. More generally, if $X_{baseline}$ has a routine component of size k, we can state that:

$$\sum \Lambda_i^{routine} = 1 + 2 + \cdots + \frac{k}{2} + \frac{k}{2} + \frac{k}{2} - 1 + \frac{k}{2} - 2 + \cdots + 1 = \left\lceil \frac{k^2}{4} + \frac{k}{2} \right\rceil,$$

where k is the total number of symbols in the routine component of the sequence.

Thus, we can rewrite Equation 3 to compute the entropy of the baseline sequence as follows:

$$S_{baseline} \approx \frac{n \log_2(n)}{\left\lceil \frac{(k+1)^2}{4} + \frac{k+1}{2} \right\rceil + m}, \tag{4}$$

where n is the size of original the sequence, m is the number of symbols in its novelty component, and k is the number of symbols in its baseline routine. In the equation above, we have to add one to the size of the routine component to account for the fact that one of the symbols in the sequence appears both in its baseline routine and in its novelty component, *i.e.*, for practical purposes, it is as if the routine component had an extra symbol.

It is also important to highlight that applying Equation 4 to an input sequence X yields the same entropy value as using Equation 1 to compute the entropy of a sequence $X_{baseline}$ such as the one in Figure 2(d), *i.e.*, a baseline sequence obtained from an input sequence X. In other words, Equation 4 is a closed-formula for the entropy of a baseline sequence.

Having determined how to estimate the entropy of the baseline sequence, we can finally tackle the problem of estimating the effect of the routine component on the predictability of an individual's mobility expressed in an input sequence X_{real}. To that end, given the entropy S_{real} of the original sequence and the entropy $S_{baseline}$ of the baseline sequence, we can estimate the deviation of routine component on S_{real} from the baseline routine as follows:

$$\Delta_{S_{routine}} = S_{real} - S_{baseline}, \tag{5}$$

In other to better exemplify this perspective, consider as an example the sequence $X = (C, A, B, B, A, D, C, B, A, A, E, D)$, also shown in Figure 2(a). The entropy S_{real} of this sequence is given by:

$$S_{real}(X) \approx \frac{n \log_2(n)}{\sum\limits_{i \leq n} \Lambda_i} = \frac{12 \log_2(12)}{19} = 2.00.$$

In turn, we can calculate the entropy $S_{baseline}$ of the corresponding baseline sequence $X_{baseline} = (A, A, A, A, A, A, C, A, B, D, E)$, which is given by:

$$S_{baseline}(X) \approx \frac{12 \log_2(12)}{\left(\frac{7^2}{4} + \frac{7}{2}\right) + 5} = \frac{12 \log_2(12)}{21} = 1.81.$$

Here, the effect of routine on the entropy of X_{real} can be estimated as $2.00 - 1.81 = 0.19$. We argue that this entropy gap, *i.e.*,, deviation from the baseline routine, concerns behavior in the routine component that is hard to predict.

Having defined our technique to assess the effect of the routine component on the predictability of one's mobility, we use it in the following sections *to understand what makes routine-related mobility easier or harder to predict.*

4 Overview of our Mobility Datasets

In this section, we offer an overview of the mobility datasets used in our study. We start by first presenting, in Section 4.1, a brief description of them, discussing their main characteristics and filtering process adopted. Next, in Section 4.2, we offer a characterization of the data, focusing on properties that may affect predictability estimate.

4.1 Datasets

Our analyses are performed on two different mobility datasets, of distinct temporal and spatial resolutions, which allow us to investigate predictability in varying spatiotemporal contexts. These datasets are representative 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

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

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 5 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 of 132 volunteers distributed among six countries located in two different continents: 67 are from the same country and represent students, researchers, and administrative staff in two universities where lectures were held. To filter out potential cross-country effects, we decided to focus on users from the same country, that is, 67 users, in all of our analyses.

CDR dataset: The second dataset consists of *Call Detail Records* (CDRs), provided by a major cellular operator in China. It 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. This dataset is collected from 642K fully anonymized mobile phone subscribers. Here, a CDR is logged every time a subscriber initiates or receives a voice call. An entry in the dataset contains the subscriber's identifier, the call start time, and the location of the subscriber at this time. Unlike traditionally analyzed CDR datasets, the locations here represent the users' centroid of the hour, within a 200 meter radius, according to the instruction of the data provider, and does not contain the area covered by each tower. Hence, the accuracy of positioning is higher than that of traditionally analyzed CDR datasets. As some users do not have data covering the whole period, we focused on those who have at least one location registered each 2 hours, on average. This filtering criterion is the same adopted by Song et al. After this filtering process, we ended up with 3,349 users, which we use in our study.

Unless otherwise noted, we will use a temporal resolution of one observation every 5 minutes for each user in the GPS dataset, and we ensure that there is at least one observation per user every 2 hours for the CDR dataset. In both datasets, the size of the side of each square grid is 200 meters.

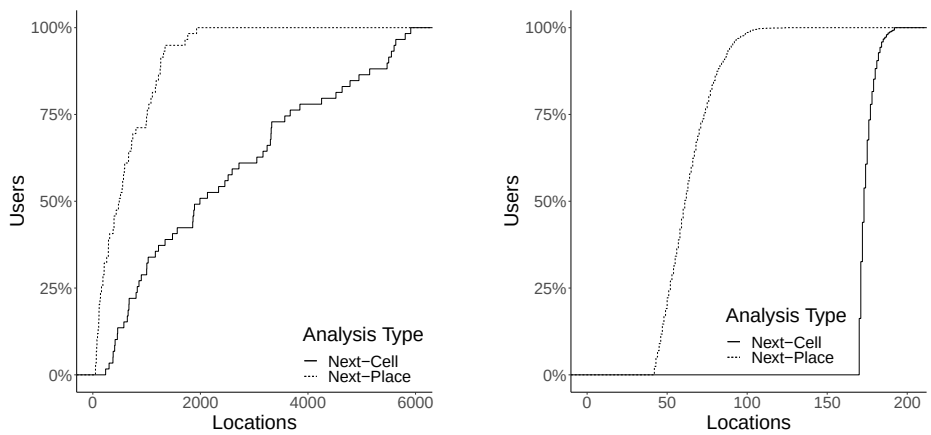
4.2 Data Characterization.

In this section, we study general properties of our datasets and discuss how they may affect predictability estimate.

We start by analyzing the total number of locations visited by users in each dataset, which corresponds to the total size n of the sequence used as input to predictability estimates. Figure 3 shows the distributions of the total number of visited locations in both GPS and CDR datasets. Moreover, for each dataset, we show distributions for the next-cell and next-place analyses, while the latter is characterized by the removal of stationarity from the data. First of all, we note the great diversity of available data (visited locations) across users in both datasets, notably in the GPS dataset. Moreover, we note also that, for the CDR dataset, in which the temporal resolution is lower (fewer observations per time unit) and the period of observation (two weeks only) is shorter, the total number of visited locations tends to be smaller (174.7, on average) when compared to the GPS dataset (2388.8, on

average). Moreover, for both datasets, the total number of visited locations is much smaller in the next-place analysis; in other words, the removal of stationarity from our datasets results in fewer total locations, as expected.

These differences in the distributions of number of locations per user in our two datasets as well as the differences in temporal and spatial resolutions, noted in the previous section, build up different relevant scenarios of analysis for our investigation. Furthermore, as predictability estimate is based solely on the underlying dataset, having datasets with such distinct properties allows us to have a broader understanding of what affects predictability.



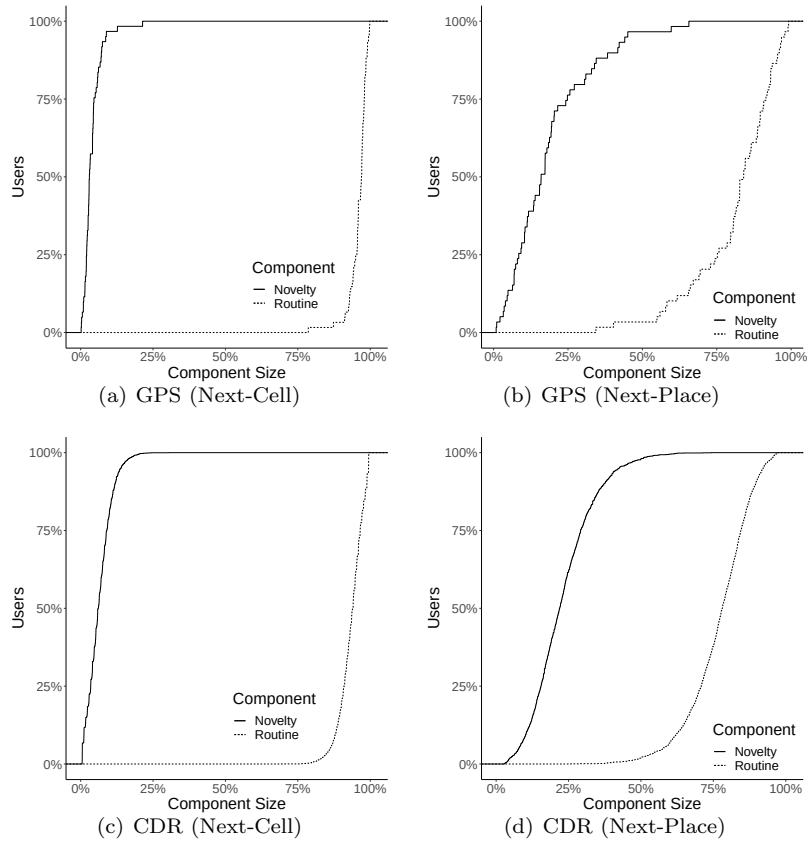
(a) GPS Dataset (b) CDR Dataset

Figure 3 Cumulative distributions of the total number of visited locations for both datasets (next-cell and next-place predictions).

Next, we analyze the novelty and routine components of a user's mobility. We do so by describing how much each of these components represent in terms of the total mobility trace of each user. Specifically, we compute, for each user in each dataset, the fractions of n, the total number of visited locations, that correspond to visits of the routine and novelty components[3], as defined in Section 3. Figure 4 shows the cumulative distributions of these fractions for both datasets, considering both next-cell and next-place analyses. Overall, the routine component dominates the locations visited, as expected. Yet, we can observe some users with a large fraction of novel visits, especially in the CDR dataset (up to 22% of all visits, in the next-place analysis). Notice also that the novelty component tends to be smaller in the next-cell prediction tasks as stationary results in a larger routine component.

Conversely, the size of the novelty component tends to be larger for next-place analyses. As such, the impact of novelty on the overall predictability will also be larger in these cases. These results corroborate previous arguments that the next-place prediction task is harder than next-cell prediction [5, 12]. As shown in the figure, we can indeed expect the next-place prediction to be harder because (i) there is no stationarity involved, so prediction is more challenging, and (ii) the size of the novelty component is larger, which also makes prediction more challenging.

[3]Note that, for a given user, these two fractions sum up to 1.



(a) GPS (Next-Cell) (b) GPS (Next-Place)

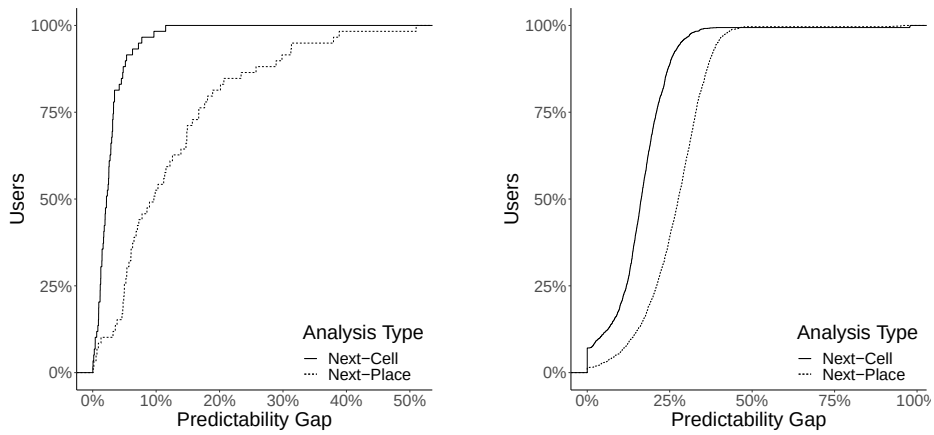
(c) CDR (Next-Cell) (d) CDR (Next-Place)

Figure 4 Cumulative distributions of the size of the routine and novelty components in our two datasets.

Focusing on the routine component, our main interest in this paper, we now assess the extent to which there is unpredictable behavior in people's routine. To that end, we apply Equation 5 to the mobility trace of each user to estimate $\Delta_{S_{routine}}$, that is the gap between the predictability of the user and the predictability of the corresponding baseline sequence (which has a completely predictable routine component). In the following we refer to this measure as simply *predictability gap*.

Figure 5 shows cumulative distributions of the predictability gap for users in both datasets and both next-cell and next-place prediction tasks. Note that the predictability gap varies considerably for users in our two datasets, showing once again great diversity of user behavior, for both prediction tasks.

Moreover, the gap tends to be smaller for next-cell prediction. For example, for next-cell prediction, the gap is on average only 2.6% and 13% in the GPS and CDR datasets, respectively. For next-place prediction, in turn, the average gap reaches 13.5% and 22.2% for the same tasks, respectively. Once again, the stationary periods make the users' routine easier to predict, which is reflected by the smaller difference between the actual predictability of the user and the predictability of the corresponding baseline sequence. As for the next-place prediction problem, because the stationary periods are removed from the users' location trace, the predictability gap is wider, indicating that the routine component is harder to predict in this case.



(a) GPS Dataset (b) CDR Dataset

Figure 5 Cumulative distributions predictability gap $\Delta_{S_{routine}}$ (Equation 5) for both datasets (next-cell and next-place predictions).

Figure 5 also shows that the predictability gap is larger in the CDR dataset, for both prediction tasks. We conjecture that this gap is due to the difference in size and temporal resolution of the traces in the two datasets. In the GPS dataset, which spans several months and has a higher temporal resolution, there is more stationarity. The CDR dataset, on the other hand, has a lower temporal resolution (i.e., 1 location every two hours, on average), and thus captures less stationarity. As mentioned, the greater amount of stationarity leads to smaller predictability gaps.

In sum, as the results discussed in this section show, despite great diversity across datasets, prediction tasks and users, there is still a fair amount of unpredictability in the routine component of human mobility. Thus, in the following, we delve further into this component by analyzing what affects the predictability of a person's routine.

5 Investigating the Predictability of a Person's Routine

In this section, we study the predictability of the routine component of human mobility. Our study is composed and driven by a series of analyses targeting both prediction tasks, namely next-cell and next-place. Recall that for the next-cell prediction task we consider the whole dataset, including stationary periods, but in the next-place prediction task we remove stationarity from the user's history of visited locations.

Concretely, we propose to use simple and easy-to-interpret *proxy metrics* that capture different factors related to a person's mobility to help us understand the predictability of the routine component of human mobility. We employ three metrics, two of which were proposed previously [**?**] and one is a novel contribution of this work. We show that these metrics can indeed be used to explain the entropy (and thus the predictability) of one's routine-related mobility by building regression models and showing that they fit reasonably well our data. By doing so, we offer valuable tools to interpret and understand the predictability of routine-related mobility.

In the following, we start by presenting the proposed proxy metrics (Section 5.1). Next, we discuss our approach to build regression models (Section **??**). Finally, we present our experimental evaluation of the use of these metrics to explain the predictability associated with routine-related mobility (Section 5.2). *Throughout this section, whenever we refer to a sequence of visited locations, we are indeed considering the extracted routine component of an original complete sequence (i.e., the subsequence with symbols in white background in Figure 2-(c)).*

5.1 Proxy Metrics

In this section, we present three proxy metrics that can be used to help us understand the predictability of the routine component of human mobility. Two of these metrics, namely regularity and stationarity, were previously used to help explain the predictability of one's mobility, considering the person's *complete mobility trace* [12]. Here, we take a different perspective, narrowing our focus to the routine component of one's mobility. Moreover, we also introduce a third metric, which together with regularity and stationarity, help us to better explain the predictability associated with that component. We note that all three metrics, though here related to the predictability associated with one's routine-related mobility, can be applied to the more general case of a complete mobility trace (as done in [**?**] for regularity and stationarity.

Before introducing our new metric, we first present the definition of regularity and stationarity [12]:

Definition 5.1 ***Regularity:*** Given a time-ordered sequence $X = (x_1, x_2, \ldots, x_n)$ of locations visited by a person, the regularity of the sequence is given by: $reg(X) = 1 - n_{unique}/n$, where n_{unique} is the number of distinct locations in X.

Definition 5.2 ***Stationarity:*** Given a time-ordered sequence $X = (x_1, x_2, \ldots, x_n)$ of locations visited by a person, the stationarity of the sequence is given by: $st(X) = st_{trans}/(n-1)$, where st_{trans} is the number of *stationary transitions* in X. A *stationary transition* is one where the previous location is equal to the next one, i.e., the location x_{i-1} is the same as x_i. Clearly, stationarity is not defined for the next-place prediction task.

We here argue that, although useful, these metrics alone do not fully explain the predictability of a person's routine. Consider, for instance, the following two sequences $X_1 = (H, W, S, H, W, S, H, W, S, H)$ and $X_2 = (H, S, W, H, W, S, H, S, W, H)$, which represent the routine components of two original mobility traces. These two sequences have the same regularity, as the total number of symbols and the number of unique symbols are the same in both of them. That is, $reg(X_1) = reg(X_2) = 0.7$. They also have the same stationarity $st(X_1) = st(X_2) = 0$, as there are no consecutive repetitions of symbols – no stationary transitions – in them. However, due to the recurring pattern HWS in X_1, X_1 is more predictable than X_2, where there is greater variation in the order of visited locations. Indeed, the entropy of X_1, computed using Equation 1, is 1.50 whereas the entropy of X_2 is 2.18.

To capture additional patterns affecting the entropy (and thus predictability) associated with the routine component of a mobility trace, we introduce another

metric, called *diversity of trajectories*. This metric helps us identify the mixture of patterns within the sequences – such as the pattern HWS in sequence X_1 and the varying patterns in X_2 – which can make them easier or harder to predict. We here define the diversity of trajectories as follows:

Definition 5.3 ***Diversity of Trajectories:*** Given a time-ordered sequence $X = (x_1, x_2, \ldots, x_n)$ of locations visited by a person, the diversity of trajectories associated with X, $div(X)$, is given by the number of distinct trajectories in X. More specifically, if we see X as a string, the diversity of trajectories is the number of distinct substrings in X.

Notice that this definition gives us an important measure of a person's mobility, and it is also related to how the entropy estimator in Equation 1 works. According to this estimator, the entropy of the sequence is proportional to the number of distinct subsequences in the original sequence. Thus, it is expected that the more diverse a person's routine is, the higher its entropy (and consequently lower predictability). Indeed, considering the aforementioned sequences X_1 and X_2, we find that $div(X_1) = 0.49$, and $div(X_2) = 0.76$.

In order to illustrate that these metrics capture important aspects of the predictability of one's routine, we compute the Spearman's rank correlation coefficient between each metric and the entropy associated with the routine component of each mobility trace in our datasets. The results are shown in Table 3, columns 6 and 10. Note the absence of correlations between stationarity and entropy for the next-place prediction, since this metric is not defined for that task.

As these results show, there is a strong correlation between each of the metrics and the entropy of one's routine: whereas both regularity and stationarity are negatively correlated with entropy, diversity of trajectories if positively correlated. Moreover, note that the latter is even more strongly correlated with entropy than regularity in all scenarios.

We also measured the pairwise correlation between the three metrics. Table 3 shows the Spearman's correlation coefficient for each pair of metric, for both datasets and prediction tasks. As we can see, there is a strong correlation between stationarity and diversity in the next-cell prediction task in both datasets, and additionally, there is a strong correlation between regularity and stationarity in the CDR dataset. We also observe some complementarity between the metrics, especially in the next-place prediction task.

In the next section, we propose to use these metrics as proxies to entropy (and thus predictability). To that end, we employ regression-based analysis to investigate the extent to which these metrics are able to explain the entropy of the routine components of individual mobility traces in our datasets.

5.2 Experimental Results

In this section, we build regression models of increasing complexity, each of which uses some of the metrics discussed in Section 5.1 as proxies to the entropy of a person's routine. We use these models to fit the entropy of a person's routine using the proxy metrics described in Section 5.1. We then compare the fitted entropy

		GPS				CDR			
		Regularity	**Stationarity**	**Diversity**	**Entropy**	**Regularity**	**Stationarity**	**Diversity**	**Entropy**
Next-Cell	**Regularity**	1	0.35	-0.17	-0.46	1	0.58	-0.66	-0.70
	Stationarity	0.35	1	-0.74	-0.78	0.58	1	-0.95	-0.94
	Diversity	-0.17	-0.74	1	0.54	-0.66	-0.95	1	0.98
	Entropy	-0.46	-0.78	0.54	1	-0.70	-0.94	0.98	1
Next-Place	**Regularity**	1	—	0.25	-0.41	1	—	-0.16	-0.53
	Stationarity	—	—	—	—	—	—	—	—
	Diversity	0.25	—	1	0.15	-0.16	—	1	0.84
	Entropy	-0.41	—	0.15	1	-0.53	—	0.84	1

Table 3 Pairwise Spearman's correlation coefficient between each proxy metric as well as between each metric and the entropy, computed for the routine component each user's mobility trace.

with the actual entropy of a person's routine and show that our metrics can indeed explain most of the variability in the entropy associated with it. We also evaluate the importance of each of the metrics to the entropy (and thus predictability) of one's routine. Collectively these results offer a fundamental knowledge to help explain the predictability associated with a person's routine and, by doing so, understand what makes one's routine more or less predictable.

We present our results first for the next-cell prediction task (Section 5.2.1) and then for the next-place prediction task (Section 5.2.2).

5.2.1 Proxy Metrics and Entropy: Next-Cell Prediction Task

In this section, we evaluate several regression models that rely on the metrics described in Section 5.1 to fit the entropy of a person's routine in the next-cell prediction problem.

Our first model uses only the two previously proposed metrics, namely, the regularity *reg* and the stationarity *st* of the input sequence) to fit the entropy of one's routine. The resulting model, called *RS model*, is given by:

$$H(X) \approx \alpha + \beta \times reg + \gamma \times st + \nu \times \mu + \epsilon, \tag{6}$$

where α is the intercept of the regression line and ϵ is the regression error, and μ is a variable that accounts for the interaction between highly correlated variables, according to Table 3, and is given by the produc of those variables.

Our second model, called *RSD model*, uses, in addition to regularity and stationarity, s the diversity of trajectories *div* as third predictor variable, leading to the following formula:

$$H(X) \approx \alpha + \beta \times reg + \gamma \times st + \delta \times div + \nu \times \mu + \epsilon, \tag{7}$$

We evaluate the quality of each model for each dataset by the *adjusted* coefficient of determination (*adjusted* R^2). As shown in Table 4, both models fit the data quite well, especially for the CDR dataset which is much larger.

Moreover, adding the diversity of trajectories as a predictor in the RSD model does not improve model accuracy, for neither dataset, as both models have the same R^2 for both datasets. This suggests that, at least for the next-cell prediction task, the diversity of trajectory plays a less important role on entropy (thus predictability),

	GPS dataset	CDR dataset
	Adjusted R^2	Adjusted R^2
RS Model	0.786	0.939
RSD Model	0.783	0.960

Table 4 Variation in entropy explained by each of the proposed regression models (adjusted R^2) for both of our datasets, in the next-cell prediction task. The RS model is the model that uses regularity and stationarity, and the RSD model is the one where diversity of trajectories is also used, along with regularity and stationarity.

and any impact it may have on it is captured by regularity and stationarity. Indeed, from Table 3, we observe that the diversity of trajectories is highly correlated with stationarity in the GPS dataset, and with both regularity and stationarity in the CDR dataset.

To better understand the role of each metric in explaining the entropy of the routine-related mobility, we zoom in on our RSD model, and analyze the coefficients of the regression. We start our investigation with the GPS dataset, for which our RSD model is shown in Equation 8:

$$H(X) \approx 6.87 - 8.44 \times reg + 1.54 \times st + 3.98 \times div \times -3.96\mu \tag{8}$$

From Table 4, we observe that, for the GPS dataset, the model with diversity of trajectories did not produce better fittings in terms of the adjusted coefficient of determination (adjusted R^2) than the simpler RS model. In fact, the p-value for the diversity of trajectories indicates that this variable is not significant (p-value $= 0.34$) for the model. We conjecture that this behavior is due to the fact that diversity of trajectories is strongly correlated with stationarity, and thus stationarity alone might be providing enough information for the model to fit the entropy of one's routine.

To illustrate the interplay between stationarity and diversity, consider a stationary period $X_s = (A, A, A, A, A, A)$ in one's routine. The diversity of trajectories for this period would be $6/21 = 0.28$, corresponding to the subsequences A, AA, AAA, $AAAA$, $AAAAA$, and $AAAAAA$, but all of those trajectories correspond to a stationary period. As the temporal resolution of our GPS dataset is high (one observation every five minutes) there are many stationary periods in it, thus highlighting this overlap in the behavior capture by stationarity and diversity.

Indeed, a simpler model (the RS model which does not use diversity of trajectories), shown in Equation 9, produced equivalent results:

$$H(X) \approx 10.2 - 7.80 \times reg - 2.42 \times st \tag{9}$$

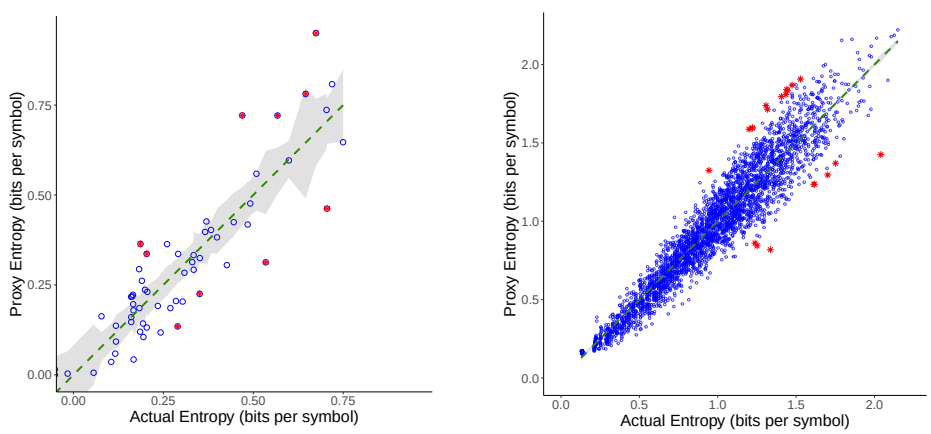
We note that the p-value for both coefficients in the model depicted by Equation 9 are significant (p-value $< 1\text{e}{-5}$). A comparison of the results of models RS and RSD suggests that, for the next-cell prediction task in the GPS dataset, a simpler model that uses only regularity and stationarity might be enough.

The situation is different for our CDR dataset. In Equation 10, we show the coefficients of our RSD model for the CDR dataset:

$$H(X) \approx -14.1 - 2.00 \times reg + 16.0 \times st + 19.4 \times div \times -19.0\mu \tag{10}$$

All of the coefficients of the model in Equation 10 are significant (p-value $<$ 1e$-$26). Furthermore, we note that our new metric, diversity of trajectories, slightly improved the performance of the model, compared to our RS model, as shown in Table 4. We conjecture that our new metric was able to improve the model for the CDR dataset because, as the period covered by the data is shorter and the temporal resolution is smaller (fewer observations per time unit), stationarity alone is not able to capture as much information as it did on the GPS dataset. Thus, our new metric provides useful information to fit the entropy of people's routine.

Our discussion so far offers an average view of how the metrics relate to entropy. We now delve further by looking at this relationship for individual users. To that end, Figure 6 shows a scatter plot (each dot is a user) of the real entropy versus the entropy estimated by the model, here called *proxy entropy*, for both datasets. These plots were built considering the complete RSD model. The closer to the diagonal the points are the more accurately the model captures the real entropy of the corresponding users. As shown in the figure, most dots (users) lie close to the diagonal in both graphs, suggesting good model fittings, but the results are better for the CDR dataset, which is consistent with the larger *adjusted* R^2. One possible reason is the larger sample (i.e., number of users) present in the CDR dataset, which favors a tighter model fitting.



(a) GPS Dataset (b) CDR Dataset

Figure 6 Entropy (in bits per symbol) fitted by the regression model (y-axis) versus actual entropy (x-axis), for both datasets in the next-cell prediction problem. The green, dashed line shows the regression line and the gray area shows the confidence interval. The red points are considered outliers and will be discussed separately. For the GPS dataset, we consider the RS model.

However, for both datasets, there are a few dots that are farther away from the diagonal, shown in red in Fig. 6. These outliers are examples of users for which the

regression model was not able to provide very accurate entropy estimates. To better understand why it happened, we manually inspected our dataset and selected 10 of these outliers from the GPS dataset, and 20 outliers from the CDR dataset for further investigation.

In the GPS dataset, we observed that most of the cases where model provided a lower entropy estimate than the actual entropy correspont to users with long (routine) mobility traces, *e.g.*, more than 1,000 locations. As mentioned, the entropy estimator shown in Equation 1 is sensitive to the size of the input traces, and produces better (lower) estimates as the size of the input sequence grows. Thus, for users with long mobility traces, our model overestimated the entropy.

Similarly, we observed cases where our model underestimated the entropy correspond to highly regular and stationary users whose mobility trace is not long enough for the entropy estimator in Equation 1 to converge, so there is a gap between the entropy (computed using Equation 1) and the fitted entropy (computed using the metrics). The same situation was observed in the CDR dataset. We manually inspected twenty users for whom the model did not perform well and found that some of them had fewer than 40 total observations after our filtering.

In order to validate our hypothesis, we added a variable n to our models and evaluated their *adjusted* coefficient of determination. We found that, for the GPS dataset, the RS model augmented with the size n of the sequence yielded an adjusted R^2 of 0.839. As for the CDR dataset, adding an extra variable n did not increase the adjusted R^2, and the extra variable was less significant than the others (p-value equal to 0.04).

Finally, we experimented with adding yet another variable, also related to one's routine, to our best models: the baseline entropy, given in Equation 4. We found that this extra variable increased the adjusted R^2 of the GPS dataset to 0.849, but did not improve the model for the CDR dataset.

5.2.2 Proxy Metrics and Entropy: Next-Place Prediction Task

We now turn our attention to the next-place prediction task. We note that the models used to fit the entropy in this prediction task are the same models discussed in Section 5.2.1, with a single modification: the only difference is that, by definition, there is no stationarity in the next-place prediction problem, therefore the stationarity term is removed from all of our three models. Additionally, we added a variable n that accounts for the size of the input sequence, as discussed in Section 5.2.1.

Because of the lack of stationarity, this prediction task is harder compared to next-cell prediction [5]. In the latter, a large portion of the accuracy in prediction comes from the fact that people tend to stay for long periods of time in the same location. Thus, models that guess that the user will be at the same location in the next time bin have a higher chance of making a correct prediction. As there is no stationarity in the next-place prediction problem, models have to cope with the difficulty of effectively guessing the next *distinct* location where the user will go.

This difficulty can be seen when we compare values of the *adjusted* R^2 in Table 4, in the previous section, to those in Table 5, which summarizes the performance of our models for the next-place prediction task. Clearly, unlike observed for the next-cell prediction, our newly proposed measure, diversity of trajectories, slighly

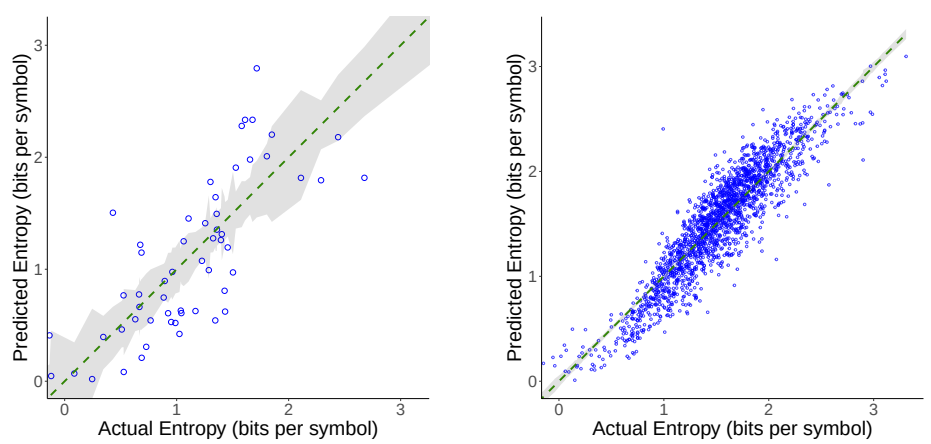
	GPS dataset	CDR dataset
	Adjusted R^2	Adjusted R^2
R Model	0.672	0.735
RN Model	0.723	0.801
RND Model	0.739	0.852
RNDB Model	0.750	0.855

Table 5 Variation in entropy explained by each of the proposed regression models (adjusted R^2) for both datasets, in the next-place prediction task. The R model is the model that uses regularity, and the RD model is the one where diversity of trajectories is also used, along with regularity. We also include results for the RDN model, which in addition to regularity and diversity also uses the size n of ones routine, and the RDNB model, which adds information about the baseline entropy of one's routine.

improves the performance of the model in the GPS dataset, and produces significant performance gains in the CDR dataset, in the current scenario. These improvements suggest that this measure is capturing important aspects of human mobility that were not captured only by regularity. We also note that the diversity of trajectories is more important for the CDR dataset, providing greater improvements to model accuracy in that case.

We further note the importance of our newly proposed metric by analyzing the coefficients of regression of the models. As shown below, though regularity has once again the largest effect on the entropy estimate, the effect of diversity of trajectories is also quite important in this task. We note that all model coefficients are statistically significant with p-value < 0.05. Additionally, as the correlation between diversity and regularity is low in the next-place prediction task, we observe greater complementarity between these metrics, justifying the performance gains.

Our results also suggest that metrics have different importance depending on the type of dataset (as evidenced by the coefficient of regression of our models). This has important implications in terms of prediction because it suggests that prediction strategies have to be tailored not only to the type of prediction task, but also to the type of dataset.



(a) GPS Dataset (b) CDR Dataset

Figure 7 Entropy (in bits per symbol) fitted by the regression model (y-axis) versus actual entropy (x-axis), for both datasets in the next-place prediction problem. The green, dashed line shows the regression line.

Figure 7 shows scatter plots of the fitted entropy of our RND model versus the real entropy for both datasets. Once again, we found that users with few observations also tend to present poor performance in terms of entropy fitting, as was also observed for the next-cell prediction in Section 5.2.1.

Thus, for both next-cell and next-place prediction, our regression models were able to capture most of the variability in people's routine, as evidenced by the R^2 of the models and the entropy fittings shown in Figure 6 and Figure 7. We also note that, for the next-place prediction problem, which is a harder prediction task than next-cell prediction, our new metric (diversity of trajectories) improves our ability of the models to explain the entropy of people's routine-related mobility (increase in adjusted R^2 values of 7.2% and 9.9%, for the GPS and CDR dataset, respectively). We also observed that adding information about one's baseline entropy can improve the performance of the model, in 1.1% and 0.3% in the GPS and CDR datasets, respectively.

We end this section by arguing for the importance of using proxy metrics to understand entropy (and predictability) in human mobility. The state-of-the-art predictability technique relies on sophisticated entropy estimates, as explained in Section 2. As previous work [12] argued, these entropy estimates are difficult to explain, in the sense that it is hard to relate an entropy value to what resulted in that value, in terms of mobility patterns. By using proxy metrics that capture specific mobility patterns and relating them to entropy, we can better understand and explain what affects the entropy of a person's mobility. In this paper, we have shown that three such metrics are enough to explain most of the variability in the entropy of a person's routine mobility.

6 Conclusions and Future Work

In this paper, we proposed to view human mobility as consisting of two components, routine and novelty, with distinct properties. We showed that this alternative view of one's mobility allows us to identify unpredictable behavior in each of these components, and we focused on analyzing and understanding what affects the predictability of one's routine. To that end, we proposed a technique to assess how much one's routine deviates from a baseline routine which is completely predictable, therefore estimating the amount of unpredictable behavior in one's routine.

Furthermore, we relied on previously proposed metrics, as well as a newly proposed one, to understand what affects the predictability of a person's routine. Our experiments show that our metrics are able to capture most of the variability in one's routine in two different prediction tasks: next-cell and next-place prediction. Our new metric, diversity of trajectories, in the next-place prediction task, was able to increase the adjusted R^2 of our regression models by 7.2% and 9.9%, on our GPS and CDR, respectively, compared to the state-of-the-art.

Our results also showed that routine behavior can be largely explained by three types of patterns: (i) stationary patterns, in which a person stays in her current location for a given time period, (ii) regular visits, in which people visit a few preferred locations with occasional visits to other places, and (iii) diversity of trajectories, in which people change the order in which they visit certain locations.

As future work, we envision exploring other metrics that may capture the variability in one's mobility patterns as well as to use our new technique to further understand mobility behavior.

Appendix A

In Section 2, we argued that an entropy estimate is the crux of predictability, and we also mentioned that the state-of-the-art predictability technique uses an entropy estimator, defined in Equation 1, according to which the entropy of an input sequence X is given by:

$$S_{real}(X) \approx \frac{n \log_2(n)}{\sum_{i=1}^{n} \Lambda_i}.$$

The term $\sum_{i=1}^{n} \Lambda_i$ records the sum of the sizes of the smallest subsequences starting at position i that do not appear before in the input sequence.

In Section 3.1, we argued that every new (previously unseen) symbol will produce a subsequence that has not appeared before in X. We also argued that that, for a sequence of size n containing $m \leq n$ distinct symbols, the contribution of such symbols to the term $\sum_{i=1}^{n} \Lambda_i$ will be exactly m.

Recall that, in Section 3.2, when describing our technique to isolate the effect of novelty on the predictability of a sequence, we moved the symbols in the novelty component to the back of the sequence. In this section, we argue that it is safe to do so because the contribution of each new (previously unseen) symbol to the term $\sum_{i=1}^{n} \Lambda_i$ does not depend on the position of such symbols in an input sequence X.

To illustrate that, we will focus on how Λ_i is computed, for a given i. Let q be the largest subsequence starting at position i that *does appear* before in X. In practice, $\Lambda_i = |q| + 1$ [25]. Suppose that we want to insert a new (previously unseen) symbol s into q and that we want to measure the impact of this new symbol on $\sum_{i=1}^{n} \Lambda_i$. There are three cases to consider:

(i) We can prepend s to q;
(ii) We can append s to q;
(iii) We can insert s somewhere inside q.

For case (i), we note that this case is equivalent to case (ii), as prepending s to q has the same effect as appending s to a subsequence p that appears immediately before q in X.

For case (ii), given that q is the largest sequence that starts at position i and appears before in X, appending s to q will result in the smallest subsequence that starts at position i and *does not* appear before in X, therefore $\Lambda_i = |q|+|s| = |q|+1$, *i.e.*, the contribution of s to Λ_i will be 1.

For case (iii), to see that the contribution s to $\sum_{i=1}^{n} \Lambda_i$ when we insert this symbol into q, it helps to break q into two subsequences r and t with $q = r + t$, where $|q| = |r| + |t|$, and r and q are subsequences of q.

Given that q has appeared before in X, both r and t will also have themselves appeared before. For instance, if we have $q = AABCAEDFBA$, and we make $r = AABCA$ and $t = EDFBA$, as both of these subsequences are part of q and q

as a whole appears before in X, both r and t must also have appeared before in the input sequence X.

The insertion of symbol s into q can be seen as concatenating s to r—prepending s to t has the same effect. Notice that r is a subsequence that appears before in X. When we append s to r, as s is a symbol that does not appear before in X, we are forming a new subsequence which is the smallest subsequence that does not appear previously in X, resulting in $\Lambda_i = |r| + 1$.

The subsequence t, which was part of q will still contribute to $\sum_{i=1}^{n} \Lambda_i$, but instead of appearing as part of Λ_i, it will be incorporated into a sequence u, appearing immediately after t, and will account to the term Λ_{i+1}, instead of Λ_i.

Thus, we have showed that no matter where the symbols in the novelty component appear in the input sequence, their contribution to $\sum_{i=1}^{n} \Lambda_i$ will be the same, therefore our strategy to move these symbols to the back of the sequence in order to focus on the routine component is a valid one.

Funding
This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001, and by the National Council for Scientific and Technological Development – Brasil (CNPq). This work is also supported by the STIC AmSud MOTIf project, and was performed in the context of the EMBRACE Associated Team of Inria.

Availability of data and materials
Due to privacy consideration and non-disclosure agreements with the data owners, we cannot make our datasets publicly available as they contain sensitive personal info and traces. We understand and appreciate the need for transparency in research, therefore we made the code used in the paper available at XXXXXXX, and for more information about access to the datasets, please send your queries to Aline Viana, at aline.viana@inria.fr.

Ethics approval and consent to participate
Data collection was approved by the data owners, and written informed consent has been obtained for all study participants.

Competing interests
The authors declare that they have no competing interests.

Author's contributions
Designed the study: DT, AV, MA, and JA. Collected and analyzed the data: DT. All authors wrote and approved the final manuscript.

Author details
[1] Inria, 91120, Palaiseau, France. [2] École Polytechnique/IPP, 91120, Palaiseau, France. [3] Federal University of Minas Gerais, 31270-901, Belo Horizonte, Brazil.

References

1. Zheng, Y., Capra, L., Wolfson, O., Yang, H.: Urban computing: Concepts, methodologies, and applications. ACM Transactions on Intelligent Systems and Technology **5** (2014)
2. Ma, S., Zheng, Y., Wolfson, O.: T-share: A large-scale dynamic taxi ridesharing service. In: Proc. IEEE International Conference on Data Engineering (2013)
3. Hasan, S., Zhan, X., Ukkusuri, S.V.: Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. In: International Workshop on Urban Computing (2013)
4. Lu, X., Wetter, E., Bharti, N., Tatem, A.J., Bengtsson, L.: Approaching the limit of predictability in human mobility. Scientific Reports (2013)
5. Cuttone, A., Lehmann, S., González, M.C.: Understanding predictability and exploration in human mobility. EPJ Data Science (2018)
6. Moon, G., Hamm, J.: 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, pp. 86–97 (2016)
7. Song, C., Qu, Z., Blumm, N., Barabási, A.-L.: Limits of predictability in human mobility. Science **327**(5968) (2010)
8. Herrera, J.C., Work, D.B., Herring, R., Ban, X., Jacobson, Q., Bayen, A.M.: Evaluation of traffic data obtained via gps-enabled mobile phones: The mobile century field experiment. Transportation Research Part C: Emerging Technologies **18**(4) (2010)
9. Beiró, M.G., Panisson, A., Tizzoni, M., Cattuto, C.: Predicting human mobility through the assimilation of social media traces into mobility models. EPJ Data Science **5**(1) (2016)

10. Smith, G., Wieser, R., Goulding, J., Barrack, D.: A refined limit on the predictability of human mobility. In: 2014 IEEE International Conference on Pervasive Computing and Communications (PerCom), pp. 88–94 (2014). IEEE
11. Teixeira, D., Alvim, M., Almeida, J.: 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, pp. 8–14 (2019). doi:10.1145/3283590.3283592. http://doi.acm.org/10.1145/3283590.3283592
12. Teixeira, D.d.C., Viana, A.C., Alvim, M.S., Almeida, J.M.: Deciphering predictability limits in human mobility. In: Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. SIGSPATIAL '19, pp. 52–61. ACM, New York, NY, USA (2019). doi:10.1145/3347146.3359093. http://doi.acm.org/10.1145/3347146.3359093
13. Hess, A., Hummel, K.A., Gansterer, W.N., Haring, G.: Data-driven human mobility modeling: A survey and engineering guidance for mobile networking. ACM Computing Surveys **48**(3) (2016)
14. Silveira, L.M., Almeida, J.M., Marques-Neto, H.T., Sarraute, C., Ziviani, A.: Mobhet: Predicting human mobility using heterogeneous data sources. Computer Communications **95** (2016)
15. Cho, E., Myers, S.A., Leskovec, J.: Friendship and mobility: User movement in location-based social networks. In: Proc. International Conference on Knowledge Discovery and Data Mining (2011)
16. Brockmann, D., Hufnagel, L., Geisel, T.: The scaling laws of human travel. Nature **439** (2006)
17. Simini, F., González, M.C., Maritan, A., Barabási, A.-L.: A universal model for mobility and migration patterns. Nature **484**(7392), 96–100 (2012)
18. Gonzalez, M.C., Hidalgo, C.A., Barabasi, A.-L.: Understanding individual human mobility patterns. Nature **453** (2008)
19. Mucceli, E., Carneiro Viana, A., Sarraute, C., Brea, J., Alvarez-Hamelin, J.I.: On the Regularity of Human Mobility. Pervasive and Mobile Computing **33**, 73–90 (2016)
20. Ikanovic, E.L., Mollgaard, A.: An alternative approach to the limits of predictability in human mobility. EPJ Data Science **6**(1), 12 (2017). doi:10.1140/epjds/s13688-017-0107-7
21. Cover, T.M., Thomas, J.A.: Elements of Information Theory. John Wiley & Sons, ??? (2012)
22. Li, M., Vitányi, P.M.B.: Handbook of theoretical computer science (vol. a), pp. 187–254. MIT Press, Cambridge, MA, USA (1990). Chap. Kolmogorov Complexity and Its Applications. http://dl.acm.org/citation.cfm?id=114872.114876
23. Feder, M., Merhav, N., Gutman, M.: Universal prediction of individual sequences. IEEE transactions on Information Theory **38**(4), 1258–1270 (1992)
24. Lempel, A., Ziv, J.: On the complexity of finite sequences. IEEE Trans. Inf. Theor. **22**(1), 75–81 (2006). doi:10.1109/TIT.1976.1055501
25. Kontoyiannis, I., Algoet, P.H., Suhov, Y.M., Wyner, A.J.: Nonparametric entropy estimation for stationary processes and random fields, with applications to english text. IEEE Transactions on Information Theory (2006)
26. Kulkarni, V., Mahalunkar, A., Garbinato, B., Kelleher, J.D.: Examining the limits of predictability of human mobility. Entropy (2019)
27. Merhav, N., Feder, M.: Universal prediction. IEEE Transactions on Information Theory **44**(6), 2124–2147 (1998)
28. Zhao, K., Khryashchev, D., Freire, J., Silva, C., Vo, H.: Predicting taxi demand at high spatial resolution: Approaching the limit of predictability. In: Proc. IEEE International Conference on Big Data (2016)
29. Li, M., Westerholt, R., Fan, H., Zipf, A.: Assessing spatiotemporal predictability of lbsn: a case study of three foursquare datasets. GeoInformatica **22**(3), 541–561 (2018). doi:10.1007/s10707-016-0279-5
30. Bagrow, J.P., Liu, X., Mitchell, L.: Information flow reveals prediction limits in online social activity. Nature Human Behaviour **3**(2), 122–128 (2019). doi:10.1038/s41562-018-0510-5
31. Xu, T., Xu, X., Hu, Y., Li, X.: An entropy-based approach for evaluating travel time predictability based on vehicle trajectory data. Entropy **19**(4) (2017). doi:10.3390/e19040165
32. Zhou, X., Zhao, Z., Li, R., Yifan Zhou, Zhang, H.: The predictability of cellular networks traffic. In: 2012 International Symposium on Communications and Information Technologies (ISCIT), pp. 973–978 (2012)
33. Ding, G., Wang, J., Wu, Q., Yao, Y., Li, R., Zhang, H., Zou, Y.: 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)
34. Jeong, J., Leconte, M., Proutiere, A.: Cluster-aided mobility predictions. In: IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications, pp. 1–9 (2016). IEEE