Predictability in Human Mobility: Interpretability, Extensions and Applications

Douglas Do Couto Teixeira

INSTITUT POLYTECHNIQUE DE PARIS

UFMG UNIVERSIDADE FEDERAL DE MINAS GERAIS

Inria

NNT : 2021IPPAX051

Thèse de doctorat

Predictability in Human Mobility: Interpretability, Extensions and Applications

Thèse de doctorat en cotutelle de l'Institut Polytechnique de Paris (France) et de l'Université Fédérale de Minas Gerais (Brésil)
préparée à Inria (Saclay) et Université Fédérale de Minas Gerais

École doctorale n°626 École doctorale de l'Institut Polytechnique de Paris (EDIPP)
Spécialité de doctorat : Informatique

Programme d'études supérieures du département d'informatique de l'Université Fédérale de Minas Gerais (UFMG)
Spécialité de doctorat : Informatique

Thèse présentée et soutenue à Belo Horizonte, Brésil, le 2 août 2021, par

DOUGLAS DO COUTO TEIXEIRA

Composition du Jury :

Katia Obraczka Professor, University of California Santa Cruz	Président, Examinateur
Razvan Stanica HDR, Maître de conférences, INSA de Lyon	Rapporteur, Examinat.
Luís Henrique Maciel Kosmalski Costa Professor, Universidade Federal do Rio de Janeiro (UFRJ)	Rapporteur, Examinat.
Antônio Alfredo Loureiro Full Professor, Universidade Federal de Minas Gerais (UFMG)	Examinateur
Jussara M. Almeida Full Professor, Universidade Federal de Minas Gerais (UFMG)	Directeur de thèse
Aline Carneiro Viana HDR, Directeur de recherche, Inria Saclay - Ile de France	Co-directeur de thèse
Pedro Olmo Vaz de Melo Professor, Universidade Federal de Minas Gerais (UFMG)	Invité

Abstract

Predicting mobility-related behavior is an important yet challenging task. On one hand, factors such as one's routine or preferences for a few favorite locations may help in predicting their mobility. On the other hand, several contextual factors, such as variations in individual preferences, weather, traffic, or even a person's social contacts, can affect mobility patterns and make its prediction significantly more challenging.

A fundamental approach to study mobility-related behavior is to assess how *predictable* such behavior is, deriving theoretical limits on the accuracy that a prediction model can achieve given a specific dataset. This approach focuses on the inherent nature and fundamental patterns of human behavior captured in that dataset, filtering out factors that depend on the specificities of the prediction method adopted.

However, the current state-of-the-art method to estimate predictability in human mobility, proposed by Song *et al.*, suffers from three major limitations. First, it has low interpretability, which makes it difficult to trace the causes of given predictability values. Second, it views one's mobility as one monolithic entity, thus preventing us from understanding the impact of one's routine on predictability. And third, it lacks flexibility to incorporate external factors which are known to help mobility prediction (i.e., contextual information). In this thesis, we revisit this state-of-the-art predictability technique, aiming at tackling these limitations as well as at providing techniques to use predictability information in practical applications.

Keywords: human mobility, prediction, entropy, predictability

Contents

Chapter 1

Introduction

Several services such as traffic control, ubiquitous computing, place recommendation, and contextual advertisement depend on our ability to predict the whereabouts and mobility patterns of individuals [72]. Despite the number of applications that benefit from it and the various strategies to tackle the problem, human mobility prediction remains an intrinsically challenging task, mainly due to the heterogeneity and complexity of human behavior. For instance, both internal factors (the person's individual preferences and personality traits that govern her decision-making processes) and external factors (hour of the day, weather, location of friends, etc.) may impact one's decision to visit a particular place.

The impact of such factors, which varies across individuals, implies that a maximum prediction accuracy of 100% may not be achievable for any given person. Yet, current state-of-the-art mobility prediction models are not evaluated taking into account such factors and how they affect the maximum accuracy that can be achieved for each person.

In contrast, Song *et al.* [58] proposed to tackle that problem by studying *predictability* in individual human mobility. Predictability refers to the maximum accuracy a prediction model can achieve when trying to foresee the next location a person will visit, given a dataset of visited locations. Studying predictability is important because, as we will explain later, it offers fundamental insights to understand patterns of human mobility, to improve prediction models, and to enhance location-based systems.

Our goal in this thesis is to investigate the state-of-the-art technique for estimating predictability, understand how it works, show some of its shortcomings, and then propose ways to address these shortcomings.

1.1 Overview of predictability

The state-of-the-art technique for estimating predictability was proposed by Song *et al.* [58], and it exploits the concept of entropy as a measure of how complex (or, inversely, how predictable) a person's mobility patterns are.

In a nutshell, this technique estimates the entropy of the person's mobility trace, and subsequently uses this value to obtain a predictability estimate in the range $[0, 1]$, with 0 meaning completely unpredictable, and 1 meaning totally predictable. Thus, if a person's predictability is 0.8, it means that 20% of her mobility is considered to be unpredictable, according to Song *et al.*'s technique.

The focus of predictability analysis is to detect patterns as they appear in the data. In other words, this approach abstracts away from any specific prediction strategy and concentrates instead on the inherent nature and fundamental patterns of human behavior, as captured by the available data. Unlike particular comparisons of alternative prediction models on different datasets, Song *et al.*'s approach is 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 and has direct applications to mobility prediction.

For instance, Song *et al.*'s predictability technique can be used to solve two important problems in the evaluation of mobility prediction models. Usually, such models are evaluated against a maximum accuracy of 100%, but this approach has two problems: (i) given the complexity and heterogeneity of human behavior, a prediction accuracy of 100% may not be achievable for any given person, and (ii) different people exhibit different behavior, which means that prediction models should be evaluated on a *per-user* basis, instead of treating every person as equal. As mentioned, using each person's predictability to evaluate the performance of prediction models solves both of these problems, as every person's predictability is (likely) to be lower than an ideal value of 100% and tailored to that person's behavior.

Given the alleged benefits of using predictability to evaluate prediction models and the more fundamental nature of the technique, one may be skeptical about whether predictability values are actually attainable or just another ideal target to aim for. Although valid, this skepticism is unwarranted, given that Song *et al.*'s predictability values can indeed be achieved, as shown in previous work [39].

Predictability also has many (still unexplored) applications of its own. For instance, it can be used as a tool to assess the confidence in predictions, which is useful in the analysis of highly unpredictable individuals, particularly when a misprediction has a high cost. Let's say, for example, that the predictability estimated for a user is

40%. Then, a model will mispredict the user's next location *at least* 60% of the time. That is, the confidence on the result of a prediction model can be considered low as the risk of mispredictions is reasonably high.

Predictability can also be employed in outlier identification: in a particular dataset, users who exhibit levels of predictability very different from the rest are likely to be outliers, and therefore may deserve special attention. In a disease spreading scenario, less predictable users may also be of interest since they tend to visit a large variety of distinct locations, and therefore might exhibit higher risk of infection.

Predictability can also be used as a baseline to decide whether the cost of more complex models is worth in a specific scenario. The motivation for this is that the maximum accuracy given by predictability is actually achievable by simpler models (such as Markov-based models). Thus, it is only worth using a more complex model (such as a neural network) if it offers higher accuracy. Additionally, in certain services such as place recommendation, predictability can be used as a *measure of the susceptibility* of less predictable users to novelty and diversity. In Chapter 7 we discuss how some of these practical applications of predictability could be implemented and deployed.

1.2 Shortcomings of the state-of-the-art predictability technique

Song *et al.*'s predictability technique is an invaluable tool for both theoretical and practical studies, and it has been employed in several fields [71, 73, 16, 5, 39]. It does, however, suffer from three shortcomings, which we discuss below and address in this thesis.

1.2.1 Low interpretability

Motivation The first shortcoming of Song *et al.*'s technique is its low interpretability. As mentioned, Song *et al.*'s technique is able to capture patterns in a person's mobility and translate them to a value, *i.e.*, the more predictable the person's mobility the higher their predictability value, and vice-versa. However, the knowledge of a person's predictability value does not give us insight into what resulted in that particular value.

This difficulty comes from how Song *et al.*'s technique works. This technique is based on a sophisticated compression algorithm [35], which approximates the entropy of the sequence by the compression-rate of the input data. It is thus difficult to keep track of what the compression algorithm is really capturing in terms of mobility patterns.

The end result is that it is hard to understand what types of mobility patterns make one's predictability higher or lower, that is, what makes one's mobility easier or harder to predict. While many previous studies relied on Song *et al.*'s technique, to our knowledge, a thorough approach to understand and interpret predictability values is still lacking. In light of the previous observations, the first research question this thesis aims to address is the following.

Research Question 1 (RQ1) *Would it be possible to trace a given predictability value back to its causes,* i.e.*, to interpret/explain predictability values and to understand what makes a person's predictability higher or lower?*

Addressing RQ1 Our first goal is to tackle the low interpretability of Song *et al.*'s technique, by designing more robust, flexible and easy-to-interpret measures of predictability.

Ideally, a theoretical measure, such as predictability, should offer insights into aspects of human mobility that have not been revealed before. It should help uncover new patterns of mobility-related behavior that could be used to improve prediction strategies and drive the design of more robust approaches. Yet, if we do not understand the patterns it captures, it becomes harder to build prediction models that leverage such patterns. Understanding what affects a person's predictability can shed light into new avenues of improvement for mobility prediction.

One of the approaches to understand what affects predictability is to try to untangle mobility patterns from the output of the compression algorithm used by Song *et al.*'s technique. Here we take a different approach: we propose the use of simple proxy metrics that (i) are directly related to entropy/predictability, and (ii) allow us to to capture mobility patterns in a more intuitive and easy to interpret way, thus allowing us to understand what makes a person's mobility easier or harder to predict.

1.2.2 Viewing human mobility as a single component

Motivation The second shortcoming in Song *et al.*'s technique is the difficulty to use it to study different components of an individual's mobility. Although previous work [57, 48, 3] modeled individual human mobility as consisting of two types of visits (explorations and preferential returns), Song *et al.*'s work as well as subsequent studies derived from it [39, 15, 56, 42, 60, 62] viewed individual human mobility as a whole, monolithic entity. It is therefore difficult to use Song *et al.*'s technique to study different

components of human mobility separately. Thus, the second research question this thesis aims to address is the following.

Research Question 2 (RQ2) *Would it be possible to use Song* et al.*'s technique to study the predictability of different components of individual human mobility?*

Addressing RQ2: In addition to trying to understand predictability as a whole, we also take steps towards understanding the predictability of different components of an individual's mobility. As mentioned, Song *et al.*'s technique, as originally proposed, does not allow us to do that, as it views human mobility as a single entity.

To tackle that, we propose a way to study predictability by splitting one's mobility into two distinct components: routine and novelty, where the novelty component consists of visits to a place for the first time, and the routine component consists of every visit that does not belong to the novelty component.

The motivation for these two components is that visits that occur in each of them have different properties: one's routine component is less influenced by some external factors than one's novelty component. For instance, one may go to work regardless of the weather, but if it rains, one may decide to stay home instead of going to a new restaurant. These different properties suggest that novelty and routine should be analyzed and evaluated separately. We hypothesize that, by dividing human mobility into these two components, the mobility patterns that affect each of them should become clearer, compared to looking at human mobility as one, monolithic entity. Specifically, we focus on studying predictability of the routine component, as that can lead to new insights for prediction strategies that rely on the history of visited locations.

1.2.3 Lack of flexibility to incorporate contextual information

Motivation The third shortcoming in Song *et al.*'s predictability technique is the fact that it does not allow for the use of external information which is known to help mobility prediction (i.e., contextual information). As we will discuss in Section 2.1.5, mobility-related behavior may be influenced by several factors (daily or weekly routine, traffic conditions, weather, and so on), some of which have been shown to help prediction accuracy [15]. However, Song *et al.*'s technique takes as input only the person's mobility trace, completely disregarding these factors.

Although previous work has argued for the benefit of exploiting contextual information on predictability [15], to our knowledge, no one has shown how to incorporate

it into predictability estimates nor effectively quantified its impact on predictability. Thus, the third research question this thesis aims to address is the following.

Research Question 3 (RQ3) *Would it be possible to take* contextual information *into account when using Song* et al.*'s technique, as well as to investigate the pros/cons related to the use of contextual information jointly with predictability?*

Addressing RQ3 Song et al.'s technique [58] relies on a person's history of visited locations to estimate their predictability, and, by doing so, it is able to capture frequently visited locations and trajectories. Fundamentally, the introduction of contextual information leads to entropy measures based on a joint probability distribution of two random variables, namely *location* and *context*. However, as we discuss in Chapter 5, Song *et al.*'s technique does not explicitly consider a probability distribution as a component of its algorithm, so adding context to it becomes a challenge.

In this thesis, we tackle that by devising ways to estimate predictability with contextual information by using different entropy estimators—which allow for the use of contextual information—, and also by devising strategies to incorporate such information into compression-based entropy estimators. Additionally, we study the impact of context on predictability when these strategies are employed.

1.3 Contributions of this thesis

Building on the motivation and directions presented in the previous section, we describe below our contributions towards addressing each of our RQs as well as list the results and publications we obtained for each RQ.

RQ1 Our results show that most of the variability in an individual's predictability can be explained by simple, easy to interpret proxy metrics: stationarity, regularity, and diversity. Our decision to use metrics that capture a person's predictability was motivated by the fact that Song *et al.*'s predictability technique is based on a sophisticated compression algorithm, which makes it hard to look at the output of the algorithm and reason what caused such output. We validated our metrics by proposing regression models that use them as proxies of one's predictability. Our results, which encompass two prediction tasks (described in Section 2.1.3), show that our proposed metrics are able to capture up to 93.5% of the variability of a person's predictability.

RQ2 Towards addressing RQ2, we devised a novel technique to isolate an individual's mobility into two components (called novelty and routine), so as to estimate the impact of each of these components on predictability. We then used our technique to estimate the impact of each component on predictability, and to zoom in on the routine component, aiming at understanding what affects the predictability of one's routine. Our results, which extend and deepen the investigation started in RQ1, show that most of the variability in the predictability of one's routine can be explained by the amount of distinct locations one visits (regular behavior), the amount of time one spends at preferred locations (stationary behavior), and the order in which one visits certain sequences of locations (diverse behavior). Additionally, we show that these three types of behavior account for up to 96% of the variability in one's routine.

RQ3 We obtained two important results with respect to RQ3. First, we show that by using different entropy estimators than the one used by Song *et al.*, it is possible to incorporate contextual information into predictability estimates. We propose two novel techniques to incorporate contextual information directly into the entropy estimator used in Song *et al.*'s work. Second, we evaluate the impact of contextual information both when different entropy estimators are used, as well as when our two techniques are employed. Our results show that contextual information does not always lead to higher predictability, and we provide a few hypotheses to explain these results. Despite limitations related to mobility sequence sizes and context availability, our findings in this RQ open up different avenues for research in the topic of predictability.

The results listed above (and detailed in Chapters 4, and 5, and 6) are summarized in the following publications:

- On Estimating the Predictability of Human Mobility: The Role of Routine. Douglas Teixeira, Jussara M. Almeida, Aline C. Viana. EPJ Data Science. *Under review.* (**RQ2**).

- The Impact of Stationarity, Regularity, and Context on the Predictability of Individual Human Mobility. Douglas Teixeira, Aline C. Viana, Mário S. Alvim, Jussara M. Almeida. ACM Transactions on Spatial Algorithms and Systems. 2021. (**RQ1 & RQ3**).

- Deciphering Predictability Limits in Human Mobility. Douglas Teixeira, Aline C. Viana, Mário S. Alvim, Jussara M. Almeida. ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. 2019. **Best paper candidate.** (**RQ1 & RQ3**).

- On the Predictability of a User's Next Check-in Using Data from Different Social Networks. Douglas Teixeira, Mário S. Alvim, Jussara M. Almeida. ACM SIGSPATIAL Workshop on Prediction of Human Mobility. 2018. (**RQ3**).

- An Empirical Study of Human Mobility Patterns. Douglas Teixeira, Jussara M. Almeida. Brazilian Symposium on Computer Networks and Distributed Systems (SBRC). 2018. (**RQ1**).

We also obtained other results, not directly related (but still relevant) to this thesis:

- On Car-Sharing Usage Prediction with Open Socio-Demographic Data. Michele Cocca, Douglas Teixeira, Luca Vassio, Marco Mellia, Jussara M Almeida, Ana Paula Couto da Silva. Electronics. 2020.

1.4 Organization of this document

The rest of this document is organized as follows. In Chapter 2 we provide the necessary background for this thesis, discussing human mobility, the two prediction tasks we target, explaining the theoretical foundations of Song *et al.*'s technique, and positioning our work with respect to the literature on predictability. Chapter 3 explains and characterizes the datasets used in our experiments. In Chapters 4, 5 and 6 we discuss the results we obtained for each of our research questions (RQs). Chapter 7 summarizes our contributions and discusses future directions for predictability studies.

Chapter 2

Background

In this chapter, we position this thesis with respect to the literature and provide the necessary background for our work. We start by making a distinction between mobility modeling and mobility prediction. We then move on to define the the *individual human mobility prediction* problem (Section 2.1.2), as predictability is a measure of the maximum accuracy that a model can obtain when predicting the next location someone will visit. Subsequently, we explain the prediction tasks—ways to frame the mobility prediction problem—targeted in this study (Section 2.1.3).

In order to describe how the mobility prediction problem is addressed in practice, we provide a brief overview of several types of prediction strategies (Section 2.1.4) and discuss factors that play a role on human mobility (Section 2.1.5). Subsequently, we provide the necessary background on predictability (Section 2.2.1), describing its theoretical foundations, and then position our work with respect to the literature on predictability (Section 2.2.2), highlighting the gaps in the literature and discussing how our work fills some of those gaps.

2.1 Mobility Modeling and Prediction

In this section, we discuss the difference between mobility modeling and mobility prediction, and how prediction strategies rely on mobility modeling to infer people's mobility patterns. We start by explaining mobility modeling (here called *explanatory models*), in Section 2.1.1. We briefly mention patterns of human mobility that have been used to explain people's movements, as well as several ways to categorize human mobility in terms of mobility patterns. We then move on to discussing mobility prediction (in Section 2.1.2), and how it relates to our work.

2.1.1 Mobility Modeling

In human mobility, there are *explanatory models* and *predictive models*. Explanatory models try to derive statistical properties of human mobility, while predictive models rely on existing mobility patterns (properties of human mobility) to predict the whereabouts of individuals.

There are a number of studies that investigate explanatory models of human mobility. For instance, Brockman *et al.* [8] try to model the dynamic spatial redistribution of individuals as scale-free random walks known as Lévy flights. Gonzales *et al.* [21] argue that human trajectories show a high degree of temporal and spatial regularity, and characterize the mobility of individuals in terms of their radius of gyration, and the probability of a person returning to a previously visited location. Other examples of explanatory models are the gravity model and the radiation model [54], which try to explain the migration patterns of individuals. Wang *et al.* [67] nicely summarize classic explanatory models and what properties of human mobility they capture.

Karamshuk *et al.* [32] also survey and categorize several statistical properties of human mobility. These properties are divided into three categories: temporal, spatial, and social. Temporal properties include the frequency with which people visit a given location, their probability of returning to a previously visited location, or how much time they spend in each visited location [21]. Spatial properties include how close to Lèvy flight are human movements [8], the radius of gyration [21] of people's movements. Social properties include whether the users are considered isolated [21], in groups [70], or whether contacts between users are considered [13].

Explanatory models have been categorized in several other dimensions. For instance, Asgari *et al.* [4] survey data collection techniques and mobility patterns of individuals, classifying previous studies into three major types: trajectory-based, dynamic proximity networks, and flow on networks. Other studies adopt different categorizations, such as the one presented by Treurniet *et al.* [64], which focuses on the elements (spatial constraints, pause time, motion, etc.) of human mobility that are captured by each model. Yet another categorization is proposed by Hess *et al.* [27] to classify mobility models in terms of features and general strategies (modeling view, evaluation method, and so on) adopted by each model.

Explanatory models are important for mobility prediction because predictive models often rely on statistical properties of human mobility to make predictions. Having made the distinction between explanatory and predictive models, we now proceed to explain human mobility prediction, the prediction tasks we investigate in this study, and several types of predictive models.

2.1.2 Mobility prediction

Human mobility prediction is a research topic with broad and important applications in areas such as urban planning, traffic engineering, epidemiology, recommender systems, and advertisement, to name a few [72, 40, 25]. For instance, knowing the amount of people that routinely go to certain regions in a city can be used for better traffic forecast and urban planning, and knowing a person's next location can be used to offer better route suggestions, to recommend places, and to provide location-aware advertising.

From these applications, one can infer that there are two types of mobility prediction: volume (aggregate) and individual prediction. Aggregate prediction is a coarse-grained approach: it works at a population level and focuses in predicting the direction in which groups of people will flow. Examples of studies in this area include the work of Brockman *et al.* [8], in which the authors investigate human travelling by analysing the circulation of bank notes in the United States, and the work of Simini *et al.* [55], which investigates migration flows between regions using the radiation model.

In individual prediction, the goal is to provide a fine-grained approach to human mobility by focusing in forecasting the whereabouts of individuals. An example of individual mobility prediction is the work of Gonzales *et al.* [21], which shows that human trajectories exhibit spatiotemporal regularities, with frequent visits to a few preferred locations interspersed with occasional visits to other locations. Throughout this document, unless otherwise noted, whenever we mention mobility prediction, we are referring to individual human mobility prediction, which is the focus of our study.

The *individual human mobility prediction problem* can be defined as follows.

Definition 2.1.1. Given a time-ordered sequence of locations $X = (x_1, x_2, \ldots, x_{n-1})$ that a person visited in the past, we wish to predict the next location $x_n \in X$.

Notice that the definition of the *mobility prediction problem* deals with symbols in a sequence, but location data is usually collected as latitude/longitude pairs. Predicting the exact latitude and longitude of users is rather challenging so, to simplify the problem, the geographical area can be divided into a grid of square cells, and latitude/longitude coordinates can be converted to cell identifiers. Thus, each x_i in X becomes a unique identifier of a cell in the grid. The mobility prediction problem is thus to guess the next symbol (identifier) x_n in X.

As mentioned, our focus in this doctoral thesis is on individual human mobility, regardless of the mode of transportation the individual chooses to move among places. As long as his or her mobility trace can be represented by a sequence of symbols in a given alphabet, as described in Definition 2.1.1.

2.1.3 Prediction tasks

The *mobility prediction problem* can be rendered under different prediction tasks, depending on the properties of x_n. In this study, we will focus on two particular prediction tasks, namely *next-cell* and *next-place* prediction [28, 15].

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 2.1.2. ***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 2.1.3. ***Next-place prediction:*** Predict the next location $x_n \in X$, where $x_n \neq x_{n-1}$. Notice that here we wish 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, while performing predictions, one would ignore every next location that is equal to the previous one. 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 thesis, whenever we refer to a particular dataset for the next-place prediction task, as we discuss in Section 3.2, we are referring to the dataset after we filter out stationarity. Notice that, while filtering out stationary periods, we remove symbols from the sequences, therefore producing smaller sequences compared to the next-cell prediction task. As detailed in Chapter 5, the lack of stationarity and the smaller size of the sequences makes this prediction task harder than next-cell prediction.

2.1.4 Prediction strategies

The study of human mobility prediction has received considerable attention in the literature and many previous studies have proposed prediction strategies by employing a plethora of different techniques. We note that a comprehensive review of the several types of mobility prediction models can be found elsewhere [67, 64]. For the purposes

of this thesis, we will provide a brief description of studies that are representative of some of the main strategies used in mobility prediction. We will also make a distinction between the type of model for which Song *et al.*'s technique applies and those in which it may not apply.

Compression-based models Compression-based strategies try to infer sequences of commonly visited locations in terms the sequences that appear often in an input, and can in turn be used to compress the input sequence. This idea is used by Pulliyakode *et al.* [49] to propose a compression-based model to predict people's future location in small datasets. Song *et al.* [59] compare the performance of Markov-based models with compression based models, finding that Markov models outperform the compression-based strategies they evaluated. Other compression-based strategies rely on more sophisticated algorithms, such as the Lempel-Ziv [35] algorithm, to approximate a k-order Markov predictor.

Markov models In this type of prediction strategy, transitions between locations are modeled as a Markov chain, and the frequencies of each transition are used to predict where the person will go next. Markov models have been used in conjunction with predictability since Song *et al.*'s work [58], in which the authors build Markov models to try to reach the maximum accuracy obtained by their technique. Subsequently, Xin Lu *et al.* [39] use Markov chains to predict the next location of individuals based on their current and past locations, and show that Markov models can indeed reach the predictability limits devised in Song *et al.*'s work. Libo *et al.* [59] also evaluate Markov-based methods for location prediction and compare their performance with other types of models.

Graph-based models This type of model views locations as nodes and transitions between locations as edges in a graph. An example of the use of this strategy for mobility prediction is Dong *et al.*'s work [18], in which the authors propose a structure called *leap graph*, where an edge (or a leap) corresponds to actual user mobility. They also evaluate a Markov-based model that uses the leap graph to predict users' mobility in a network. They show that their model can substantially improve the performance of content prefetching and base station selection during handover. Silveira *et al.* [53] also propose alternative prediction strategies and compare them to leap graph and another strategy called SMOOTH [43] on heterogeneous data sources coming from social networks and mobile phone usage, showing that their proposed strategies perform better and are more robust than the baselines. Terroso *et al.* [63] also propose a graph-

based model that requires no prior training and evaluate its performance on geo-tagged data from a social network.

Time-series models Time-series models have also been used to perform mobility prediction. The reasoning behind these models is to modify the input time-series, preserving some its properties, while making it easier to predict the observations (locations) in the sequence [67]. Several strategies can be used to obtain a stationary time-series. For instance, auto-regressive (AR) models [65] identify periodicity in a time series in order to make it more regular. Other strategies are based on computing a moving average (MA) of the input time-series [67]. More sophisticated approaches, such as Autoregressive Moving Average (ARMA) and Autoregressive Integrated Moving Average (ARIMA) combine these two strategies in order to obtain more regular, stable time-series on which patterns can be inferred. Wang *et al.* [66] employ these strategies to build a higher-order time series model to predict future visited locations.

Machine learning models Machine-learning models have several uses in mobility prediction. For instance, Cuttone *et al.* [15] use logistic regression models that rely on the users' current and past locations as well as contextual information (hour of the day, day of the week, the user's home location, etc.) to predict their next location. Moon *et al.* [42] use a Long Short-Term Memory (LSTM) network-based model to predict next locations a person will visit in a period of up to six hours. Ozturk *et al.* [46] use deep learning for mobility prediction in the context of 5G networks, and Ghouti *et al.* [20] use deep neural networks to propose improvements in the quality of service in Mobile Ad-Hoc Networks.

Universal versus non-universal predictors To put it simply, a *universal predictor* is a prediction model that does not require prior training and learns patterns on-the-fly as it processes the input. An example will make things clearer. Suppose, for instance, that one wishes to predict the next outcome in a sequence coin tosses. Because it is expected that the number of 0s and 1s in an unbiased coin tend to be roughly the same as the number of tosses goes to infinity, if the number of 0s is less than the number of 1s so far, the model predicts that the outcome will be 0. Otherwise it predicts 1. As the number of 0s and 1s changes at each coin toss, the model adjusts these numbers on-the-fly and makes predictions based on which of them is lower. In the case of human mobility prediction, one could employ a very simple model that always predicts the user's next location as the location they visited more frequently in the past. As time

goes on, the frequency with which the user visits locations may change, and the model updates its predictions accordingly.

Conversely, a *non-universal predictor* is a model that *does* require prior training. For this type of predictor to work, several parameters have to be tuned, and to tune a model's parameters, we feed it with real data, compare its predictions with the actual outcome (present in the input data) and then adjust its parameters so as to fit the desired output to the corresponding input data. Machine Learning (*e.g.*, linear regression, logistic regression, decision trees, neural networks, etc.) are representatives of this type of model.

In this study, we will focus on universal predictors because Song *et al.*'s technique is valid only for this type of model, as we explain in Section 2.2.1. It is important to note, however, that albeit simpler, this type of model presents remarkable performance in mobility prediction (*e.g.*, Xin Lu *et al.* [39] report accuracy values of up to 95% in predicting people's next locations). These models also have the advantage of being more interpretable and less computationally intensive than other approaches such as neural networks.

2.1.5 Factors that influence mobility prediction

A lot of factors play a role on mobility prediction. It is important to note that all of these factors are captured by the data, which is central to mobility prediction and predictability. We defer to Section 3.1 the discussion of the overall influence of different data types on mobility prediction, and in this section we discuss other factors such as the spatiotemporal resolution of the data and contextual information. While in this section we provide a qualitative discussion of the impact of these factors on predictability, in Chapter 4 and Chapter 6, we actually quantify their effect on predictability.

As explained in Section 2.1.3, in order to obtain a sequence of identifiers (instead of dealing with latitude/longitude coordinates), we tessellate a grid onto the geographical area, thus dividing it into square cells of a given size. As a result, the choice of the size of the cells influences mobility prediction. If the area is broken into a grid of larger cells, most of the user's activity tends to be confined within fewer cells, with two opposing effects. On one hand prediction accuracy, increases, since it becomes relatively easier to correctly infer which cell the user will visit next. On the other hand, prediction *utility* degrades, since the bigger the cell the user is predicted to visit, the less informative the corresponding prediction is. In the extreme case of a grid with a single cell, prediction is always trivially correct but it is also of little use. Hence, by adopting grids with higher granularity (i.e., smaller cells), we increase prediction

utility, but at the possible cost of hurting accuracy. *What are the trade-offs between these two dimensions?*

Temporal resolution has a similar effect on mobility prediction. It has been shown that stationary periods, i.e., periods in which the individual stays in the same location, lead to higher prediction accuracy [15, 62, 61], depending on the prediction task. If the temporal resolution is high (more observations per time unit), there will be more observations in the same location, which will lead to longer stationary periods. If, however, the temporal resolution is lower, the more likely the person is to change locations between observations.

In general, we expect prediction accuracy to be inversely proportional to spatial resolution (the smaller the cell size, the lower the accuracy of models), and directly proportional to the temporal resolution (the more observations per time unit, the higher the accuracy of models).

There are also other factors that influence mobility prediction. For instance, it has been shown that *context* can play an important role in helping to forecast the whereabouts of individuals [15, 62]. As an example, depending on whether it rains in a particular day, one may decide to stay at home and watch a movie, or go out to visit a park. Similarly, contextual information associated with one's *friends* (*e.g.*, their current locations) has been shown to help predict one's next location [13, 53]. Thus, whenever possible, several types of *contextual information* (day of the week, hour of the day, the weather, the location of a person's friends, etc.) should be taken into account when predicting someone's next location.

The mobility prediction problem is thus affected by several factors. Predictability, which is a measure of the maximum accuracy that a given prediction strategy can achieve on a particular dataset, is also influenced by these factors.

2.2 Predictability in human mobility

In this section, we present the theoretical foundations of how to estimate predictability in human mobility. In particular, we revisit Song *et al.*'s technique in the light of more fundamental theoretical concepts and well-established measures of complexity. To the best of our knowledge, no previous work has summarized the roots of predictability—tracing the equivalences between entropy and compressibility and showing why entropy is a good approximation for the complexity of a sequence of symbols—as we do here. We believe this effort brings insights into how predictability estimate works from a much more fundamental perspective, which is valuable to understand its challenges

and improve on it.

We end this section with a discussion of relevant literature on predictability, positioning our study with respect to the literature gaps we aim to address.

2.2.1 Foundations of predictability

In 2010, Song *et al.*. [58] proposed a technique to estimate the predictability of a sequence of visited locations by relating predictability to the entropy of the sequence. Concretely, estimating the predictability of an input sequence of locations is a multi-step process: the entropy of the sequence is first estimated and then used, together with the size of the sequence, to compute a value π_{max}, referred to as the predictability associated with the input sequence. This process is illustrated in Figure 2.1.

$$X = (x_1, x_2, x_3, \ldots, x_{n-1}) \quad \rightarrow \boxed{\text{Entropy Estimate}} \rightarrow \cdots \rightarrow \Pi_{max}$$

Figure 2.1: The high-level process of estimating predictability.

As we can see, estimating the entropy of the sequence of locations is the crux of the predictability technique. Thus, there is an assumption that entropy is related to complexity, which in turn is related to predictability.[1] In other words, 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 (randomness) is related to entropy.

There is another conceptual leap we need to make in order to understand Song *et al.*'s technique, which is the fact that entropy is related to *compressibility*, i.e., to how *compressible* the input sequence is [19, 35]. For instance, sequences with many repeated symbols are highly compressible. Intuitively, if a sequence has many repeated symbols it is highly compressible and thus 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.

Furthermore, the entropy, which can be defined as the average uncertainty in the outcomes of a random variable [14], is a good approximation for the complexity of the sequence because the entropy of a sequence of symbols is a lower bound on its compressibility [14, 36]. 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.*

[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.

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 increasingly accurate 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 (dependencies among visits). This third, more robust entropy estimator was originally proposed by Kontoyiannis *et al.* [33]. According to its definition, the entropy S_{real} of an input sequence of locations X of size n can be approximated by:

$$S_{real} \approx \frac{n \log_2(n)}{\sum_{i \leq n} \Lambda_i}, \tag{2.1}$$

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 redundancy, and therefore has low entropy, i.e., it is more predictable. Throughout the rest of this thesis, 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 2.1.

To illustrate how Equation 2.1 works, consider the following example. Suppose that a person visits a sequence of locations represented by the following sequence of symbols: $X = (H, W, H, W, S, H, W, H, W, R)$. In Table 2.1, we illustrate how the input sequence X is processed and how each Λ_i, *i.e.*, the length of the shortest sub-sequence starting at position i which does not appear from 1 to $i-1$ in sequence X, is obtained. We illustrate the process of computing each Λ_i by showing (i) the value of i, (ii) the sub-sequence $X_{1:i}$ where the symbols from 1 to $i-1$ appear in black and the rest of the symbols are shown in gray, (iii) L_i, the shortest shortest sub-sequence starting at position i which does not appear from 1 to $i-1$ in sequence X, and (iiii) Λ_i, which is simply $|L_i|$.

The sub-sequences L_i are obtained via pattern-matching in the following way. For a given value of i, we first obtain the longest sub-sequence l_i that *does* appear from 1 to $i-1$ in sequence X. Then, L_i is just l_i followed by the next symbol in X. In Table 2.1, the sub-sequences l_i are the underlined part of the sequences L_i.

i	$X_{[1:i]}$	L_i	Λ_i
1	HWHWSHWHWR	H	1
2	HWHWSHWHWR	W	1
3	HWHWSHWHWR	HWS	3
4	HWHWSHWHWR	WS	2
5	HWHWSHWHWR	S	1
6	HWHWSHWHWR	HWHWR	5
7	HWHWSHWHWR	WHWR	4
8	HWHWSHWHWR	HWR	3
9	HWHWSHWHWR	WR	2
10	HWHWSHWHWR	R	1

Table 2.1: An example illustrating the innerworkings of Equation 2.1 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 sub-sequences that start at position i and do not appear from 1 to $i-1$, and Λ_i is the length of each L_i.

Applying Equation 2.1 to the example in Table 2.1, we have

$$S_{real} \approx \frac{n \log_2(n)}{\sum_{i \leq n} \Lambda_i} = \frac{10 \log_2(10)}{23} = 1.44, \tag{2.2}$$

where $n = |X| = 10$ is the size of the input sequence X, and $\sum_{i \leq n} \Lambda_i = 23$ is the sum of the values in the last column of Table 2.1.

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.*

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 on prediction accuracy 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 [19, 35, 41]. 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. Examples of this type of predictor are supervised machine learning algorithms (e.g., neural networks) [20, 46, 34] that are specialized to a particular dataset.

Using Fano's Inequality [14], which provides a way to compute error bounds for certain phenomena, Song et al.. derived a formula to compute the *predictability* of a sequence. This formula is based on the intuition that, if a user with entropy S moves between N locations, her predictability Π_{max}, which is a function of S and N, is given by:

$$S = -H(\Pi_{max}) + (1 - \Pi_{max}) \log(N - 1), \tag{2.3}$$

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 [58, 56, 71]. In particular, Smith *et al.* [56] provided a detailed, thorough derivation of the formula above.

We illustrate the computation of Π_{max} using our previously introduced toy example. For an entropy value S and a given number of *distinct* locations N, we have to implicitly solve Equation 2.3 to obtain Π_{max}. Plugging the formula for $H(\Pi_{max})$ into Equation 2.3 and applying it to our toy example, where we have $S = 1.44$ and $N = 4$, we obtain

$$-\Pi_{max} \log_2(\Pi_{max}) - (1-\Pi_{max}) \log_2(1-\Pi_{max}) + (1-\Pi_{max}) \log(4-1) - 1.44 = 0, \tag{2.4}$$

which gives us $\Pi_{max} \approx 0.669$.[2]

[2]To see how Equation 2.4 is solved for our toy example, simply type `solve 0 = -p log[2, p] - (1 - p) log[2, 1 - p] + (1 - p) log[2, 4 - 1] - 1.44 for p` in a computational engine such as Wolfram Alpha (`https://www.wolframalpha.com/`).

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

Definition 2.2.1. 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 $x_n \in X$.

2.2.2 Literature on predictability

Song *et al.*'s 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.* [39] 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.

Smith *et al.* [56] 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.* [28] 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.* [15] 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.* [37]. build on Song *et al.*'s technique to assess spatiotemporal predictability in location-based social networks. Bagrow *et al.* [5] 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.* [71] 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 in scenarios such as travel time estimates [69], cellular network traffic [73], and radio spectrum state dynamics [16].

Previous work also evaluated the robustness of Song *et al.*'s technique in several aspects. For instance, Kulkarni *et al.*. evaluate the assumptions [34] made by Song *et al.*, showing that under certain conditions the limits established by Song *et al.*. could be surpassed—see Sections 2.1.4 and 2.2.1 for more details on the type of model on which Song *et al.*'s technique works. Some of its mathematical minutia have also been scrutinized [68], the argument being that some details in the formula to estimate

predictability could be improved. We have incorporated these improvements in our discussions and implementations.

As we argue in Chapter 1, Song *et al.*'s technique has three major shortcomings. First, because of the innerworkings of its entropy estimator, the technique has low interpretability. Second, it views human mobility as a single entity, therefore making it hard to study separate components of an individual's mobility. Third, it does not allow for the use of contextual information when estimating the predictability of a person's mobility. In this study, we focus on addressing these three shortcomings. We now discuss previous work related to them.

2.2.2.1 Understanding predictability in human mobility

Earlier work examined how predictability varies according to features of the data, to the prediction task under study, and how certain types of mobility patterns influence predictability. In this section, we discuss these studies and position our contributions in relation to theirs.

Smith *et al.* [56], for instance, studied how predictability varies for different temporal and spatial resolutions, showing that predictability is directly proportional to the temporal sampling rate (the more frequent the rate at which the user's locations are recorded, the higher the predictability) and inversely proportional to the spatial resolution of the dataset (the smaller the cells in the spatial grid, the lower the predictability).

Previous work has also computed the limits of predictability for the two different prediction tasks, namely next-cell and next-place prediction [28, 15], showing that the predictability for the next-place prediction problem is lower than that of the next-cell prediction problem. This shows, as the authors argue in their study, that the next-place prediction problem is harder than the next-cell prediction problem.

Cuttone *et al.* [15] and Ikanovic *et al.* [28] also showed that predictability is affected by stationary patterns, *i.e.*, periods in which the person stays at his or her current location for a long time. In their study, they argue that predictability is directly proportional to the amount of stationarity present in the data. In this thesis, we confirm these previous findings, namely the influence of spatiotemporal resolution on predictability (Section 4.2.3), the effect of different prediction tasks (Section 4.2), and the impact of stationarity on predictability.

We also argue that stationarity alone is not sufficient to explain predictability, and propose alternative metrics that, together with stationarity, help us better understand what affects predictability (Section 4.1). Additionally, we propose regression

models that use our proposed metrics to fit the entropy of the mobility patterns of individuals, showing that our proposed metrics are able to explain the vast majority of the variability in predictability for both next-cell and next-place prediction (Section 4.2).

2.2.2.2 Understanding predictability of components of human mobility

Although previous work [57, 48, 3] modeled individual human mobility as consisting of two types of visits (explorations and preferential returns), previous studies on predictability [58, 39, 15, 56, 42, 60, 62] viewed individual human mobility as a whole, monolithic entity.

In Chapter 5, we propose a strategy to separate a person's mobility into two components: *novelty and routine*, which map explorations and preferential returns, respectively. 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. To our knowledge, we are the first to propose a strategy to investigate the predictability of different components of human mobility. We direct our focus to the study of the predictability of the routine component of human mobility, as discussed in Section 5.1.

Our approach is different from previous work about predictability [15, 38] in two important aspects. First, our goal is different from that of those prior studies, where the authors investigated how the exploration (or novelty) part of a person's mobility trace impacts predictability. In contrast, we here focus primarily on the routine component with the goal of showing that there are patterns in one's routine that are also hard to predict, and therefore affect predictability. In other words, we look at a person's mobility trace from a different perspective, being thus complementary to those prior studies. Rather than quantifying predictability for various sizes of the novelty component (as previous work), we here take this component "as is", and look instead at how much the person's routine deviates from a baseline routine which is completely predictable.

To do that, we propose to create a baseline sequence, as explained in Section 5.1, which has the same size, and the same number of exploration visits as the original sequence. Since the baseline sequence has a completely predictable routine component, by comparing it with the person's actual mobility trace, we can assess how much the person's routine deviates from this completely predictable one. One of the contributions of our work is a closed-formula that allows us to compute the entropy of the baseline sequence, which is in turn used to compute the predictability gap, the difference between the predictability of the baseline sequence and the original sequence.

Second, given that our goal is different from that of previous work, our findings are also different. Previous work stressed the fact that exploration is hard to predict and therefore its amount in a given mobility trace impacts predictability. In contrast, we here show that one's routine also contains behavior that is hard to predict, according to the state-of-the-art predictability technique. This hard-to-predict behavior in one's routine is reflected in the predictability gap, as shown in Section 5.2.1.

Furthermore, we conduct a thorough analysis of routine-related mobility, using the metrics proposed in Chapter 4: *regularity*, *stationarity*, and *diversity*. As we show in Section 5.2.3, these metrics help us to understand what affects the predictability of the routine component of a person's mobility, providing insights into the type of patterns that make one's routine easier or harder to predict.

2.2.2.3 Predictability and contextual information

In this section, we discuss previous attempts to examine the role of external factors, which we here call contextual information, *e.g.*, day of the week, hour of the day, the weather, the location of a person's friends, etc., on predictability, as well as how to measure the impact of such information on predictability estimates. We also discuss how our work complements and expands on previous studies.

Previous work has showed that context is useful for prediction. For instance, Cho *et al.* [13] show that social relationships can explain 10% to 30% of people's movements. Jeong *et al.* [29] use the locations of other users in a network to predict a particular user's locations, showing that this technique significantly outperforms existing predictors, and in particular those that only exploit individual past trajectories.

Cuttone *et al.* [15] also show that context is useful for prediction by building a prediction strategy that considers context which performs better than a baseline which does not consider such information. They therefore claimed that context should be useful for predictability, but they do not show *how* to evaluate the role of context on predictability. In fact, to evaluate the impact of context on predictability estimates one would first have to know how to incorporate context into predictability estimates.

Indeed, to our knowledge, there have been very few attempts to do so. One such attempt was Smith *et al.*'s work [56]. Smith *et al.*'s work showed that the limits of predictability can be refined by excluding from locations that are far away from the user's current position from the set of possible next locations. Thus, they do not directly incorporate this information into Song *et al.*'s technique, but apply prior filtering to the set of next locations so as to eliminate those that are unlikely to be visited next.

Another attempt to use context to help predictability is recent work by Bagrow *et al.* [5], where the authors estimated how much knowing the contents of the tweets of a person's friends helps in predicting the contents of this person's tweets. Although in a different domain, this work is relevant because it shows a way to use context with Song *et al.*'s technique without having to filter the data.

In the case of human mobility, their strategy would be equivalent to estimating the predictability of a person's locations based on the predictability of her friends locations. The drawback of their approach is that it works only for some types of context. For instance, when estimating the predictability of a person's locations based on her friends locations, we are only dealing with locations. Thus, both the target sequence (the person's locations) and the context (her friends locations) are of the same type.

As we will argue in Chapter 5, it is quite challenging to extend Song *et al.*'s method to directly incorporate other types of contextual information (e.g., weather, time of the day). In Sections 6.3.1 and 6.3.2, we propose two strategies to do that and evaluate the impact of context when using these strategies to estimate predictability. Additionally, we evaluate alternative ways to use context with predictability by investigating different entropy estimators and showing how context can be incorporated to them.

2.3 Summary

In this chapter, we provided the necessary background to understand predictability in human mobility as well as provided a brief overview of individual human mobility. We first explained and defined the mobility prediction problem. We then explained the two prediction tasks we target in this study, as well as the several types of prediction models proposed in the literature. We also explained how entropy, compressibility and predictability are related as well as the type of predictor for which Song *et al.*'s predictability technique works.

In the next chapter, we discus the impact of different data sources on predictability as well as explain and characterize the datasets used in this thesis.

Chapter 3

Datasets

Recall from Chapter 2 that the predictability Π_{max} of a sequence X of locations is computed based solely on the data from which X is extracted. As such, Π_{max} is an expression of human behavior, as revealed by the data. Thus, properties of the data are of key concern to understanding Π_{max} values. In this chapter, we discuss the role of data on predictability estimates (Section 3.1) and explain the datasets used in this doctoral dissertation (Section 3.2).

3.1 Predictability and Data

As explained in Section 2.2.1, Song *et al.*'s predictability technique disregards particular prediction strategies and focuses rather on the data to obtain an upper bound on the prediction accuracy that can be achieved for a given dataset. Mobility datasets have thus great impact on predictability estimates, and different types of dataset capture different aspects of a person's mobility. In this section, we review some of the most popular types of mobility datasets as well as describe the impact of different data features on predictability estimates. We also discuss the advantages and disadvantages of each type of dataset for studying predictability and explain the limitations that they impose on predictability estimates.

Indirectly Collected Mobility Data

In this section, we discuss a broad category of mobility data that we call *indirect data*, in the sense that a person's mobility is not measured directly (there is no device attached to person's body recording their position), but rather through an indirect artifact. Examples of this type of data are census data, tax revenue data, travel surveys, and

bank notes, among others [6]. We here describe one of the representatives of this category, namely census data.

Census data was one of the first types of data used to perform mobility prediction [47]. It is collected periodically and usually contains nation-wide data. The data is usually collected by government employees who go door to door with surveys containing questions about the socio-demographic and economic status of household members. Examples of mobility-related information collected through this process are: location of current residence, previous residence, and workplace, as well as means of transportation from the person's home to their workplace.

In the U.S., this data is compiled and made available by the United States Census Bureau[1] in the form of aggregate commuting flows that indicate the counties of origin and destination of people's commute [6]. Two of the biggest drawbacks of census data are (i) the fact that the data collection process is expensive and time-consuming, and (ii) the long periodicity with which the data is collected (typically every 10 years) [67]. Additionally, because the data is made available on an aggregate level, it is not possible to use it to predict human mobility at the user level.

Impact on predictability As a consequence of its granularity (aggregate), census data is not well-suited for studying predictability as we do here, *i.e.*, at the user level. In order to study human mobility at the individual level, it is necessary to collect (or simulate) user level mobility data. Before the popularization of smartphones and other cheap sensors, which allow for direct data collection at the user level, researchers relied on synthetic mobility data, which simulate human mobility using synthetic traces generated according to some pattern [44, 9, 43]. This type of data, although useful for studying predictability, suffers from the obvious fact that it does not truly reflect a real person's mobility. Therefore, it is difficult to draw meaningful conclusions about human mobility based on synthetic data.

Call Detail Records (CDRs)

Call Detail Records (CDRs) are data related to mobile phone calls and text-messages, and they are one of the most used types of mobility datasets [58, 39, 53, 27], and they are collected as follows. Whenever a user makes or receives a call, the phone company relays the call to the nearest phone tower and registers the user's activity in a record that has roughly the following format: `caller-id, caller-cell-id, datetime, duration`. There can also be additional fields, depending on the CDR. Common extra features are

[1] `http://www.census.gov/hhes/commuting/data/commutingflows.html`

the type of the activity being registered (call or text message) and also the identifier of the user at the other end of the interaction. Thus, a CDR dataset consists of several records in the above format, for each user.

One of the features that make CDRs attractive for mobility studies is the fact that they provide a fine-grained view of the mobility of users. They also have some disadvantages, which are mainly due to their temporal and spatial resolution. The number of phone towers in an area is usually proportional to the number of people who live in that area, thus there may be an uneven distribution of phone towers in urban and rural areas, which may hinder one's ability to study mobility patterns in rural areas. Furthermore, even in urban areas, some areas have many more towers per square kilometer than others, and even in the areas with many towers, each tower covers a large area (typically more than $1km^2$), therefore not allowing a fine-grained spatial view of people's mobility patterns.

Aside from that, because each call detail record is only generated when the user receives or makes a call, CDRs are dependent on user activity, and the records are not sampled at a uniform temporal rate. The dependence on user activity can generate biases [50] and as sometimes users can stay for long periods without placing or receiving a call, CDRs usually do not allow for a fine-grained view of people's mobility in time.

Impact on predictability These characteristics can have impacts on predictability estimates. For instance, as CDRs depend on user activity to log their mobility, what is revealed in the data may not offer a realistic picture of the user's mobility. In other words, the data may not constitute a good sample of the user's mobility, which will result in distorted predictability values. The time between each location record can also be an issue for predictability estimates, as the user may have moved to many locations between two calls, but those locations would not be taken into account for predictability purposes.

Social Media Data

Social media is another popular source of mobility data [31, 60, 25, 52]. When a user posts something on a social network such as Twitter, Facebook, Instagram, or Foursquare, there is usually an option to associate a location to the content of the post. These posts typically consist of a picture or text, and a location where the activity was recorded. This information usually assumes the following format: `user-id, lat, lon, timestamp, content`.

The ability to have access to people's locations when they use social media also makes this type of data attractive to mobility studies, and the large number of users in these services contributes to the usefulness of social media data for understanding mobility patterns. It has been shown that social media data can indeed be used to infer certain mobility patterns [25]. Unlike CDRs, social media data provides a more accurate location of the users' activities, as it is registered by the GPS system in the user's mobile device.

There are, however, a couple of issues with using social media data for studying human mobility and, more specifically, predictability. The first is that, as in the case of CDRs, social media data is activity-dependent because a record is generated only when the user posts something on social media. Furthermore, it has been argued that social media data is biased for studying mobility because the sampling pattern is skewed (people do not post content from every location they visit), and because there might be a bias in what type of people use the social network [31].

In an attempt to circumvent some of the problems with social media datasets, some people have tried using data from multiple sources so as to reduce the irregular sampling rate of this type of data [60, 31, 25]. Nowadays, people use their credentials in one social media website to register for other websites, and when they do so, their activity is registered in multiple places.

For instance, people often use their Twitter account to register on Foursquare or Instagram, and when doing so, they allow their posts on these services to be also registered on Twitter. Thus, by collecting a user's feed, it is possible to gather their information in other social networks as well. We have already shown [60], however, that in some cases data from different social networks does not reinforce mobility patterns, but rather captures different patterns that were not present in the original data. As a consequence, we have shown that using different data sources does not necessarily lead to higher predictability.

Although this strategy of capturing a user's mobility in different data sources is sound, it is difficult to apply it to datasets of different types. For instance, one may wish to capture a user's mobility patterns on CDRs and on social media at the same time, but the major problem is how to identify the same user in both datasets. As explained, it is relatively easy to identify a user in different social networks, but it is much more challenging to do so in datasets of different nature.

Impact on predictability The limitations of the use of social media data for studying predictability are similar to those of CDR datasets. First, as the record of one's mobility depends on one performing an action (posting on social media), the locations that are

recorded are only those posted on social media. Furthermore, the biases associated to the type of location where people usually post on social media are also reflected on their predictability estimates. Second, as people only tend to post on social media from a few selected locations, the data may not constitute a good sample of their mobility, could lead to distorted predictability estimates. Additionally, as people tend to post less frequently from places they visit often, this type of dataset is not well-suited for capturing people's routine.

Global Positioning System Data (GPS)

A GPS dataset typically consists of location samples from a set of users, and this data type solves some of the problems with CDRs and Social Media data. The location of each user is obtained from the device's GPS system at a uniform temporal rate, usually every couple of minutes. As the user's position is precise (up to the accuracy of the device's GPS system) and the temporal sampling rate is uniform, GPS datasets usually provide a fine-grained spatiotemporal view of users' mobility patterns.

The problem with GPS data is that it is usually hard to get such data for a large volume of users. For instance, recent studies [56, 1, 15, 28] that used GPS datasets consist of a few thousand users, at most. Given the other advantages of GPS data (regular sampling rate, and accurate locations) this type of data is very attractive for studying predictability.

Impact on predictability For the purposes of studying predictability, GPS datasets have many desired characteristics, as we highlight in Table 3.1. The only drawback of this type of dataset is that it is usually hard to get such data for many users over a long period of time.

Ideal Dataset for Studying Predictability

The ideal type of dataset i.e., the one that would allow for a more comprehensive and nuanced study of human mobility would have: high temporal and spatial resolution of a large number of users over a long period of time. In practice, however, this type of dataset is hard to find. Some types of dataset have some of those attributes, but lack others. In other words, mobility datasets provide only a window to a person's mobility. Table 3.1 lists some attributes of the types of mobility datasets we have described and also lists the characteristics of the ideal dataset for mobility studies.

In this dissertation, we use a GPS and a CDR dataset, as we will discuss in Section 3.2. We chose these two datasets for two main reasons. The first one is that,

Dataset Type	User-Level Mobility	Regular Sampling	High Temporal Resolution	High Spatial Resolution	Long Period Covered	Many Users
Indirect Data	✗	✓	✗	✗	✓	✓
Social Media	✓	✗	✗	✓	✓	✓
CDR	✓	✗	✓	✓	✓	✓
GPS	✓	✓	✓	✓	✓	✗
Ideal Dataset	✓	✓	✓	✓	✓	✓

Table 3.1: A summary of the most popular types of mobility datasets and a comparison with the Ideal Dataset of mobility datasets. We consider a dataset high temporal resolution if it usually contains many observations per day, and we consider a dataset as having many users if it usually has thousands of users.

by looking at Table 3.1, they are the ones which have most of the desired attributes for studying predictability. Second, they do not have the same drawbacks, which reduces the possibility of some type of behavior not being shown in the data. Our CDR dataset has an additional shortcoming compared to what is shown in Table 3.1: its period of observation is rather short (two weeks), as we will describe in Section 3.2. However, it tends to be less biased than social media data, for the reasons described in Section 3.1, therefore we chose to use it instead of using social media data. On the flip side, our CDR dataset does not suffer from a common drawback in this type of data, as it has a regular sampling rate, which is another desired attribute for studying predictability.

3.2 Our Datasets

Our study is composed and driven by a series of analyses performed on two different mobility datasets, of distinct temporal and spatial resolutions, which allow us to study the impact of spatiotemporal factors on Song *et al.*'s technique. These datasets, which are summarized in Table 3.2 and discussed below, are representatives of two categories of datasets often used in mobility studies (GPS and CDR datasets), as mentioned in Section 3.1.

As discussed in Section 3.1, GPS data has many desired properties for studying predictability: regular sampling rate, precise location records, and in the case of our dataset, the period of observation is also long. Our CDR dataset offers us another view of mobility data: it has many users, and unlike other CDR datasets, it offers a regular sampling rate (one observation every two hours, on average). It also has a period of observation shorter than our GPS dataset, which allows us to study short-term predictability. We believe that these characteristics make these two datasets of great importance and relevance for studying predictability. We now proceed to discuss

them in more details.

	GPS dataset	CDR dataset
Number of users	67	2,780
Period covered	18 months	2 weeks
Temporal resolution	5 minutes	1 hour
Spatial resolution	200 meters	200 meters

Table 3.2: Summary of our GPS and CDR datasets.

3.2.1 GPS Dataset

The first dataset is a high temporal and spatial resolution dataset consisting of GPS traces. This dataset was obtained through an Android mobile phone application, called MACACOApp[2]. Users who volunteered to install the app allowed it to collect data such as uplink/downlink traffic, available network connectivity, and visited GPS locations from their mobile devices. These activities are logged with a fixed periodicity of 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.

3.2.2 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

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

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.

3.2.3 Data preprocessing

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

Additionally, our preprocessing methodology for the GPS dataset is similar to that of Song *et al.*'s work, where the authors overlay a grid of square cells onto the geographical region, and consider every cell as a distinct location. Observations of a user's position that happen inside the same cell are considered to be the same location. This strategy is different from other strategies [24, 15] which identify *movements* and *stop locations*, and then consider as actual locations only those labeled as *stops*.

This preprocessing strategy also has implications on the next-place prediction task, which was originally defined [15] taking into account *movements* and *stops*. In our case, we consider every distinct location that appears in a user's mobility trace as an actual visit, and not only *stop locations*. In practice, this makes mobility traces larger in the next-place prediction task, which is important for predictability purposes, as entropy estimators tend to yield more reliable estimates for longer sequences.

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.

3.2.4 Data Characterization

In this section, we discuss properties of our datasets that are relevant to our study of predictability by showing relevant data for both next-cell and next-place prediction, which, as mentioned in Section 2.1.3, are the prediction tasks that we focus on throughout this dissertation.

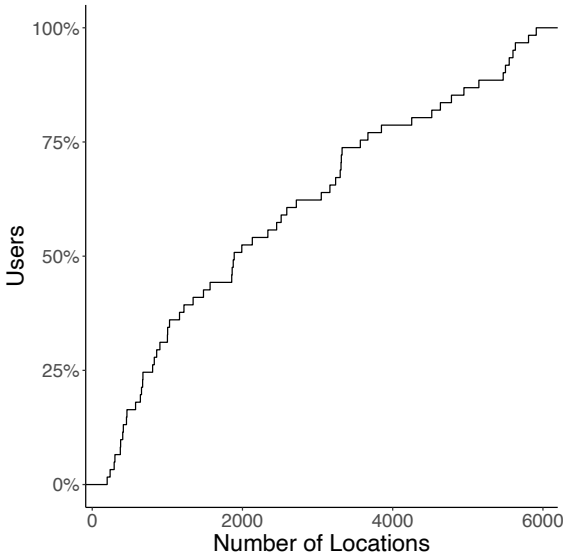
As mentioned in Section 2.1.3, whenever we talk about a dataset for next-place prediction, we are referring to the dataset after we filter out stationary periods. Thus, we expect that the number of locations to be lower in the next-place datasets, as evidenced in Figures 3.3 and 3.4.

In Figure 3.1 we show the distributions of total number of locations and total number of unique locations visited by a user. Recall that entropy and predictability are strongly dependent on these two metrics, as defined in (Equations 2.1 and 2.3). The total number of locations is important because the entropy estimator used by Song *et al.* converges to the actual entropy of the sequence as the length of the sequence goes to infinity. As shown in Figure 3.1(a), at least 50% of the users in our GPS dataset visited more than 2,000 locations. The variability in the number of locations is due to the fact that different users in our GPS dataset were active in the data for different periods of time.

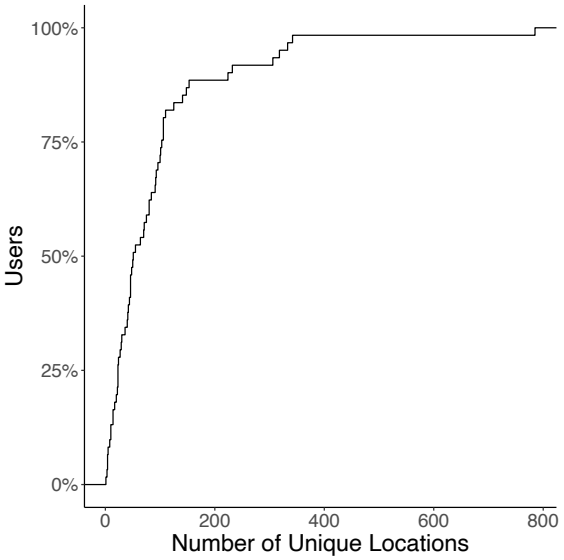
As we will discuss in Section 4.1.2, the number of unique locations is also important for predictability. In general, we expect that the more unique locations a user visits, the higher the entropy of her mobility and the lower her predictability. As shown in Figure 3.1(b), most of the users in our GPS dataset have less than 400 unique locations in their mobility trace.

In Figure 3.2, we show distributions of the total number of locations and number of unique locations visited by a user for our CDR dataset, in the next-cell prediction task. In this dataset, we have low temporal resolution (fewer observations per time unit) and a shorter period of observation (two weeks only), according to Table 3.2. As a consequence, the total number of visited locations and the number of unique locations are smaller when compared to the GPS dataset. These differences in the total number of locations and unique locations in our two datasets allow us to conduct our analyses in different scenarios. Specifically, we can evaluate how Song *et al.*'s technique as well as our proposed techniques work for different types of datasets with distinct spatiotemporal resolutions.

We now turn our attention to distributions of total number of locations and total number of unique locations visited by a user in the next-place prediction task. Figure 3.3 shows the cumulative distributions for the total number of visited locations

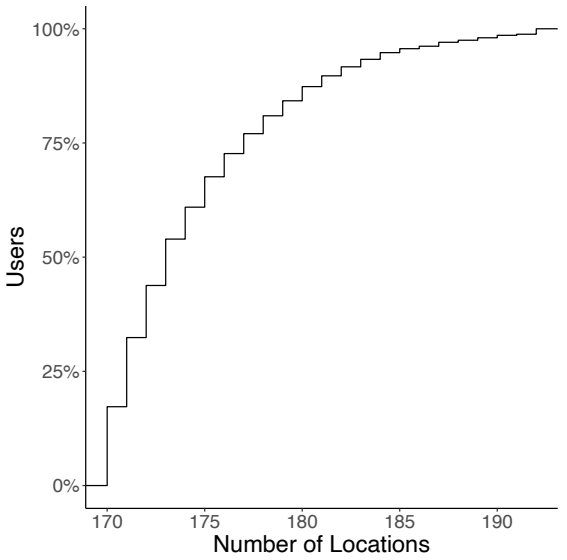


(a) CDF of visited locations (GPS)

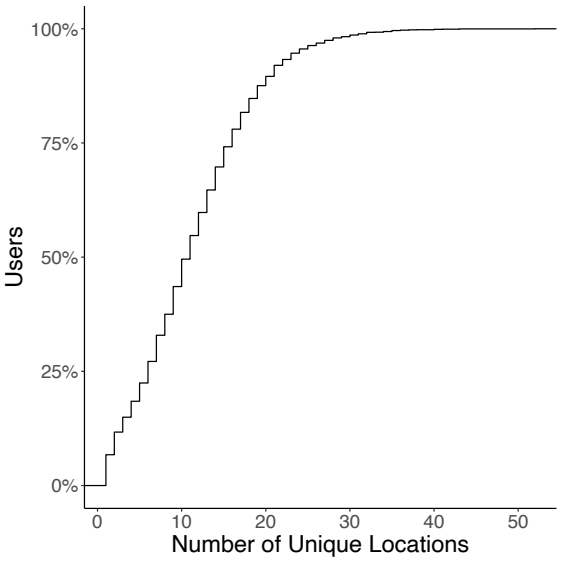


(b) CDF of unique locations (GPS)

Figure 3.1: Cumulative distribution of the total numbers of visited locations and unique locations in the GPS dataset for the next-cell prediction task.



(a) CDF of visited locations (CDR)

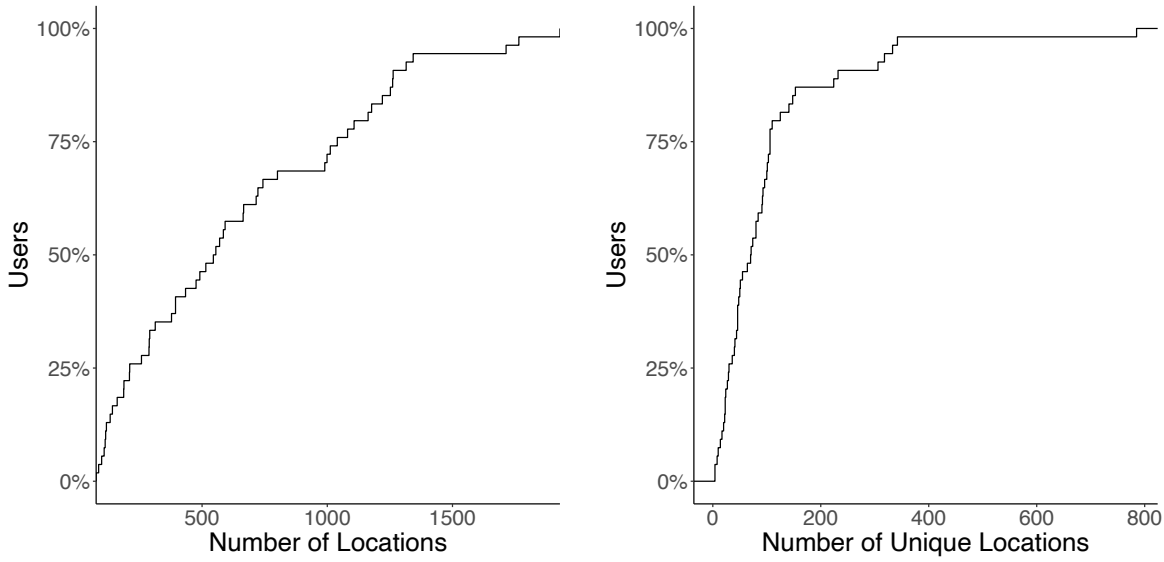


(b) CDF of unique locations (CDR)

Figure 3.2: Cumulative distribution of the total number of visited locations and unique locations in the CDR dataset for the next-cell prediction task.

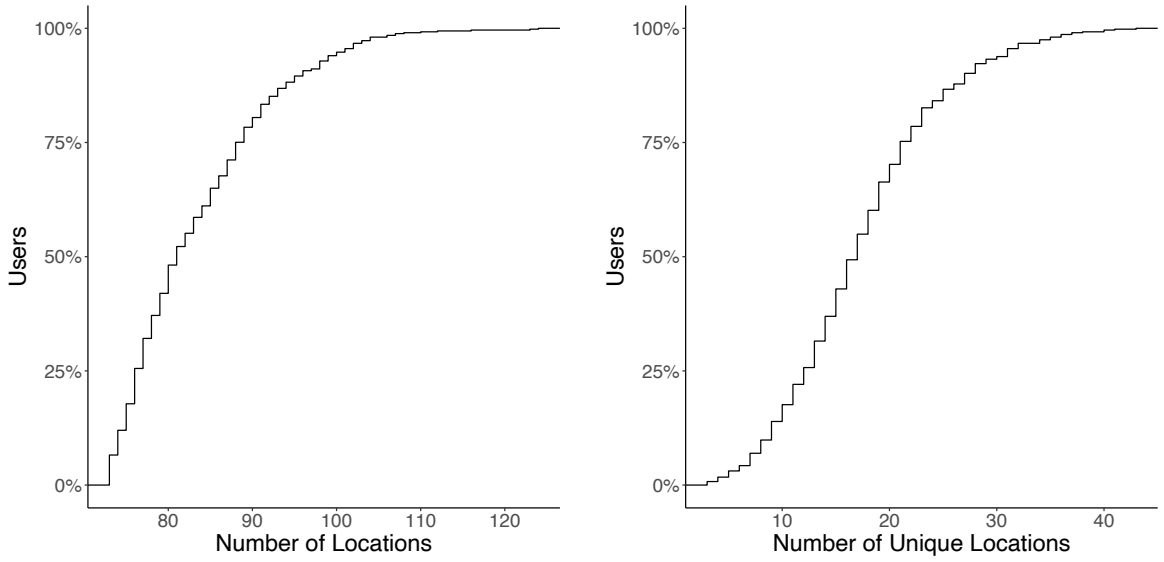
and unique locations for the GPS dataset. We note that the total number of visited locations is smaller in this prediction task, compared to the values in Figure 3.1(a). Thus, the removal of stationarity from our dataset results in fewer total locations, but the number of unique locations seems to remain unchanged (Figures 3.1(b) and 3.3(b)). The same phenomenon can be observed in Figure 3.4, which shows the distributions

of total number of locations and number of unique locations for the CDR dataset, in the next-cell prediction task.



(a) CDF of visited locations (GPS) (b) CDF of unique locations (GPS)

Figure 3.3: Cumulative distribution of the total number of visited locations and unique locations in the GPS dataset for the next-place prediction task.



(a) CDF of visited locations (CDR) (b) CDF of unique locations (CDR

Figure 3.4: Cumulative distribution of the total number of visited locations and unique locations in the CDR dataset for the next-place prediction task.

The number of locations in a person's mobility trace raises important issues, as we will discuss in more details in Chapter 4 and Chapter 6. Specifically, the predictability technique is greatly impacted by the size of the input sequence, as shown

in Equation 2.1. This is an issue that permeates several of our discussions about predictability, as it impacts predictability estimates. For instance, it is hard to obtain a robust predictability estimate for small input sequences. Throughout this thesis, we made significant efforts to identify these cases, and we discuss their impact on our results. This is an issue that appears more in our CDR dataset, as our GPS dataset has significantly longer input sequences.

However, as we will discuss later, studying predictability of small input sequences also reveals important properties of the state-of-the-art predictability technique, and sheds some light on the types of mobility patterns that can be uncovered with small and large input sequences. For instance, previous work [48] has shown that the number of unique locations visited is higher in small sequences. Conversely, one's routine becomes more apparent as the number of places visited grows.

3.3 Summary

In this chapter, we described several types of mobility data sources, discussing their properties and how these properties relate to predictability estimates. We also characterized both of our datasets, showing that they cover distinct time periods (18 months versus 2 weeks), have different temporal resolutions, and different number of users. These differences, allied with the fact that we are targeting two distinct prediction tasks, poses quite a challenge to our analyses and techniques. On the flip side, these same differences allow us to draw more general conclusions, *i.e.*, ones that apply to different data types, temporal resolutions, and for users with short or large mobility traces.

In the next chapter, we describe our investigation of the first research goal of this dissertation: understanding and interpreting predictability estimates in human mobility.

Chapter 4

Understanding Predictability in Human Mobility

In this chapter, our goal is to understand what affects the predictability of a person's mobility. To that end, we propose to use three simple metrics (regularity, stationarity, and diversity) that capture important aspects of human mobility (Section 4.1). We then use the proposed metrics as proxies to predictability. Specifically, we build regression models that use these metrics to fit the entropy of a person's mobility in the next-cell and next-place prediction tasks (Sections 4.2.1 and 4.2.2, respectively).

Recall that Song *et al.*'s technique [58] estimates the predictability of a sequence of locations based on the entropy of the sequence. Specifically, their work established limits on the predictability of a sequence of locations based on three estimates of the entropy of the sequence. The first estimate is the Shannon entropy [51] of a uniform distribution on possible locations, shown in Equation 4.1. This estimate is known to yield the highest possible entropy value for an input sequence, thus establishing a lower bound on predictability.

$$S_{uniform}(X) = \log_2(n), \tag{4.1}$$

where n is the total number of locations in the input sequence.

The entropy bound of Equation 4.1, however, can be refined if a non-uniform probability distributions on locations, which take into account the relative frequency with which a user visits each region, is considered. In fact, prior work has shown that people often visit a few places and occasionally go to previously unvisited locations [21, 30]. Thus, we consider the distribution p_{freq} on X in which, for each $x_i \in X$, $p_{freq}(x_i)$

is the frequency with which location x_i was visited in the observed time frame. The corresponding entropy S_{freq} is obtained as follows:

$$S_{freq} = -\sum_{x_i \in X} p_{freq}(x_i) \log_2 p_{freq}(x_i). \tag{4.2}$$

Despite being more general than the first approach, S_{freq} is still not a completely adequate estimation of the "real" entropy of a sequence of visited locations. This occurs because S_{freq} does not capture the full temporal patterns of people's location history. For instance, if a person visits the sequence of locations A, B, and C several times, S_{freq} would only consider the number of times each location was visited, but not the fact the the sequence ABC appears multiple times—a fact that is useful if the person is at locations A or B and we are trying to predict her next location.

Yet, it is possible to derive a probability distribution on locations that captures such temporal correlations, which is done in the third, more precise variation of entropy used in Song *et al.*'s work, which estimates the entropy using a distribution that accounts for both the frequency of visitations as well as temporal patterns. The third estimator, described by Kontoyiannis et al. [33], is related to the Lempel-Ziv compression algorithm [35] and to the Lempel-Ziv measure of the complexity of a sequence [35]. According to its definition, the entropy S_{real} of an input sequence of locations X of size n can be approximated by:

$$S_{real} \approx \frac{n \log_2(n)}{\sum_{i \leq n} \Lambda_i}, \tag{4.3}$$

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.

For ergodic, stationary processes, this estimator is said to converge to the entropy rate of the source as the size of the input goes to infinity [14]. This estimator does not require the underlying probability distribution of the symbols of the source. As such, it is suitable for computing the entropy of mobility traces, for which we may never know the true underlying probability distribution. Note that different values of entropy yield different limits of predictability: while the first two variations of the entropy work by directly manipulating the underlying probability distribution of the locations, the third, more precise one, leverages the relation between entropy and compressibility to estimate the entropy of the input sequence, which poses challenges to interpretability of predictability values.

As argued in Section 2.2.1, predictability is directly related to a good estimate of

the entropy, therefore we need to understand what affects the entropy of an individual's mobility. However, it is hard to do so only by looking at the estimator shown in Equation 4.3. In this chapter, we propose indirect ways to understand what affects predictability, and we show that these indirect ways, while being simpler and easier to interpret, explain most of the variability in the entropy of one's mobility.

4.1 Proxy Metrics

In this section, we present three metrics that capture key aspects of mobility patterns and show that they can effectively be used as proxies to understand predictability estimates. We start by arguing that analyzing the entropy estimate itself, particularly the more precise one based on compressibility, which is focus of our study, is quite challenging, as the result of the method is hard to interpret.

Thus, we look for simpler and easier to understand proxy metrics, which can be used in its place to understand predictability in human mobility. Specifically, we employ three *simple* metrics that help explain what affects predictability in a sequence of locations visited by a user, as captured by Song *et al.*'s estimate.

4.1.1 Stationarity

The first metric, called the *stationarity* of a sequence of locations, is related to the number of observations for which the person stays continuously in the same location. Given a time-ordered sequence $X = (x_1, x_2, \ldots, x_i, x_{i+1}, \ldots, x_n)$ of observations of a person's location, we say that a *stationary transition* occurs at time i if $x_i = x_{i+1}$. Thus, the stationarity of sequence X is the ratio of stationary transitions over the total number of transitions in X. More formally, we can define the stationarity of a sequence as follows:

Definition 4.1.1. ***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.

For example, sequence $X = (1, 1, 2, 2, 3, 3, 4, 4)$ contains a total of seven transitions, four of which are stationary. Therefore, the stationarity of the sequence $st(X)$ is $st(X) = 4/7 = 0.57$. Intuitively, if a person stays at the same location for a long period

of time, there will be many consecutive repeated symbols in the sequence. Sequences with many consecutive repeated symbols are easier to compress, therefore the higher the stationarity of a sequence, the lower its entropy.

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

4.1.2 Regularity

In order to capture the aforementioned phenomenon, we introduce another metric, called *regularity*, that also helps explain the entropy of a person's observed location history. The *regularity* of a sequence captures the preferences of a person to return to previously visited locations. It is defined as one minus the ratio between the number of *unique* symbols and the length of the sequence.

Definition 4.1.2. ***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.

For instance, the regularity of input sequence $X = (1, 2, 2, 3, 3, 3, 4, 4, 4, 4)$ is given by $reg(L) = 1 - 4/10 = 0.6$. If we compute the *regularity* of the two aforementioned example sequences (X_1 and X_2), we obtain $reg(X_1) = 1 - 4/8 = 0.5$ and $reg(X_2) = 1 - 2/8 = 0.75$, which helps explain why X_2 has lower entropy than X_1 (X_2 is more regular than X_1). Intuitively, the more regular a sequence, the fewer distinct symbols it has, and sequences with few distinct symbols are easy to compress. Therefore, the higher the regularity, the lower the entropy of a sequence.

4.1.3 Diversity

Although useful, regularity and stationarity do not fully explain the predictability of a person's mobility. Consider, for instance, the following two sequences $X_1 = (1, 2, 3, 1, 2, 3, 1, 2, 3, 1)$ and $X_2 = (1, 3, 2, 1, 2, 3, 1, 3, 2, 1)$, which represent two 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 123 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 2.1, is 1.50 whereas the entropy of X_2 is 2.18.

To capture additional patterns affecting the entropy (and thus predictability) associated with a given 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 123 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 4.1.3. ***Diversity:*** 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 2.1 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 $div(X_2) = 0.76$.

To compute *diversity*, we count the number of distinct substrings of size $1 \leq i \leq n$, where n is the size of the input string, and divide that number by the total number of substrings in the input string. For a string of size n, there are a total of $\sum_{i=1}^{n} = n(n+1)/2$ substrings. Given that there is a closed-formula for computing the total number of substrings in a given string, the challenge is computing the number of distinct substrings in it. The naive solution is to generate all substrings and count the number of distinct ones. Unfortunately, this solution is slow for large input strings, as its asymptotic complexity is $O(n^2)$. More efficient solutions rely on the *longest common prefix* (LCP) array or the *suffix array* of the input string [22].

We note that our choice of metrics (regularity, stationarity and diversity) comes from experimental observations of how Song *et al.*'s technique works. Intuitively, such metrics capture three key and complementary components of a person's mobility patterns: *the ratio between previously visited places and new places*, *the amount of time spent in each place (stationary transitions)*, and *the number of distinct sequences of locations visited (diversity)*. Although the importance of stationarity to predictability has been noted before [15], using regularity and diversity to help understand predictability

and thoroughly evaluating the three metrics in two different prediction tasks is a novel contribution of our work. These results are discussed next.

4.2 Discussion of results

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

4.2.1 Next-cell prediction

In this section, we evaluate the extent to which regularity helps understand and interpret predictability results in the next-cell prediction task. Recall from Section 2.1.3 that in next-cell prediction, given a sequence $X = (x_1, x_2, \ldots, x_{n-1})$, we are interested in estimating the maximum achievable accuracy when trying to predict x_n, the next symbol in sequence X.

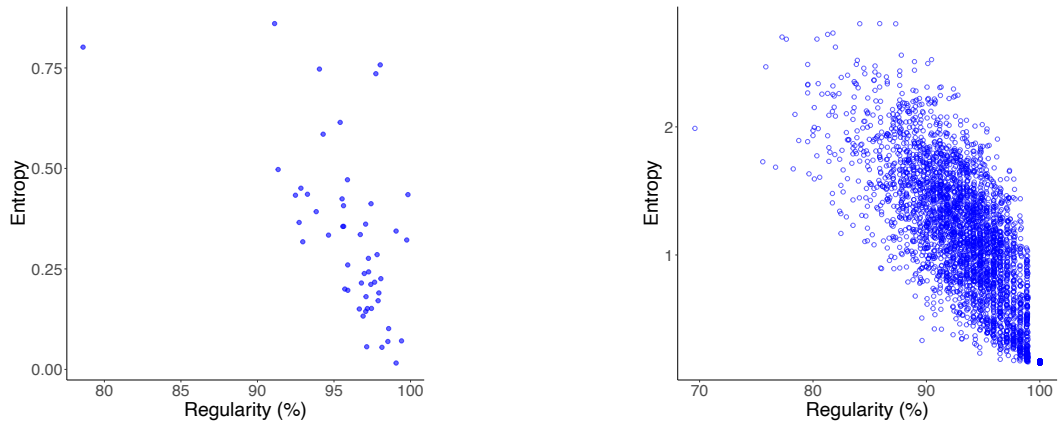
We begin our discussion by first illustrating the relationship between entropy and each of our proposed metrics. Figure 4.1-(a) shows scatter plots with the relationship between regularity and entropy for the GPS and CDR datasets, respectively. Similar plots for stationarity and entropy are shown in Figure 4.1-(b), and for diversity and entropy in Figure 4.1-(c).

In Figure 4.1, we observe that the three metrics have different relationships with entropy. While the relationship between regularity and entropy is more varied, both stationarity and diversity have a clearer relationship with entropy: stationarity seems to vary linearly with entropy, and diversity exhibits a non-linear relationship.

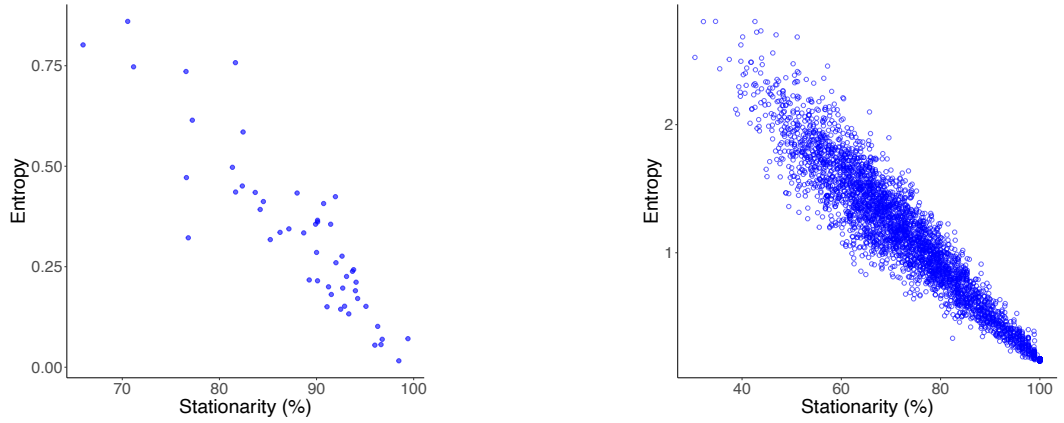
The next step in our analyses is to investigate the relationship between the metrics among themselves, as well as the correlation between the metrics and entropy. To do that, we show, in Table 4.1, the Spearman correlation coefficient between each of our three metrics. In practice a correlation greater than 0.5 in absolute value indicates a strong correlation between two variables.

	GPS				CDR			
	Regularity	**Stationarity**	**Diversity**	**Entropy**	**Regularity**	**Stationarity**	**Diversity**	**Entropy**
Regularity	1	0.50	-0.25	-0.55	1	0.63	-0.72	-0.74
Stationarity	0.50	1	-0.75	-0.88	0.63	1	-0.95	-0.94
Diversity	-0.25	-0.75	1	0.69	x -0.72	-0.95	1	0.99
Entropy	-0.55	-0.88	0.69	1	-0.74	-0.94	0.99	1

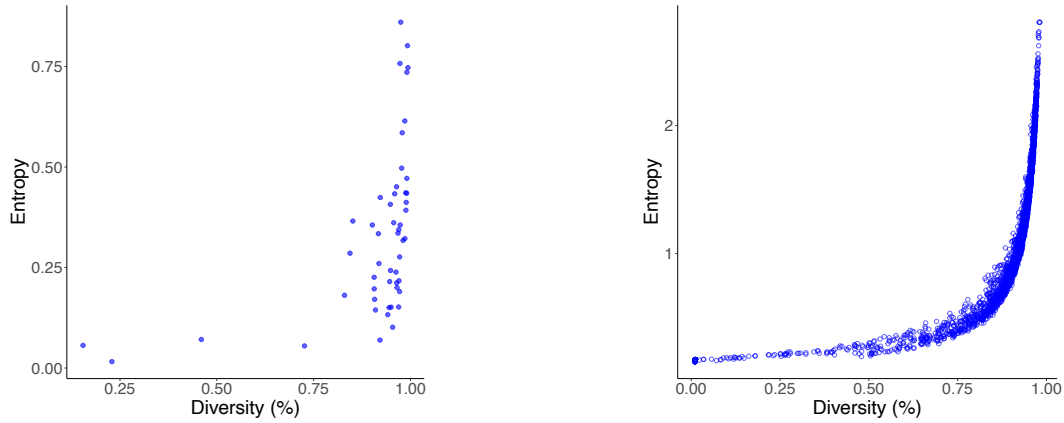
Table 4.1: Pairwise Spearman's correlation coefficient between each proxy metric as well as between each metric and the entropy, computed for each user's mobility trace.



(a) Regularity vs. Entropy (GPS and CDR)



(b) Stationarity vs. Entropy (GPS and CDR)



(c) Diversity vs. Entropy (GPS and CDR)

Figure 4.1: Relationship between our three metrics and entropy. Notice that the plots are at different scales. Notice also that users have similar values for one of the metrics, but very different entropy values, indicating that one of the metrics alone is not able to fully explain entropy.

From Table 4.1, we observe that the metrics are themselves reasonably well correlated, but as illustrated in Figure 4.1, none of them is fully able to explain the entropy.

Indeed, a visual inspection of Figure 4.1 reveals that there are several users who, despite having similar values of one of the metrics, have very different entropy values. This suggests that each metric, in isolation, cannot explain entropy. Investigating such cases, we found that large differences in entropy for users with similar regularity could often be explained by great differences in stationarity or diversity, and vice-versa.

In order to verify the hypothesis that each metric alone cannot reasonably explain predictability, we analyzed the extent to which each metric alone versus the three in conjunction can explain the predictability of a sequence of locations. To that end, we employed a regression analysis by fitting the entropy $H(X)$ of a sequence X as a function of: (i) regularity $reg(X)$ alone, (ii) stationarity $st(X)$ alone, (iii) diversity $div(X)$ alone, and (iv) as a function of the three metrics in conjunction, for all users in each dataset. For the latter, we experimented with different regression functions and the one that led to the best fitted model is given by:

$$H(X) \approx \alpha \cdot reg(X) + \beta \cdot st(X) + \gamma \cdot div(X) + \delta \cdot reg(X) \cdot st(X) \cdot div(X) + \epsilon, \quad (4.4)$$

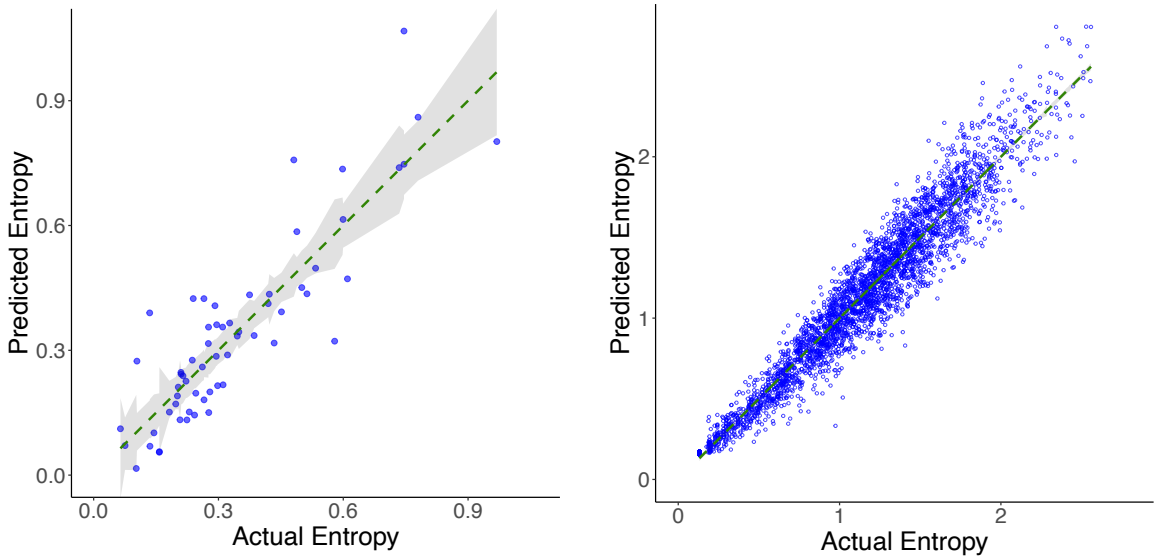
where α, β, γ, and *delta* are the coefficients of regression and ϵ is the regression error. Furthermore, it was necessary to consider the interaction between the three metrics because there is a confounding effect between them—the correlation among them is non-negligible. This function was chosen to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

This model, albeit simple, is able to explain a large fraction of the total variation in the entropy values *in both datasets*. It also shows better entropy fittings when compared to three other models that employ only regularity or stationarity or diversity, as shown by the adjusted R^2 of the models listed in Table 4.2. Additionally, as we further discuss in Section 4.2.3, we experimented with different spatial resolutions. We found that the model in Equation 4.4 also performed well for other spatial resolutions that we tested.

Figure 4.2 shows scatter plots of the actual entropy (x-axis) versus entropy estimated by Equation 4.4 (y-axis) for all users in both datasets. Notice that most dots (users) lie close to the diagonal, especially in the larger CDR dataset. Therefore, our three metrics can indeed be used as proxies for the purpose of studying predictability in human mobility.

	$reg(X)$	$st(X)$	$div(X)$	$reg(X)$, $st(X)$ and $div(X)$
GPS dataset	0.322	0.763	0.180	0.770
CDR dataset	0.566	0.903	0.492	0.935

Table 4.2: Adjusted R^2 of four different regression models, each of which using a combination of our metrics, for both the GPS and CDR datasets in the next-cell prediction task.



(a) GPS dataset (b) CDR dataset

Figure 4.2: Entropy (in bits per symbol) predicted by the regression model (y-axis) versus actual entropy (x-axis), in bits per symbol, for the next-cell prediction task. The green, dashed line shows the regression model (Equation 4.4) and the gray area shows the confidence interval. As there are many more data points in the CDR dataset, the confidence interval area is narrower and almost invisible in the plot.

It is important to note that, as shown in Table 4.2, the model that uses only stationarity performed almost as well as the one that uses the three metrics, in both datasets. This fact illustrates the importance of stationarity for the next-cell prediction task, and suggests that most of the predictability (i.e., achievable prediction accuracy) stems from stationary behavior in the next-cell prediction task, in the two datasets that we evaluated here. In the next section, we will discuss what happens when stationarity is taken out of the equation, i.e., in the next-place prediction task.

4.2.2 Next-place prediction

In this section, we evaluate the extent to which regularity helps understand and interpret predictability results in the next-place prediction task. Recall from Chapter 3 that in the next-place prediction task, there is no stationarity. In other words, given a sequence $X = (x_1, x_2, \ldots, x_{n-1})$, we are interested in estimating the maximum achievable accuracy when trying to predict x_n, where x_n is different from x_{n-1}.

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

To answer the first of these questions, we examine the Spearman correlation coefficient between regularity, diversity entropy in the next-place prediction task. We found that the correlation is between regularity and entropy is -0.80 and -0.79, for the GPS and CDR dataset, respectively. And the correlation between diversity and entropy is 0.88 and 0.98 for the GPS and CDR dataset, respectively. Contrast these values with the corresponding values for regularity in the next-cell prediction task: -0.55 and -0.74 for the GPS and CDR dataset, respectively, and for diversity: 0.69, and 0.99, for the GPS and CDR dataset, respectively. Thus, regularity indeed plays a larger role in next-place prediction than it does in next-cell prediction. And diversity plays a larger role in the GPS dataset and a similar role in the CDR dataset.

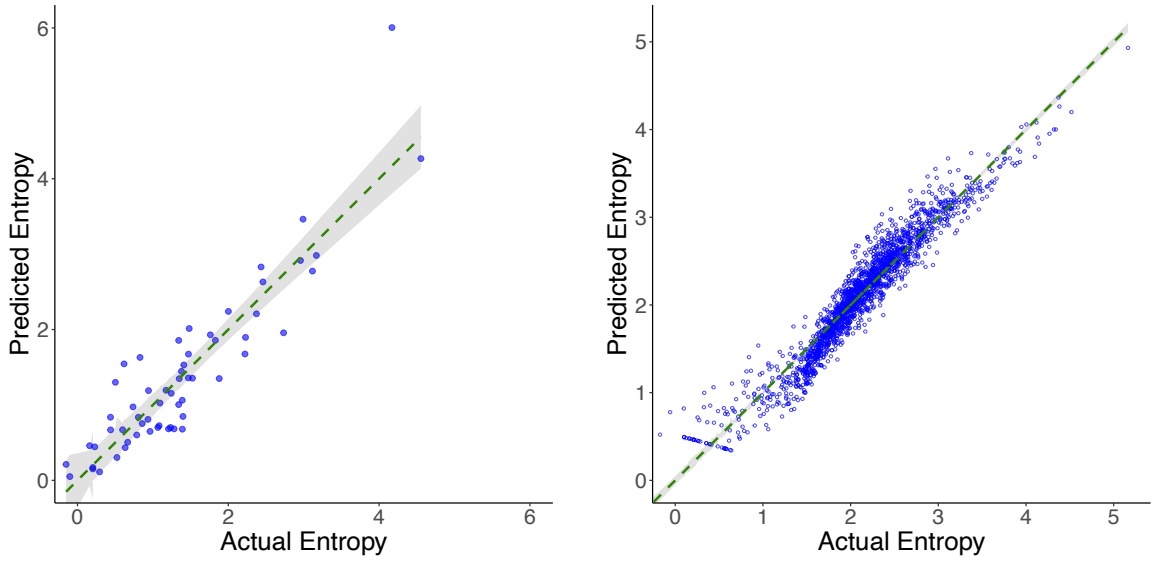
To answer the second question, we build a regression model that uses regularity and diversity to fit the entropy of next-place prediction. Our regression model is as follows:

$$H(X) \approx \alpha \cdot reg(X) + \beta \cdot div(X) + \gamma \cdot reg(X) \cdot div(X) + \epsilon, \tag{4.5}$$

where α and *beta* are the coefficients of regression and ϵ is the regression error.

We evaluate this model in both of our datasets and discover that the adjusted R^2 is 0.855 and 0.913 for the GPS and CDR datasets, respectively. Figure 4.3 shows the entropy fittings for the resulting model.

From the R^2 of the model as well as from Figure 4.3, we observe that regularity and diversity can also explain most of the variability in the entropy of the next-prediction task. Indeed, the importance of these metrics increase in the next-place prediction task (as evidenced by their stronger correlation with the entropy), and they are able to capture most mobility patterns in the next-place prediction task.



(a) GPS dataset (b) CDR dataset

Figure 4.3: Entropy (in bits per symbol) predicted by the regression model (y-axis) versus actual entropy (x-axis), in bits per symbol, for the next-place prediction task. The green, dashed line shows the regression model (Equation 4.5) and the gray area shows the confidence interval. As there are many more data points in the CDR dataset, the confidence interval area is narrower and almost invisible in the plot.

4.2.3 Metrics and Dataset Characteristics

Previous studies [56, 15, 11] have shown that the estimate of predictability in mobility is influenced by the temporal and spatial resolutions of the data. Specifically, greater predictability is expected as temporal resolution increases (more observations per time period) or spatial resolution decreases (larger cells). We now revisit and expand this discussion by looking into how both factors affect regularity, stationarity, and diversity, as well as by examining how data sparsity can affect the metrics.

4.2.3.1 Temporal resolution

Our experiments show that a decrease in temporal resolution makes the average stationarity decrease as well. This occurs because longer time intervals between measurements make it more likely for those measurements to occur at different locations—there is a higher chance that the user moved in a longer time interval. The decrease in stationarity leads to an increase in entropy and thus, lower predictability (on average). In our experiments, we observed an decrease in stationarity of about 6% when the temporal varied from one observation every five minutes to one observation every hour, in the GPS dataset.

The same logic can be applied to the relation between diversity and entropy. Longer time intervals between measurements tend to increase diversity, as there is a higher chance that the user moved in a longer time interval. Thus, we observe that, intuitively, stationarity and diversity vary in opposite ways with respect the temporal resolution.

A less obvious observation is that a decrease in temporal resolution reduces average regularity. This is due to the fact that lower temporal resolution means fewer observations being made overall, which leads to shorter sequences. Recall that regularity is related to the ratio between the number of unique symbols and the size of the sequence. Therefore, a reduction in the size of the sequence will cause a decrease in regularity. In our experiments, by varying the temporal resolution from one observation every five minutes to one observation every hour, the regularity of the GPS dataset decreased by around 17% and 5% for the next-cell and next-place prediction tasks, respectively. In general, less regular sequences will have larger entropy, as Figure 4.1 shows.

4.2.3.2 Spatial resolution

We now turn our attention to the relation among spatial resolution, regularity, stationarity, and diversity. As with the temporal resolution, it is only possible to perform this analysis on the GPS dataset: as it has high spatial resolution, we can tesselate grids of arbitrary size on the target geographical area.

A decrease in spatial resolution means that the cells in the spatial grid are larger, which means that more measurements are going to be made inside the same cell, thus increasing average stationarity and decreasing diversity, as shown in Table 4.3. A decrease in spatial resolution also causes an increase in regularity, as it will be less likely that a person moves outside a larger cell.

Given that more observations of the person's location are going to be made inside the same cell, there will be more repetition of symbols in the person's mobility trace. More repetitions means more compressibility, which results in lower entropy, and therefore higher predictability.

4.2.3.3 Data sparsity

Another important aspect of the interplay among our three metrics is data sparsity. In both of our datasets, the current location of users is measured in a fixed time interval (five minutes for the GPS dataset, and one hour for the CDR dataset). This gives us a uniform way of observing people's locations.

Spatial Resolution (m)	Average Stationarity (%)	Average Regularity (%)	Average Diversity (%)
200	96.0	88.5	90.7
300	96.5	88.9	90.4
400	96.9	89.7	90.2
500	97.2	92.2	89.0
600	97.4	92.5	88.5
700	97.7	92.7	87.7
800	97.9	93.6	87.2
900	98.0	93.7	87.1
1000	98.1	94.0	86.8

Table 4.3: Variation of regularity, stationarity, and diversity according to the spatial resolution of the data for the GPS dataset, in the next-cell prediction task.

However, as mentioned in Section 3.2, some mobility dataset such as social media data are activity dependent, which means users' locations are only measured when the user takes specifc actions (post on social media, for instance). This type of dataset does not exhibit the same uniformity in measuring users' locations as our datasets. As a result, activity dependent datasets may not be able to capture important aspects of users' routines. This type of dataset will exhibit less stationarity, less regularity, and more diversity, when compared to datasets with observations in a fixed time interval.

Previous work [12] proposed strategies to try to reconstruct the users' mobility trace from sparse data. This type of approach will rely on well-known mobility patterns to try to infer where a given person was at a moment for which her location is unknown in the dataset. This type of strategy is also useful because, as previous work argued [58], as data sparsity increases, entropy estimates (and therefore predictability) start to degrade.

4.3 Summary

In this chapter, we investigated proxy metrics (regularity, stationarity, and diversity) that help us make sense of, *i.e.*, interpret, predictability values. Our results show that these metrics capture most of the variability in one's mobility. We also argued that the reduction in entropy (and corresponding increase in predictability) seen as the spatial resolution increases comes at a cost, *i.e.*, there is a trade-off between prediction accuracy and utility, as previously mentioned in Section 2.1.5.

As shown in this chapter, stationarity, regularity, and diversity can be used to study human mobility as a whole, but as discussed in Section 1.2.2, previous studies

proposed to view human mobility in terms of different components. In the next chapter, we investigate the predictability of different components of human mobility by splitting one's mobility into two components and using these metrics to study the predictability of an individual's routine. We validate our results by employing regression models that use these three metrics to fit the predictability of one's routine-related mobility.

Chapter 5

Understanding Predictability of Mobility Components

In this chapter, we continue our efforts towards understanding predictability by investigating the predictability of different components of human mobility. Specifically, we here study predictability by viewing human mobility in terms of two components (routine and novelty) and study their impact on predictability. We then conduct a thorough investigation of the predictability of the routine component of human mobility, as that can lead to new insights for prediction strategies that rely on the history of visited locations.

Although previous work [57] analyzed individual human mobility in terms of *exploration* and *preferential returns*, Song *et al.*'s work and subsequent studies derived from it [39, 56, 15, 60, 62] studied the predictability of a person's mobility considering one's mobility as a single monolithic entity. In this thesis, we propose to study predictability in terms of two components, and we argue that separately studying such components can reveal important insights into the predictability of one's mobility.

Previous work [57] considered an individual's mobility as a collection of visits, each of which being qualified as an exploration (visits to new places), or preferential return (visits to previously visited places). We here adopt the same strategy, and group all *exploration visits* into what we call the *novelty component* of an individual's mobility. Similarly, all visits related to *preferential returns* are grouped into what we call the *routine component* of an individual's mobility. Thus, the *novelty* component consists of locations that the person visited for the first time, and all other visits belong to 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 5.1.1, the novelty component is remarkably unpredictable, mainly because the vast majority of mobility prediction models [27, 53, 13, 7] 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 5.1, we show how to isolate the effect of the novelty component on predictability, thus allowing us to quantify the effect of routine on a person's mobility predictability. Then in Section 5.2, 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, namely regularity, stationarity and diversity, to try to understand what affects a person's routine. 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, described in Chapter 3, of different spatial and temporal granularities, as these properties have been shown to influence predictability [56, 62].

5.1 Components of human mobility

As mentioned, previous predictability 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 thesis, 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 5.1 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.

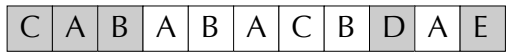


Figure 5.1: Novelty (in gray) and routine (in white) components of input sequence X.

This separation between routine and novelty is a facet of human behavior that appears not only on mobility-related decisions, but also in other scenarios [17]. For instance, in the area of Reinforcement Learning, many algorithms explore the decision space early on, and then exploit paths that lead to a maximum target value. Similarly, in human mobility, the amount of novelty visits in a person's mobility trace tends to decrease over time, as argued in previous work [57].

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 thesis, 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., [39, 53, 15, 28]). 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 [13, 29], 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 thesis, *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 5.1.1). We then present our proposed approach to capture the effect of routine-related mobility on the predictability of one's mobility (Section 5.1.2).

5.1.1 Assessing the effect of novelty on predictability

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 2.1 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 2.1 (reproduced below to facilitate the explanation):

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

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 2.1 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 2.1 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 5.1 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	HWHWR	5	✓	✓
7	HWHWSHWHWR	WHWR	4	✗	✗
8	HWHWSHWHWR	HWR	3	✗	✗
9	HWHWSHWHWR	WR	2	✗	✗
10	HWHWSHWHWR	R	1	✗	✗

Table 5.1: An example illustrating the innerworkings of Equation 2.1 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 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 5.1, 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 2.1, we can rewrite that equation as follows:

$$S_{real} \approx \frac{n \log_2(n)}{\sum\limits_{i \leq n-m} \Lambda_i^{routine} + \sum\limits_{i \leq m} \Lambda_i^{novelty}} = \frac{n \log_2(n)}{\sum\limits_{i \leq n-m} \Lambda_i^{routine} + m}, \tag{5.1}$$

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 2.1, and $\Lambda_i^{routine}$ is the effect of routine on the denominator of Equation 2.1.

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 5.1), 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.

5.1.2 Assessing the effect of routine on predictability

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 2.1, 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 5.2 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 5.2(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 5.2(a).

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

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

Figure 5.2: 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 5.1.1 and shown in 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 5.2(b), 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?*

To tackle 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 5.2(b). 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 5.1. In Section 5.1.1, 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 5.2(d), where the routine component of $X_{baseline}$ is $AAAAAA$, i.e., has size 6. Table 5.2 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 5.2: An example illustrating the innerworkings of Equation 2.1 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 5.2, 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 5.1 to compute the entropy of the baseline sequence as follows:

$$S_{baseline} \approx \frac{n \log_2(n)}{\left\lceil \frac{(k+1)^2}{4} + \frac{k+1}{2} \right\rceil + m}, \tag{5.2}$$

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 5.2 to an input sequence X yields the same entropy value as using Equation 2.1 to compute the entropy of a sequence $X_{baseline}$ such as the one in Figure 5.2(d), *i.e.*a baseline sequence obtained from an input sequence X. In other words, Equation 5.2 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:

$$\Delta_{S_{routine}} = S_{real} - S_{baseline}, \tag{5.3}$$

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 5.2(a). The entropy S_{real} of this sequence is given by:

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

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.*

5.1.3 Characterizing components of human mobility

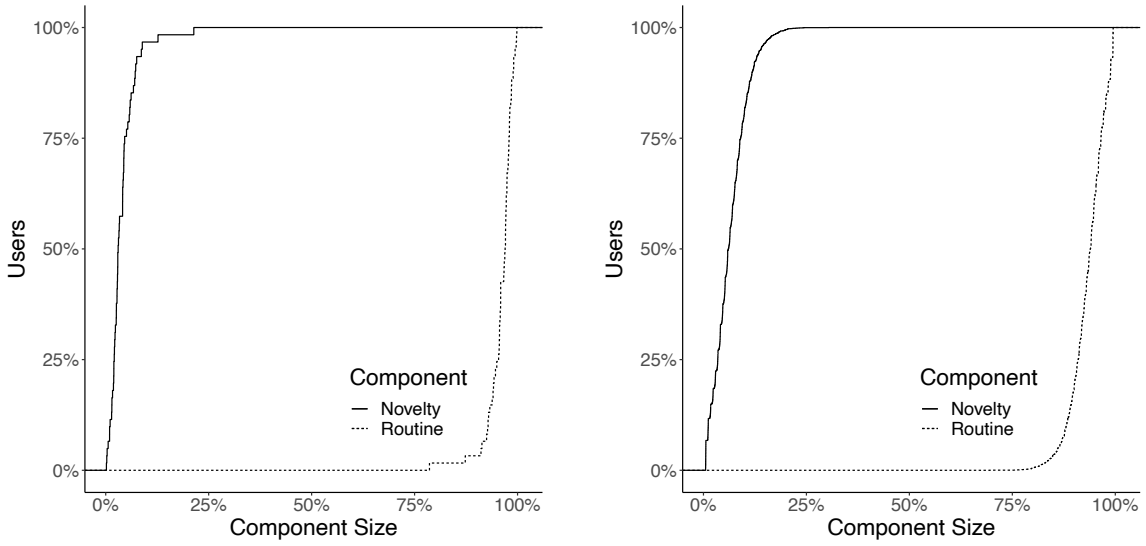
In this section, we analyze the novelty and routine components of a user's mobility in our two datasets. We do so by describing how much each of these components represent in terms of the total mobility trace of each user. Specifically, we compute, for each user in each dataset, the fractions of n, the total number of visited locations, that correspond to visits of the routine and novelty components[1], as defined in Section 5.1. Figure 5.3 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. Furthermore, we note that the routine component is larger in the GPS dataset (which encompasses a larger period of time compared to the CDR dataset), agreeing with previous work [57] which showed that the number of novelty visits decreases over time.

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

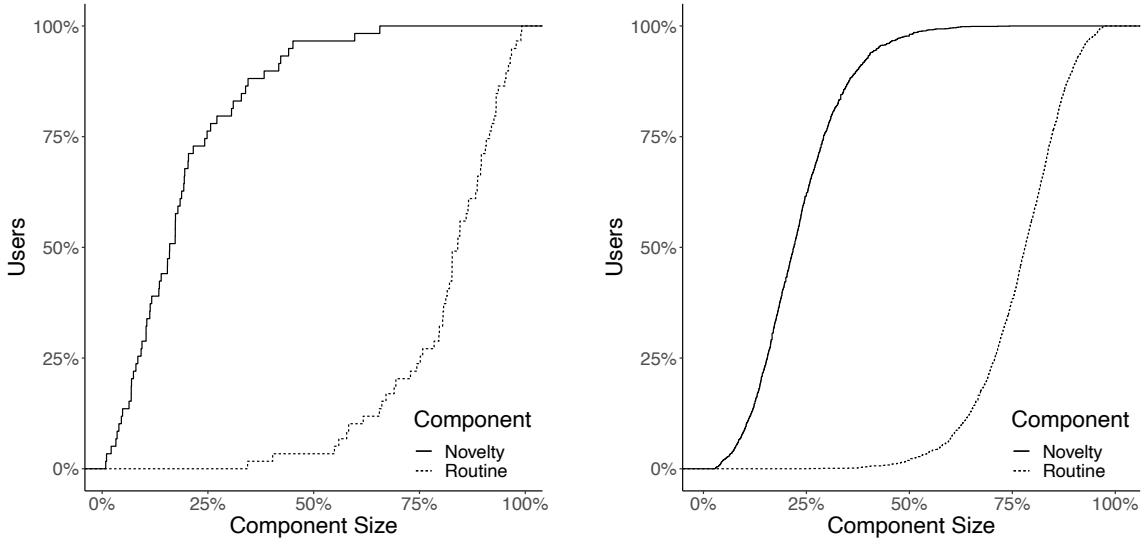
[1]Note that, for a given user, these two fractions sum up to 1.

larger in these cases. These results corroborate previous arguments that the next-place prediction task is harder than next-cell prediction [15, 62]. 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.



(a) GPS dataset (Next-Cell)

(c) CDR dataset (Next-Cell)



(b) GPS dataset (Next-Place)

(d) CDR dataset (Next-Place)

Figure 5.3: Cumulative distributions of the size of the routine and novelty components in our two datasets.

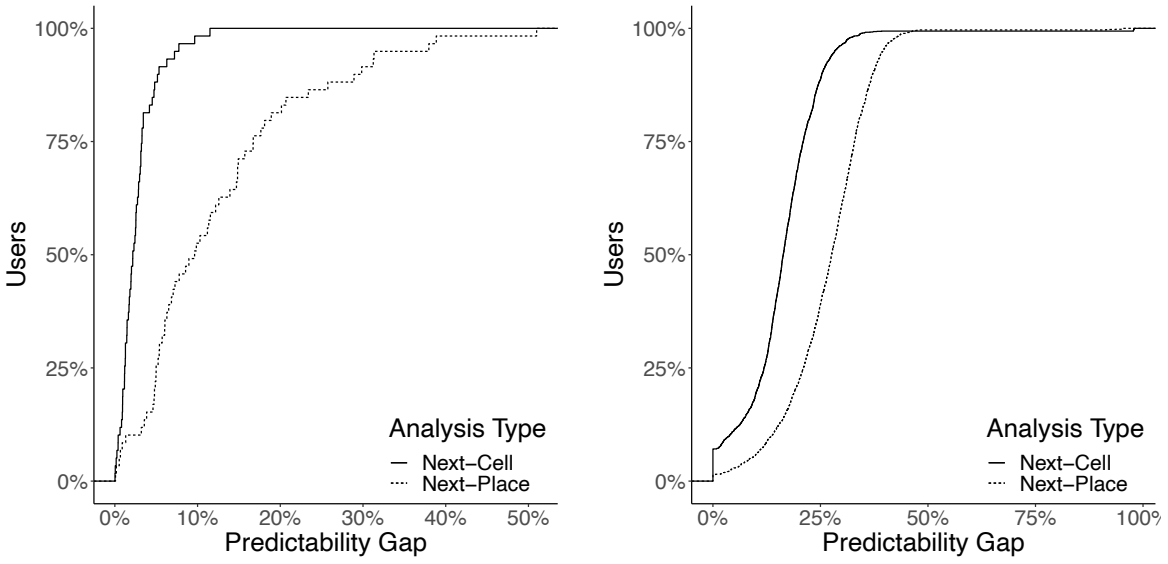
5.2 Investigating the predictability of the routine component

In this section, we study the predictability of the routine component of human mobility. We start by showing the predictability gap between users' actual routine and their baseline routine (Section 5.2.1), and then zoom in on the routine component to understand what affects its predictability.

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.

5.2.1 Predictability gap

Focusing on the routine component, our main interest in this chapter, we now assess the extent to which there is unpredictable behavior in people's routine. To that end, we apply Equation 5.3 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*.



(a) GPS Dataset

(b) CDR dataset

Figure 5.4: Cumulative distributions predictability gap $\Delta_{S_{routine}}$ (Equation 5.3) for both datasets (next-cell and next-place predictions).

Figure 5.4 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 13.0% 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.

5.2.2 Using proxy metrics to study routine

In order to better understand the predictability of the routine component of an individual's mobility, 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 the three metrics, described in Chapter 4 (regularity, stationarity and diversity). In the next section, 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 order to illustrate that these three 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 5.3, 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 5.3

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.

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

5.2.3 Discussion of results

In this section, we present our experimental evaluation of the use of the metrics previously described to explain the predictability associated with routine-related mobility (Section 5.2.3). *Throughout this chapter, 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 5.2-(b)).*

Concretely, we build regression models of increasing complexity, each of which uses some of the metrics discussed in Section 4.1 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 4.1. We then compare the fitted entropy 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 5.2.3.1) and then for the next-place prediction task (Section 5.2.3.2).

5.2.3.1 Next-cell prediction

In this section, we evaluate several regression models that rely on the metrics described in Section 4.1 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:

$$H(X) \approx \alpha + \beta \times reg + \gamma \times st + \nu \times \mu + \epsilon, \tag{5.4}$$

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 5.3, and is given by the produc of those variables.

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

$$H(X) \approx \alpha + \beta \times reg + \gamma \times st + \delta \times div + \nu \times \mu + \epsilon, \tag{5.5}$$

We evaluate the quality of each model for each dataset by the *adjusted* coefficient of determination (*adjusted* R^2). As shown in Table 5.4, 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), and any impact it may have on it is captured by regularity and stationarity. Indeed, from Table 5.3, 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 5.6:

$$H(X) \approx 6.87 - 8.44 \times reg + 1.54 \times st + 3.98 \times div \times -3.96\mu \tag{5.6}$$

From Table 5.4, 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

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

Table 5.4: 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.

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 captured by stationarity and diversity.

Indeed, a simpler model (the RS model which does not use diversity of trajectories), shown in Equation 5.7, produced equivalent results:

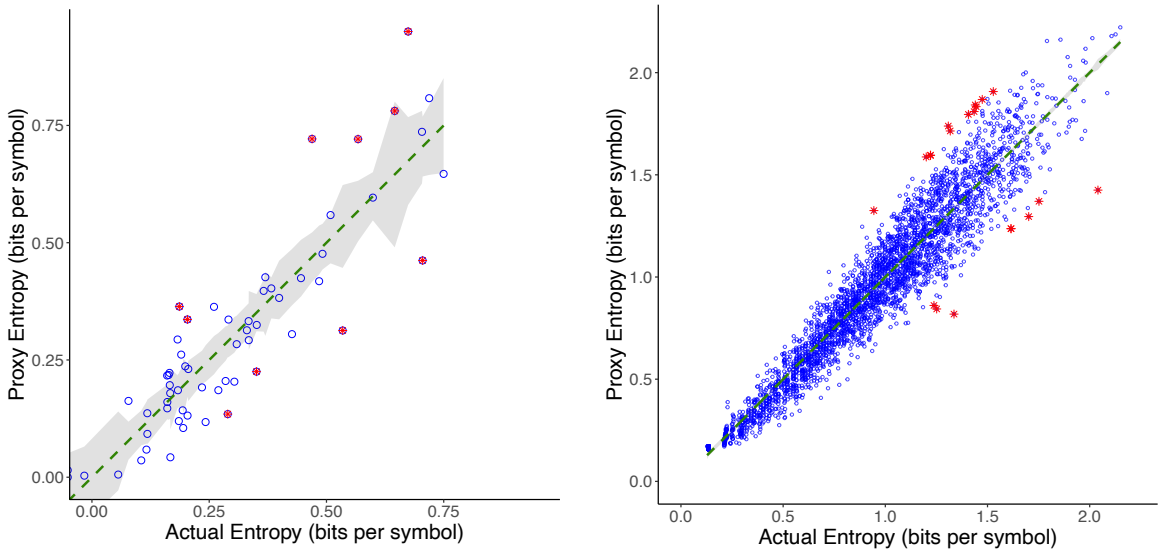
$$H(X) \approx 10.2 - 7.80 \times reg - 2.42 \times st \tag{5.7}$$

We note that the p-value for both coefficients in the model depicted by Equation 5.7 are significant (p-value $< 1 \times 10^{-5}$). 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 5.8, we show the coefficients of our RSD model for the CDR dataset:

$$H(X) \approx -14.1 - 2.00 \times reg + 16.0 \times st + 19.4 \times div \times -19.0\mu \tag{5.8}$$

All of the coefficients of the model in Equation 5.8 are significant (p-value $< 1 \times 10^{-26}$). Furthermore, we note that using *diversity* slightly improved the perfor-



(a) GPS dataset (b) CDR dataset

Figure 5.5: 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.

mance of the model, compared to our RS model, as shown in Table 5.4. We conjecture that this 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.

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 5.5 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.

However, for both datasets, there are a few dots that are farther away from the diagonal, shown in red in Fig. 5.5. 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 correspond to users with long (routine) mobility traces, *e.g.*, more than 1,000 locations. As mentioned, the entropy estimator shown in Equation 2.1 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 2.1 to converge, so there is a gap between the entropy (computed using Equation 2.1) 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 5.2. 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.

5.2.3.2 Next-place prediction

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 5.2.3.1, 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 5.2.3.1.

Because of the lack of stationarity, this prediction task is harder compared to next-cell prediction [15]. In the latter, a large portion of the accuracy in prediction comes

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

Table 5.5: 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.

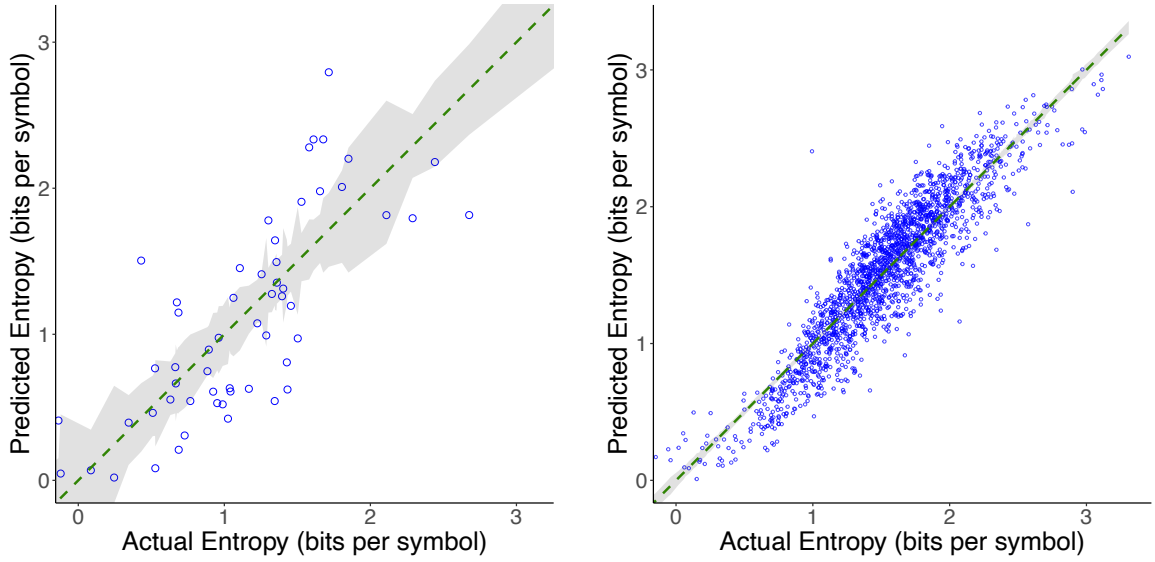
from the fact that people tend to stay for long periods of time in the same location. Thus, models that guess that the user will be at the same location in the next time bin have a higher chance of making a correct prediction. As there is no stationarity in the next-place prediction problem, models have 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 5.4, in the previous section, to those in Table 5.5, which summarizes the performance of our models for the next-place prediction task. 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 *diversity* 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.

Figure 5.6 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



(a) GPS dataset

(b) CDR dataset

Figure 5.6: 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.

also tend to present poor performance in terms of entropy fitting, as was also observed for the next-cell prediction in Section 5.2.3.1.

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 5.5 and Figure 5.6. 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 Chapter 2. As we have 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 chapter, we have shown that three such metrics are enough to explain most of the variability in the entropy of a person's routine mobility.

Open remark: Notice that if we compare the results of our models that try to approximate the entropy of the routine component with those that try to approximate the entropy of the whole sequence, we find that we obtain better results for the whole

sequence. This seems counter-intuitive, since one's routine should be easier to predict than one's whole mobility. However, it is important to point out that entropy estimates are sensitive to the size of the input sequence, *i.e.*, entropy tends to decrease as the size of the sequence increases. When analyzing a person's routine component alone, as novelty is filtered out of the sequence, we obtain smaller input sequences, therefore the entropy estimates for the routine component are not directly comparable to those of the individual's whole mobility.

5.3 Summary

In this chapter, we proposed to study predictability in terms of two components, routine and novelty, with distinct properties. We showed that this 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 proxy metrics 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 results also show 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.

The proxy metrics discussed in Chapter 4 explain most of the variability in a person's overall mobility and routine, but there seems to be something else at play here. Intuitively, one expects that external factors such as day of the week, hour of the day, weather conditions, and even socio-economic factors play a role in a person's mobility patterns. While these types of information affect people's mobility patterns, the state-of-the-art technique for computing the limits of predictability in human mobility does not take them into account. In the next chapter, we investigate how to add such types of (contextual) information into the computation of the limits of predictability.

Chapter 6

Extending Predictability with Contextual Information

Recall from Section 2.2.1 that predictability is a function of the entropy of the sequence of locations. However, given prior arguments that contextual information may indeed improve the predictability of one's mobility [15, 28], we would like to use not only the history of visited locations while computing the entropy, but also contextual information associated with each visit. In this chapter, we study different strategies to incorporate such side information into entropy (and thus, predictability) estimates, quantifying its impact on those estimates.

We start by investigating how to explore context using entropy estimators that are based on the frequency (probability) with which the locations are visited (Section 6.1). We choose to focus first on those entropy estimators, which are alternatives to the state-of-the-art compression-based approach discussed in the previous section, because extending them to incorporate context is easier and more intuitive. After quantifying the impact of context into these entropy estimators, we then move on to explore the more challenging task of adding context to the compression-based estimator used by Song *et al.* (Section 6.3).

In both cases, we consider three types of contextual information, namely day of the week, hour of the day, and weather information. The latter, obtained through an an external service[1], is only available for our CDR dataset as we were unable to gather weather information for the period and location covered by the GPS dataset. For the CDR dataset, the weather information corresponds to descriptions of the weather (clouds, rain, snow, etc.) which are mapped into 7 distinct and unique integer identifiers.

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

6.1 Adding contextual information to predictability estimates

Given $X = (x_1, x_2, \ldots, x_n)$, a time-ordered sequence of locations, and $C = (c_1, c_2, \ldots, c_n)$, a sequence of contextual information associated to each of the visits (c_i could be the weather when the person visited location x_i, for instance), we wish to measure the extent to which knowing sequence C helps estimating the entropy of X. In other words, we wish to know how much X is constrained (or influenced by) C, which can be determined by the *conditional entropy* $H(X \mid C)$ [14], computed as follows:

$$H(X \mid C) = H(X, C) - H(C), \tag{6.1}$$

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

$$H(X, C) = -\sum_{x \in X, c \in C} p(x, c) \log_2 p(x, c), \tag{6.2}$$

and $p(x)$ is the *probability mass function* of variable X given by $p(x_i) = Pr(X = x_i)$. In Equation 6.1, if X and C are independent, i.e., if C carries no information about X, it follows that $H(X, C) = H(X) + H(C)$, which leads to $H(X \mid C) = H(X, C) - H(C) = H(X)$. Once we have $H(X|C)$ we use it in Equation 2.3 to compute the predictability of sequence X constrained by the contextual information in C.

Notice from Equations 6.1 and 6.2 that the basis for entropy computation is an underlying *probability distribution.* Thus, if one has the *full* probability distribution of a sequence of symbols X, the entropy of that sequence is given by Shannon's formula: $H = -\sum p(x) \log_2 p(x)$. The same is true for the joint entropy of X and C. In real world situations, however, one usually has access to only a *sample* drawn from the underlying probability distribution. As a consequence, entropy values obtained for a sequence are *estimates* of the real entropy of the sequence. Entropy estimators that are based on the probability distribution inferred from a sample usually compensate for the effects of using such sample by adding a bias term to their probability estimates. Different estimators exploring different bias terms exist in the literature [26], but in general their entropy estimates tend to be more conservative (than the exact values) because of the added bias term.

In Section 6.2, we investigate how to add context to frequency-based entropy estimators and evaluate three representatives of this type of estimator with contextual information. In Section 6.3, we argue that the aforementioned strategy to add context

to predictability estimates does not work with Song *et al.*'s estimator, which is based on compression. We also propose new strategies to incorporate context into Song *et al.*'s estimators and evaluate the impact of this type of information on predictability.

6.2 Contextual information and frequency-based estimators

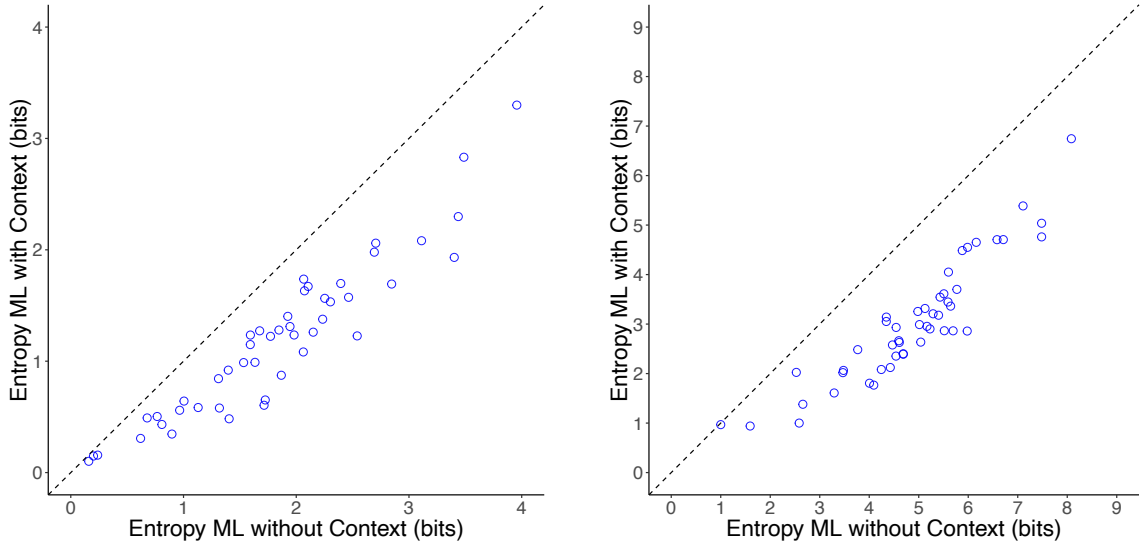
We experimented with various frequency (probability) based entropy estimators, choosing three of them that delivered the best results in our preliminary experiments.[2] The first one is called *Maximum Likelihood* (ML), which estimates entropy using the empirical frequencies of observations, and therefore is equivalent to Shannon entropy [14]. This estimator is also used in Song *et al.*'s work as a baseline for comparison against the more refined compression-based estimator, explained in Section 2.2.1. The second one, called *Miller-Madow* (MM), estimates entropy by applying the Miller-Madow bias correction [10] to Shannon entropy. The third one, called *SG*, estimates entropy using the Dirichlet multinomial pseudo-count model [2] with parameter $a = 1/n$ where n is the length of input sequence X. All of these estimators directly apply Equations 6.1 and 6.2 to compute the predictability of sequence X given sequence C.

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

Table 6.1: Evaluation of three entropy estimators in both datasets and for the two prediction tasks (next-cell and next-place). The reported average entropy values are given in bits per symbol (each location is a symbol in the input sequence). For probability-based entropy estimators, context reduces the entropy of the original sequence.

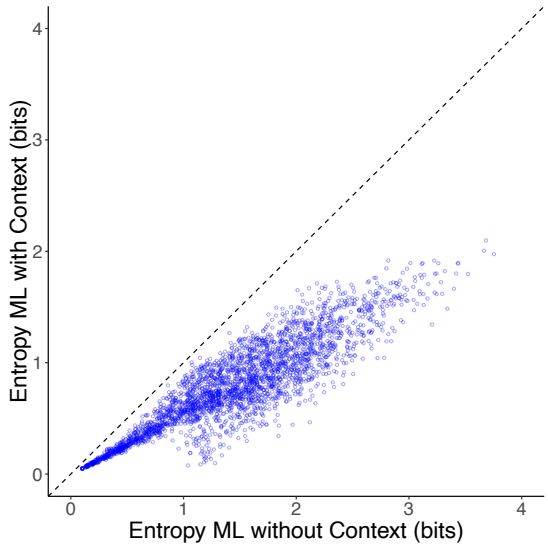
Table 6.1 shows the entropy estimates produced by these three estimators with and without context, for our two datasets, three types of contextual information, and two prediction tasks. As shown in the table, in some situations, *context does enhance predictability estimates*, i.e., entropy values with context are much lower than those without context. The gaps are larger when the hour of the day is used as context, which *suggests stronger ties between hour of the day and location.* For instance, a

[2]These three estimators, along with others, are available as off-the-shelf tools in the R package called *entropy* [26].

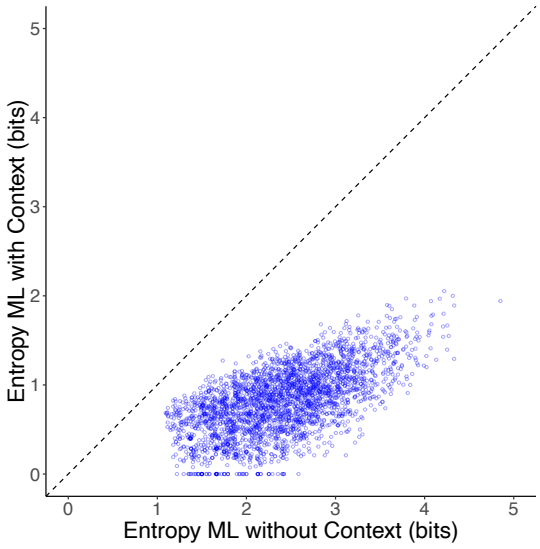


(a) GPS dataset

(c) GPS dataset



(b) CDR dataset



(d) CDR dataset

Figure 6.1: Reduction in the entropy values when contextual information (hour of the day in this case) is applied to the Maximum Likelihood estimator, in both datasets and prediction tasks.

person may visit different locations every day of the week, but almost always stays at home from midnight to early morning, or at workplace during morning and afternoon hours.

We further illustrate these enhancements by showing, in Figure 6.1, scatter plots of entropy values with and without context (hour of day) for the ML estimator (the best of the three estimators in Table 6.1). We consider the best estimator the one which produced lower entropy values. As shown in these figures, the entropy values for all users were reduced when context was used, for both datasets and prediction

tasks. These results confirm the intuition that *human mobility is constrained by several factors*. As Table 6.1 shows, some of these factors can be related to people's routine, e.g., day of the week and hour of the day, but *external* factors such as weather also have the ability to influence people's mobility.

6.3 Contextual information and Song et al.'s estimator

In Section 6.1, we showed that context reduces the entropy of probability-based entropy estimators. However, in the case of the estimator employed in Song *et al.*'s work, it is not possible to exploit contextual information by directly applying Equations 6.1 and 6.2. Recall from Chapter 2 that the algorithm used by Song *et al.* (originally proposed by Kontoyiannis *et al.*[33]) works by compressing the input sequence of symbols to estimate its entropy, therefore leveraging the relation between entropy and compressibility. In doing so, the algorithm becomes oblivious to the underlying probability distribution of the symbols in the sequence, which poses a barrier to computing the conditional entropy using Equations 6.1 and 6.2.

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

6.3.1 Sequence-splitting

Our first approach relies on splitting the original sequence X according to the contextual information into consideration and on computing the entropy for visits that occur with the same context [62]. In other words, we basically hard-code context into each sub-sequence and in the end, use the entropy of those sub-sequences to obtain the entropy of the original one.

We will illustrate how this strategy works through an example, shown in Figure 6.2. Let's assume we want to use weather as contextual information (i.e., sequence C in the figure), discretized into three different types (e.g., sunny, cloudy, rainy, represented by the symbols sun, cloud, and umbrella in the figure). To do that, we split the original sequence X into three sub-sequences, one for each type of weather, each

of which contains all of the locations visited when the weather was of the same given type. We then, run the entropy estimation algorithm in each of the three sequences, taking the weighted average of the results, to consider differences in the size of the sequences.

More formally, let $X = (x_1, x_2, \ldots, x_n)$ be the original sequence and $C = (c_1, c_2, \ldots, c_n)$ be the contextual information sequence. Moreover, let k be the number of *distinct* elements in C, i.e., different contexts in C. We split X into sub-sequences $X_1, X_2, \ldots, X_k$, so that each X_j is the (sub)sequence of locations in the original sequence X which are associated with the same type of context $c_j, j = 1..k$ in sequence C. We then apply Equation 2.1 to each X_j, taking the weighted average at the end, where the weight is the size of each sequence.

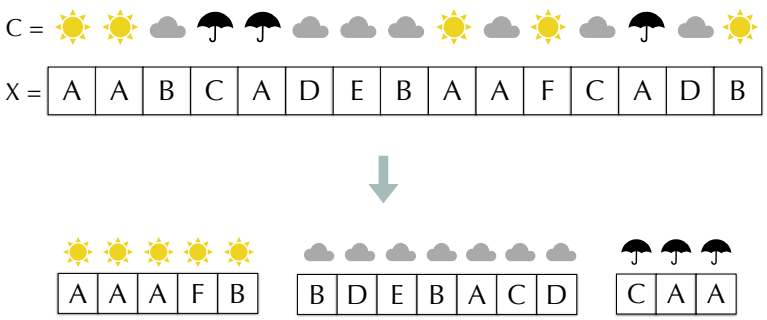


Figure 6.2: Example of our sequence-splitting strategy. We divide the original sequence into sub-sequences according to each type of context.

6.3.2 Sequence-merging

Our second strategy relies on the fact that, by combining locations and contexts in the same sequence, we can estimate their joint distribution using a compression-based estimator. In a nutshell, our sequence-merging approach is based on an analogy with Equation 6.1. We propose to estimate the conditional entropy using a compression-based entropy estimator, such as the one used by Song *et al.*Defining $H_c(X)$ as the *compression-based entropy* of sequence X, and $H_c(X, C)$ as the *joint compression-based entropy* of X and C, we have that:

$$H_c(X \mid C) = H_c(X, C) - H_c(C), \tag{6.3}$$

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

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

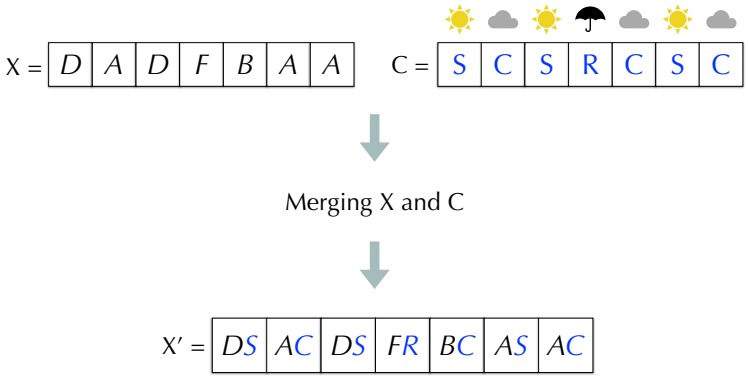


Figure 6.3: Example of our sequence-merging strategy. Each symbol $x_i \in X$ is combined with the associated context $c_i \in C$ to form sequence X'.

The key insight to computing $H_c(X, C)$, illustrated in Figure 6.3, is to merge sequences X and C into a sequence X', where each symbol is now a pair (x, c) with $x \in X$ being a location and $c \in C$ being a context. Recall that, according to Song *et al.*'s estimator, the entropy is inversely proportional to the number of repeated sub-sequences in the target sequence. As we compute the entropy of a sequence using this estimator, we keep track of every sub-sequence encountered, so that further sub-sequences can be matched against the previously discovered ones. Assuming that X is somehow related to C, i.e., there will be repeated location-context pairs throughout the new sequence X', which may help us obtain lower entropy for sequence X by using sequence C as context.

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

6.4 Discussion of results

In this section, we discuss the results for our two approaches (sequence-splitting and sequence-merging) and compare their performance to the best frequency-based estimator from Section 6.2 (the ML estimator). We are considering the best estimator the one which produced lower entropy values. Table 6.2 shows a comparison of these three approaches. For each estimator, the table shows average entropy values with and without context, for both datasets, both prediction tasks, and all types of contexts considered.

Note that for both sequence-splitting and sequence-merging, the results without context are those estimated by the original Song *et al.*'s estimator. The results for the ML estimator are the same as in Table 6.1, shown again here to facilitate comparison.

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

Table 6.2: Evaluation of our sequence-splitting and sequence-merging strategies (compared to the best estimator from Section 6.1) in both datasets and for the two prediction tasks (next-cell and next-place). The reported average entropy values are given in bits per symbol (each location is a symbol in the input sequence).

There are three key observations to make out of the results in Table 6.1. First, we note that in all cases without context, *Song* et al.*'s estimator does produce lower entropy values than the ML estimator.* In other words, it is indeed a very good entropy estimator, justifying its broad use to estimate predictability in human mobility [58]. Second, we also note that introducing context into this estimator, by applying either the sequence-splitting or the sequence merging approach, can *yield lower entropy values than the ML estimator with context* in several cases, especially for the GPS dataset.

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

are reductions on entropy values, especially if hour of the day is used as contextual information.

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

6.4.1 Context-related challenges

In this section, we discuss the challenges associated to using context with compression-based entropy estimators, and how these challenges reflect on our two proposed strategies. Specifically, we discuss challenges related to (i) sequence size, (ii) alphabet size, (iii) context variability, and (iv) how to incorporate other types of context into predictability estimates.

Sequence size. Recall from Equation 2.1 that Song *et al.*'s entropy estimate converges to the real entropy as the sequence grows to infinity, therefore being influenced by the length of the sequence. That is, the larger the sequence the better the entropy estimate. In our sequence-splitting strategy, by dividing the original input sequence, we are effectively estimating the entropy of *smaller* sequences, and *Song* et al.*'s estimator has trouble converging to the real entropy for such small sequences*, and we end up with possibly inflated entropy values.

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

Consider now a more realistic scenario of using hour of the day as contextual information. In that case, the history of locations of each user is split into 24 sequences, one for each hour. *This division results in sequences with considerably smaller sizes, which makes it harder for Song* et al.*'s estimator to converge to the real entropy.* This may be further aggravated by the splitting of longer stationary periods that span more than one hour into separate sequences, which also contributes to raising the final

entropy estimate. This explains why this approach performed poorly (i.e., its entropy with context was higher than the one without context) for the next-cell task, for which longer stationary periods greatly contribute to the real entropy of the original sequence. In the case of the CDR dataset, as the period covered by the data is smaller, and the temporal resolution is lower (fewer observations per time unit), there is reduction in stationarity, as argued in Section 4.2.3, which aleviates the problem we just described. In the case of next-place prediction in the CDR dataset, as there is no stationarity involved, we observe a reduction in entropy values when context is used in the sequence-splitting approach.

Alphabet size. Previous work [23] has shown that the compression-based estimator used in Song *et al.*'s work takes longer to converge to the real entropy of the sequence for sequences with large alphabets. Recall that, in our sequence-merging strategy, when building sequence X', each symbol $x_i \in X$ is combined with the corresponding context $c_i \in C$ to form a tuple location-context in the form (x_i, c_i). This combination of symbols produces a new sequence X' which is much more complex than the original sequence X in terms of unique symbols, as the alphabet of X' is the cartesian product of the alphabets of sequences X and C.

Song *et al.*'s estimator is based on the Lempel-Ziv compression algorithm, which is a universal compressor [14]. These compressors learn the distribution of symbols in the input sequence on-the-fly. Thus, for more complex inputs they may take longer to learn the underlying distribution of symbols. *If the input sequence is not long enough, such methods may produce inaccurate entropy estimates.* Thus, given the higher complexity of X', compared to the original sequence X, Song *et al.*'s estimator may indeed produce quite inflated estimates of entropy, as observed in Table 6.2.

As an example, consider the two following sequences: $X = (A, B, A, B, A, B, A, B, A, B, A, B)$, and $C = (1, 1, 2, 2, 3, 3, 4, 4, 5, 5, 6, 6, 7, 7)$. The entropy of X alone is 1.00, and the entropy of C is 2.66, both computed using Song *et al.*'s estimator. Yet, the entropy $H_c(X \mid C)$ is 1.14, i.e., using C to estimate the entropy of X actually increased the entropy (compared to the entropy without context). It is important to note that, though theoretically context can never increase entropy ("information cannot hurt" property [14]), it is not clear whether the use of context can actually help predictability in practice, when using compression-based entropy estimators such as Song *et al.*'s. In particular, we cannot compute the conditional entropy directly, but only estimate it through universal compressors, which may need a large sequence to learn the input distribution and start approximating the real entropy closely.

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

Incorporation of external contextual information. Recall that in our datasets there is a given piece of context associated to each location that a user visited. This is what we call *internal context*, *i.e.*, context associated to each symbol in the input sequence. It can also be desirable to incorporate what we call *external context* into entropy estimates. External context refers to contextual information that is not associated to each visit, but rather to the environment.

For instance, suppose in a given city all bars close at 10pm, and we wish to know the next location of a user after that time. A prediction model can eliminate all locations that refer to bars from the set of possible next locations. Such elimination of unlikely locations can lead to better prediction accuracy. A similar strategy to filter out unlikely next symbols can lead to lower entropy, and therefore higher predictability. Previous work [56] evaluated a similar strategy, and showed that it indeed leads to high predictability. The strategy adopted was to remove from the set of possible next locations every location that was far away from the current position of the user.

6.5 Summary

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

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

In the next chapter, we summarize our results and discuss possible future directions for research on the topic of predictability in human mobility.

Chapter 7

Conclusions, Limitations, and Future Directions

In this chapter we present a summary of our results, discuss limitations of our techniques, and provide general directions for future explorations of predictability.

Human mobility, being influenced by multiple factors ranging from a person's mood to traffic conditions to the weather, is hard to predict. Knowing how *predictable* someone's mobility can be is also a challenging task. Although Song *et al.*'s [58] and subsequent work [15, 56] made important progress towards understanding predictability, our investigations revealed significant shortcomings in their approach.

This thesis was centered around three goals, each of which aimed at addressing an important shortcoming in the state-of-the-art predictability technique. First, we provided ways to interpret/explain predictability estimates in human mobility. Second, devised techniques for studying the predictability of different components of human mobility. And third, developed techniques to extend predictability estimates with contextual information.

Our goals are translated into three Research Questions:

- Research Question 1 (RQ1): *Would it be possible to trace a given predictability value back to its causes,* i.e.*, to interpret/explain predictability values and to understand what makes a person's predictability higher or lower?*
- Research Question 2 (RQ2): *Would it be possible to use Song* et al.*'s technique to study the predictability of different components of individual human mobility?*
- Research Question 3 (RQ3): *Would it be possible to take* contextual information *into account when using Song* et al.*'s technique, as well as to investigate the*

pros/cons related to the use of contextual information jointly with predictability?

In Chapters 4, 5, and 6, we describe in details our efforts towards tackling each of these RQs, and in the following sections we summarize our results and make concluding remarks.

7.1 Conclusions

In this section, we summarize our results for each of our Research Questions proposed in Section 1.2.

RQ1: Understanding predictability in human mobility Towards tackling this research question, first delved into how Song *et al.*'s technique works, and proposed metrics to serve as proxies for predictability, showing that these metrics capture most of the variability in one's predictability. Our experiments were conducted on two datasets of distinct properties, discussed in Section 3.2, and for two different prediction tasks, described in Section 2.1.3. Our decision to use metrics that capture a person's predictability was motivated by the fact that Song *et al.*'s predictability technique is based on a sophisticated compression algorithm, which makes it hard to look at the output of the algorithm and reason what caused such output.

Previous work had showed that predictability is proportional to the amount of stationary periods, *i.e.*, periods when one stays at the same location for a given time, in an individual's mobility. We then reasoned that we could propose a metric that measures the amount of stationarity in one's mobility and use it to explain predictability (Section 4.1.1). We noticed, however, that stationarity alone is not able to explain predictability.

We hypothesized that other metrics are needed to better explain predictability. Our intuition was that the key to these metrics lies in how Song *et al.*'s entropy estimator works, *i.e.*, what patterns it captures. Our investigation showed that this estimator, being based on compression, outputs an entropy estimate that is related to the number of repeated sub-sequences (patterns) in one's mobility. We then leveraged this knowledge to propose another metric, called regularity, that, together with stationarity, could help to understand the predictability of one's mobility (Section 4.1.2). Similarly, we also proposed a third metric, called diversity, which together with stationarity and regularity, paints a clearer picture of the patterns in an individual's mobility.

In order to check the effectiveness of our metrics, we proposed regression models that use them as proxies of one's predictability. Our results, which encompass both

next-cell and next-place prediction (described in Sections 4.2.1 and 4.2.2, respectively), show that stationarity, regularity, and diversity are able to capture most of the variability in a person's predictability. For instance, in the next-cell prediction task, the adjusted R^2 of our model is 77% and 93.5% for the GPS and CDR datasets, respectively. As for next-place prediction, adjusted R^2 of our model is 85.5% and 91.3% for the GPS and CDR datasets, respectively.

Our results suggest that predictability can be captured through proxy metrics such as the ones we have proposed. Additionally, as mentioned, our experiments show that these simple metrics are capable to capture most of the variability in one's predictability. These results are encouraging and point us in the direction of investigating additional metrics that could improve the R^2 of our models as well as to using these metrics to better understand the predictability of an individual's mobility.

RQ2: Investigating predictability of components of human mobility Towards addressing RQ2, we first identified a gap in the way individual human mobility is commonly modeled in the literature and the way Song *et al.*'s predictability technique is used. While previous studies showed that individual human mobility can be modeled in terms of explorations and preferential returns, Song *et al.*'s technique views one's mobility as a single, monolithic entity, thus preventing the analysis of the predictability of different components of human mobility.

To bridge this gap, we propose a technique to break one's mobility into two components (novelty and routine), which naturally map to explorations and preferential returns. These components possess different properties: absence of visitation history in the novelty component, and higher potential for prediction, due to more regular behavior, in the routine component.

Our technique allows us to estimate the impact of each of these components on predictability, and one of its by-products is a closed-formula to estimate the impact of novelty and routine on the predictability of individual human mobility. We use this technique to isolate each component of human mobility, first providing a characterization of these components and then we focus on studying the routine component of one's mobility, which possess a higher potential for prediction.

We validate our results by applying regression models that use the three metrics proposed in Chapter 4 to explain routine-related predictability. 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 models were able to explain up to 96% of the variability in an individual's routine.

Our results also show 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.

RQ3: Extending predictability with contextual information We here describe our results towards tackling our third research question, namely to extend predictability estimates with contextual information. Concretely, this research goal can be broken down into two main objectives: (i) to propose ways to use contextual information with Song *et al.*'s technique, and (ii) to evaluate the impact of this type of information on predictability.

Towards tackling this research goal, we have (i) evaluated the impact of contextual information on predictability estimates by employing alternative, frequency-based entropy estimators, (ii) described the challenges associated with using contextual information with Song *et al.*'s estimator, (iii) proposed new techniques to seamlessly incorporate context into Song *et al.*'s estimator, and (iv) evaluated the impact of contextual information both on frequency-based entropy estimators and on Song *et al.*'s estimator.

Specifically, we found that when using frequency-based entropy estimators, context tended to lead to higher predictability. However, when using our techniques that incorporate context into the more robust, compression-based estimator originally used by Song *et al.*, the use of contextual information did not always increase predictability. Our investigations suggest that this behavior is due to the sensitiveness of the compression-based estimator to the size of the input sequences. Our results also suggest that these techniques could be used to decide whether some piece of context will be useful for prediction *before* actually using them to train a model.

7.2 Limitations

Throughout this thesis, we discussed and addressed several shortcomings of the predictability technique. However, there are some inherent limitations and issues to this technique that are also worth mentioning.

The data issue In order to obtain a robust estimate of a person's predictability, we need long data sequences. This means that we need the user to share location data for an extended period of time. Additionally, the more fine-grained the spatiotemporal

resolution, the more robust the predictability estimate. However, obtaining this type of data is often a challenge. Most datasets do not have all of these desired properties, and thus it is hard to study predictability in an ideal setting.

The privacy issue The predictability technique studied in this thesis relies on user-level location data. Such data is privacy-sensitive, as it reveals personal information about people.[1] Relying on users to share such sensitive information for research studies can make it hard to advance the research on predictability. In our datasets, user identifiers are anonymized, but ensuring people that their data will remain anonymous and ensuring such anonymity is also a challenge that one has to face when studying user-level mobility.

The computational issue In order to ensure people that their data will remain in their possession and will not be shared with other parties, one can offer to collect the data on the user's device and perform all the computations locally. Thus, the user's data would remain in their device, and only the results of the computations (in our case, the user's predictability) would be shared with an application. This practice addresses some privacy problems, but raises other issues. For instance, performing certain computations on a mobile phone can be restrictive in terms of battery consumption. Moreover, certifying the correctness of certain analyses does require that we actually inspect the data, which is not possible if the data remains in the user's device.

7.3 Future directions

In this section, we describe possible practical applications of predictability and delineate general steps that could be taken when tackling these practical uses of predictability.

Using predictability as a baseline for complex models

Overview The idea here is to use predictability values as a baseline for more complex models (non-universal predictors). The argument is that the cost of a complex model is worth it only when the model provides higher accuracy than the predictability value of a dataset.

[1] `https://www.nytimes.com/interactive/2018/12/10/business/location-data-privacy-apps.html`

As argued in Section 2.2.1, predictability values hold only when *universal predictors* are used. These are simpler, faster, and easy-to-interpret predictors that, in certain cases, lead to high prediction accuracy, as evidenced by the results of previous work [39].

The predictability technique gives us an upper bound on the accuracy that these predictors can achieve on a given dataset, and it has been shown that this upper bound is indeed reachable [39]. Here, we argue that the predictability techniques discussed in this thesis can also be used to evaluate the practical utility of more complex models that do not qualify as universal predictors.

When performing predictions, one is often tempted to employ the most sophisticated model, hoping that such model will lead to higher accuracy than simpler models. However, in general, the more complex the model, the higher its cost (in terms of computing power, time, energy, and even carbon footprint). Thus, the use of such models should be motivated by real needs.

We here argue that our predictability techniques can be used to decide whether it is worth to use complex models for a given prediction task, by using it as a baseline for the accuracy of such models. The idea is that, while evaluating a complex model, if its accuracy does exceed the maximum accuracy that a simpler model can achieve (which is given by the predictability value), the use of a complex model is not worth its cost, and one could rather invest on fine-tuning a simpler universal predictor.

For instance, let us assume we want to perform mobility prediction, and we have a range of options for which prediction model to use in production. During an initial phase, we can evaluate the performance of all the models, some of which are universal predictors, and some which are not. In this initial phase, we also compute the predictability of our dataset. If the accuracy of the non-universal predictors exceeds the predictability values, we keep them as candidates to be used in production. Otherwise, we try to tune the universal predictors so they can eventually reach their maximum possible accuracy, obtained via predictability.

Additionally, for small sequences, it might also be better to use a universal predictor, as non-universal predictors usually require more data to be trained. On the other hand, a non-universal predictor is suitable if there is enough data to train it (long input sequences), and its performance is greater than the predictability of the input sequence.

This simple idea can be used to answer a range of interesting questions:

- What are the characteristics of universal predictors that achieve the maximum accuracy obtained via predictability?

- What features do these models used to reach such accuracy? If we use the same set of features in a more complex model, do we get higher accuracy?

- Previous work has shown that the maximum accuracy obtained via predictability can be surpassed in certain cases [34], but what are the characteristics of more complex models which surpass the maximum accuracy obtained via Song *et al.*'s technique?

- What features do these models use to perform prediction? What accuracy do we get by using the same features in a universal predictor such as a Markov-based model?

Using predictability to assess confidence in predictions

Overview The idea here is to use predictability values as a measure of confidence in predictions in situations where a misprediction can have a high cost. The argument is that if the cost associated to a misprediction is high, and we have low confidence in the accuracy of the prediction, it might be better not to take the risk of a misprediction and rather take a more conservative approach.

Recall that predictability provides a value in the range $[0, 1]$, with 0 meaning completely unpredictable, and 1 meaning totally predictable. In other words, if a person's predictability is 0.8, it means that an ideal prediction model could accurately predict her next location at most 80% of the time, according to the predictability technique. In this example, looking at the 80% value in terms of the confidence in a model's prediction gives us valuable information, as it is more likely that a model will make a correct prediction for a person whose predictability is higher. As mentioned in Section 1.2, we here claim that predictability could be used as a measure of *confidence* in predictions so as to improve location-based systems that rely on the prediction of individuals' whereabouts.

One of the challenges associated to the aforementioned task is to establish a *cost model* to determine how predictability relates to confidence in predictions. Concretely, we wish to specify a function $f(x)$ whose input is a person's predictability and the output is the confidence in predictions related to that person's mobility. Determining the components and shape of $f(x)$ is a research challenge in and of itself, as there are many factors that could play a role in this function, and some of them are application-dependent.

In the following example, we illustrate a simple cost model in which $f(x)$ is a linear function of x. Suppose, for instance, that we want to predict a sequence of $n = 1,000$ events, given the following rules. For every event we make a wrong prediction, we pay a price of $x = \$20$. If choose not to predict the next event in the sequence, we pay $y = \$10$. And if we correctly predict the next event in the sequence, we pay nothing. Our goal is to minimize the amount paid at the end of the sequence of events.

In this example, the largest amount will be paid if we incorrectly predict every event in the sequence, resulting in a payment of $1,000 \times \$20 = \$20,000$, and the smallest amount paid is 0, if we correctly predict every event in the sequence. A middle ground is reached if we choose not to predict any event in the sequence, in which case we end up paying $1,000 \times \$10 = \$10,000$.

To better understand what is advantageous in this situation, consider that we correctly predict the events in the sequence 50% of the time, which would also result in a payment of $500 \times \$0 + 500 \times \$20 = \$10,000$. A prediction accuracy of 60% results in a payment of $600 \times \$0 + 400 \times \$20 = \$8,000$. Thus, if we are confident that we will correctly predict events in the sequence more than 50% of the time, it is better to always perform predictions. Conversely, if our confidence is less than 50%, it is better not to perform any predictions at all.

In the example above, our confidence in the accuracy of predictions is the key to decide whether or not to perform predictions. In the example, the threshold above which we decide to perform predictions is 50%, but this value may change depending on the values of n, x, y, and $f(x)$ which can vary in different applications. Furthermore, as mentioned above, this cost model considers that $f(x)$ is a linear function of the user's predictability, but depending on the application, $f(x)$ could be a different function.

The example above serves to illustrate that assessing the confidence in a given prediction can be useful when the cost of a misprediction is high. In these scenarios, a misprediction can result in lower quality of service or resource waste. Consider, for instance, the case of 5G networks, where there is a need for fine-grained management of user mobility [45].

In these networks, computing nodes have to be close to end-users so as to provide ultra-low latency, reliability, and scalability, and the network management system may offload certain computations to computing nodes that are closer to the end users. Our hypothesis is that mispredicting the users' next location can waste resources in this type of situation.

Suppose, for instance, that a user's next location is assumed to be near node n_i, so the network management system decides to offload computations to that node, but it turns out that the user's next location is actually near node n_j. A misprediction

such as this will waste computational resources, as the computation will have to be performed again at node n_j, and the latency of the network will increase. The problem in this case is that the cost of a misprediction is higher than making no prediction at all. At the same time, it is desirable to make predictions about the users' next location, as correct predictions can result in better quality of service for the users of the network.

Thus, we propose to investigate different cost models that would allow us to decide whether or not to trust certain predictions. For instance, depending on the output of the function $f(x)$ for a given user, the system may decide whether to (i) always offload computations for that user, (ii) to offload computations only when the load of the system is relatively low, *i.e.*, the cost of a misprediction will not affect the overall performance of the system, or (iii) not to offload the user's computations at all.

One possible direction to test the validity of this approach is to implementing it into a network simulator that takes into account mobility information. To that end, we plan to perform the following key-steps:

- Compute the predictability of each user in the system beforehand and store that information to be used later.
- Measure the average latency of the system in normal conditions, *i.e.*, when predictability information is not used to assess confidence in predictions.
- Build different cost models, with different functions that take as input a user's predictability and outputs the confidence in predictions associated to that user.
- Evaluate the impact of these different cost models on the overall latency of the system.

Specifically, this strategy can be used to answer the following research questions:

- Does the use of predictability information reduce the latency of the system?
- How does the reduction in latency relate to the confidence in predictions?
- What is the shape and components of the function that leads to lower latency overall?
- What other terms are relevant to the cost model (*e.g.*, error penalty, number of users in the system, current load of the system, and so on)?

Predictability as a measure of susceptibility

Overview The idea here is to use predictability values as a measure of how open a person is to novelty in a recommendation scenario. The argument is that less predictable people are less attached to their routine, and are thus more open to novelty, and more predictable people will tend to deviate less from their routine, and therefore will be less open to novelty.

We here describe a way to use predictability as a measure of susceptibility of users to new things and discuss how a practical application could leverage this type of information. Specifically, we argue that a possible case study to explore this idea consists of using predictability information to calibrate novelty in a place recommendation scenario. In this scenario, it is often a problem to decide which items to recommend to a user. For a given user, the recommender system has to decide whether to suggest a place the user has already been to or a place he or she never visited before but may like, based on his or her preferences.

The idea here is that more predictable users will have a more strict routine and less predictable users will be less influenced by routine. If routine does not play such a large role in someone's daily activities, they may have a higher propensity to explore new places. Thus, predictability information could be used as a way to quantify this propensity and subsequently use it to calibrate the amount of new (previously unseen) places that the system recommends to the users. These systems usually suggest to someone a list of possible places (and not only one place) so knowing whether the user is more open to new places can be useful to determine the number of new (previously unseen) places to show her.

This idea could be implemented in the following way. One could pick an existing recommender system and existing benchmarks on which this system was evaluated. One would then run the system as usual, *i.e.*, without using predictability information, and compute the amount of recommendations that the users followed. In this case, we consider that a person followed the system's recommendation if their next location is one of the locations that the system showed them. Then, one would compute the predictability of every user in the system and establish a relation between predictability and novelty in recommendation. For instance, suppose a user's predictability is 70%. The simplest approach may be to build a list of recommended places to show the user such that 70% are previously visited places and 30% are new places.

This strategy could be used to answer the following questions:

- Does the use of predictability information increase the amount of recommendations that users follow?
- Do less predictable users usually follow the system's recommendations, or to they tend to explore new locations of their own choosing?
- Conversely, are more predictable users more prone to follow the system's recommendation?

Concluding remarks In this section, we described future directions for predictability studies, focusing on possible uses of predictability in practical scenarios. The predictability technique was originally proposed as a theoretical measure, but as argued in Section 1.2 and described here, this technique also has important practical applications. We note that finding these practical applications was the result of a deep theoretical investigation and experimental evaluation the state-of-the-art predictability technique conducted in this thesis. This highlights the importance of more foundational work such as ours, which often leads to interesting avenues of research.

Bibliography

[1] M. Lin, W. J. Hsu and Z. Qi Lee (2012). Predictability of individuals' mobility with high-resolution positioning data. In *ACM Conference on Ubiquitous Computing*, pages 381–390.

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

[3] Amichi, L., Viana, A. C., Crovella, M., and Loureiro, A. A. (2020). Understanding individuals' proclivity for novelty seeking. In *Proceedings of the 28th International Conference on Advances in Geographic Information Systems*, pages 314--324.

[4] Asgari, F., Gauthier, V., and Becker, M. (2013). A survey on human mobility and its applications. *arXiv preprint*.

[5] Bagrow, J. P., Liu, X., and Mitchell, L. (2019). Information flow reveals prediction limits in online social activity. *Nature Human Behaviour*, 3(2):122–128. ISSN 2397-3374.

[6] Barbosa-Filho, H., Barthelemy, M., Ghoshal, G., James, C. R., Lenormand, M., Louail, T., Menezes, R., Ramasco, J. J., Simini, F., and Tomasini, M. (2017). Human Mobility: Models and Applications. working paper or preprint.

[7] Beiró, M. G., Panisson, A., Tizzoni, M., and Cattuto, C. (2016). Predicting human mobility through the assimilation of social media traces into mobility models. *EPJ Data Science*, 5(1).

[8] Brockmann, D., Hufnagel, L., and Geisel, T. (2006). The scaling laws of human travel. *Nature*, 439.

[9] Calegari, R., Musolesi, M., Raimondi, F., and Mascolo, C. (2007). Ctg: A connectivity trace generator for testing the performance of opportunistic mobile systems. In *Proceedings of the the 6th Joint Meeting of the European Software Engineering*

Conference and the ACM SIGSOFT Symposium on The Foundations of Software Engineering, ESEC-FSE '07, pages 415--424, New York, NY, USA. ACM.

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

[11] Chen, G., Carneiro Viana, A., Fiore, M., and Sarraute, C. (2019a). Complete Trajectory Reconstruction from Sparse Mobile Phone Data. *EPJ Data Science*.

[12] Chen, G., Viana, A. C., Fiore, M., and Sarraute, C. (2019b). Complete trajectory reconstruction from sparse mobile phone data. *EPJ Data Science*, 8(1):30.

[13] Cho, E., Myers, S. A., and Leskovec, J. (2011). Friendship and mobility: User movement in location-based social networks. In *Proc. International Conference on Knowledge Discovery and Data Mining*.

[14] Cover, T. M. and Thomas, J. A. (2012). *Elements of Information Theory*. John Wiley & Sons.

[15] Cuttone, A., Lehmann, S., and González, M. C. (2018). Understanding predictability and exploration in human mobility. *EPJ Data Science*.

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

[17] Domingos, P. (2018). The master algorithm: How the quest for the ultimate learning machine will remake our world.

[18] Dong, W., Duffield, N., Ge, Z., Lee, S., and Pang, J. (2013). Modeling cellular user mobility using a leap graph. In *Proc. 14th International Conference on Passive and Active Measurement*.

[19] Feder, M., Merhav, N., and Gutman, M. (1992). Universal prediction of individual sequences. *IEEE transactions on Information Theory*, 38(4):1258--1270.

[20] Ghouti, L. (2016). Mobility prediction in mobile ad hoc networks using neural learning machines. *Simulation Modelling Practice and Theory*, 66:104--121.

[21] Gonzalez, M. C., Hidalgo, C. A., and Barabasi, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, 453.

[22] Gusfield, D. (1997). Algorithms on stings, trees, and sequences: Computer science and computational biology. *Acm Sigact News*, 28(4):41--60.

[23] Han, Y., Sun, W., and Zheng, B. (2018). Ineffectiveness of dictionary coding to infer predictability limits of human mobility.

[24] Hariharan, R. and Toyama, K. (2004). Project lachesis: Parsing and modeling location histories. In Egenhofer, M. J., Freksa, C., and Miller, H. J., editors, *Geographic Information Science*, pages 106--124, Berlin, Heidelberg. Springer Berlin Heidelberg.

[25] Hasan, S., Zhan, X., and Ukkusuri, S. V. (2013). Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. In *International Workshop on Urban Computing*.

[26] Hausser, J., Strimmer, K., and Strimmer, M. K. (2012). *Package 'entropy'*.

[27] Hess, A., Hummel, K. A., Gansterer, W. N., and Haring, G. (2016). Data-driven human mobility modeling: A survey and engineering guidance for mobile networking. *ACM Computing Surveys*, 48(3).

[28] Ikanovic, E. L. and Mollgaard, A. (2017). An alternative approach to the limits of predictability in human mobility. *EPJ Data Science*, 6(1):12. ISSN 2193-1127.

[29] Jeong, J., Leconte, M., and Proutiere, A. (2016). Cluster-aided mobility predictions. In *IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications*, pages 1--9. IEEE.

[30] Jiang, S., Fiore, G. A., Yang, Y., Ferreira, Jr., J., Frazzoli, E., and González, M. C. (2013). A review of urban computing for mobile phone traces: Current methods, challenges and opportunities. In *Proc. 2nd ACM International Workshop on Urban Computing*.

[31] Jurdak, R., Zhao, K., Liu, J., AbouJaoude, M., Cameron, M., and Newth, D. (2015). Understanding human mobility from twitter. *Plos One*.

[32] Karamshuk, D., Boldrini, C., Conti, M., and Passarella, A. (2011). Human mobility models for opportunistic networks. *IEEE Communications Magazine*, 49(12):157–165. ISSN 0163-6804.

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

[34] Kulkarni, V., Mahalunkar, A., Garbinato, B., and Kelleher, J. D. (2019). Examining the limits of predictability of human mobility. *Entropy*.

[35] Lempel, A. and Ziv, J. (2006). On the complexity of finite sequences. *IEEE Trans. Inf. Theor.*, 22(1):75--81. ISSN 0018-9448.

[36] Li, M. and Vitányi, P. M. B. (1990). Kolmogorov complexity and its applications. In van Leeuwen, J., editor, *Handbook of Theoretical Computer Science (Vol. A)*, pages 187--254. MIT Press, Cambridge, MA, USA.

[37] Li, M., Westerholt, R., Fan, H., and Zipf, A. (2018). Assessing spatiotemporal predictability of lbsn: a case study of three foursquare datasets. *GeoInformatica*, 22(3):541--561. ISSN 1573-7624.

[38] Lin, M., Hsu, W.-J., and Lee, Z. Q. (2013). Modeling high predictability and scaling laws of human mobility. In *2013 IEEE 14th International Conference on Mobile Data Management*, volume 2, pages 125--130. IEEE.

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

[40] Ma, S., Zheng, Y., and Wolfson, O. (2013). T-share: A large-scale dynamic taxi ridesharing service. In *Proc. IEEE International Conference on Data Engineering*.

[41] Merhav, N. and Feder, M. (1998). Universal prediction. *IEEE Transactions on Information Theory*, 44(6):2124--2147.

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

[43] Munjal, A., Camp, T., and Navidi, W. C. (2011). Smooth: A simple way to model human mobility. In *Proc. 14th ACM Int. Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*.

[44] Nain, P., Towsley, D., Liu, B., and Liu, Z. (2005). Properties of random direction models. In *Proceedings IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies.*, volume 3, pages 1897–1907 vol. 3. ISSN 0743-166X.

[45] Orsino, A., Guo, W., and Araniti, G. (2017). Multi-scale mobility models in the forthcoming 5g era: A general overview. *IEEE Vehicular Technology Magazine*.

[46] Ozturk, M., Gogate, M., Onireti, O., Adeel, A., Hussain, A., and Imran, M. A. (2019). A novel deep learning driven, low-cost mobility prediction approach for 5g cellular networks: The case of the control/data separation architecture (cdsa). *Neurocomputing*, 358:479--489.

[47] Palmer, J. R., Espenshade, T. J., Bartumeus, F., Chung, C. Y., Ozgencil, N. E., and Li, K. (2013). New approaches to human mobility: Using mobile phones for demographic research. *Demography*, 50(3):1105--1128.

[48] Pappalardo, L., Simini, F., Rinzivillo, S., Pedreschi, D., Giannotti, F., and Barabási, A.-L. (2015). Returners and explorers dichotomy in human mobility. *Nature communications*, 6(1):1--8.

[49] Pulliyakode, S. K. and Kalyani, S. (2014). A modified ppm algorithm for online sequence prediction using short data records. *IEEE Communications Letters*, 19(3):423--426.

[50] Ranjan, G., Zang, H., Zhang, Z.-L., and Bolot, J. (2012). Are call detail records biased for sampling human mobility? *SIGMOBILE Mob. Comput. Commun. Rev.*, 16(3):33--44. ISSN 1559-1662.

[51] Shannon, C. E. and Weaver, W. (1998). *The mathematical theory of communication*. University of Illinois press.

[52] Silva, T. H., Viana, A. C., Benevenuto, F., Villas, L., Salles, J., Loureiro, A., and Quercia, D. (2019). Urban computing leveraging location-based social network data: a survey. *ACM Computing Surveys (CSUR)*, 52(1):1--39.

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

[54] Simini, F., González, M. C., Maritan, A., and Barabási, A.-L. (2012a). A universal model for mobility and migration patterns. *Nature*, 484(7392):96.

[55] Simini, F., González, M. C., Maritan, A., and Barabási, A.-L. (2012b). A universal model for mobility and migration patterns. *Nature*, 484(7392):96--100.

[56] Smith, G., Wieser, R., Goulding, J., and Barrack, D. (2014). A refined limit on the predictability of human mobility. In *2014 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 88--94. IEEE.

[57] Song, C., Koren, T., Wang, P., and Barabási, A.-L. (2010a). Modelling the scaling properties of human mobility. *Nature Physics*, 6(10).

[58] Song, C., Qu, Z., Blumm, N., and Barabási, A.-L. (2010b). Limits of predictability in human mobility. *Science*, 327(5968).

[59] Song, L., Kotz, D., Jain, R., and He, X. (2006). Evaluating next-cell predictors with extensive wi-fi mobility data. *IEEE Transactions on Mobile Computing*, 5(12):1633–1649. ISSN 1536-1233.

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

[61] Teixeira, D. D. C., Viana, A. C., Almeida, J. M., and Alvim, M. S. (2021). The impact of stationarity, regularity, and context on the predictability of individual human mobility. *ACM Trans. Spatial Algorithms Syst.*, 7(4). ISSN 2374-0353.

[62] Teixeira, D. d. C., Viana, A. C., Alvim, M. S., and Almeida, J. M. (2019b). Deciphering predictability limits in human mobility. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '19, pages 52--61, New York, NY, USA. ACM.

[63] Terroso-Sáenz, F., Cuenca-Jara, J., González-Vidal, A., and Skarmeta, A. F. (2016). Human mobility prediction based on social media with complex event processing. *International Journal of Distributed Sensor Networks*, 12(9):5836392.

[64] Treurniet, J. (2014). A taxonomy and survey of microscopic mobility models from the mobile networking domain. *ACM Computing Surveys*.

[65] Walker, G. T. (1931). On periodicity in series of related terms. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 131(818):518--532.

[66] Wang, H., Zeng, S., Li, Y., and Jin, D. (2020). Predictability and prediction of human mobility based on application-collected location data. *IEEE Transactions on Mobile Computing*.

[67] Wang, J., Kong, X., Xia, F., and Sun, L. (2019). Urban human mobility: Data-driven modeling and prediction. *SIGKDD Explor. Newsl.*, 21(1):1--19. ISSN 1931-0145.

[68] Xu, P., Yin, L., Yue, Z., and Zhou, T. (2019). On predictability of time series. *Physica A: Statistical Mechanics and its Applications*, 523:345 – 351. ISSN 0378-4371.

[69] Xu, T., Xu, X., Hu, Y., and Li, X. (2017). An entropy-based approach for evaluating travel time predictability based on vehicle trajectory data. *Entropy*, 19(4). ISSN 1099-4300.

[70] Zhang, C., Zhang, K., Yuan, Q., Zhang, L., Hanratty, T., and Han, J. (2016). Gmove: Group-level mobility modeling using geo-tagged social media. In *KDD: proceedings. International Conference on Knowledge Discovery & Data Mining*, volume 2016, page 1305. NIH Public Access.

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

[72] Zheng, Y., Capra, L., Wolfson, O., and Yang, H. (2014). Urban computing: Concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology*, 5.

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

Appendix A

In Section 2.2.1, 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 2.1, 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 5.1.1, 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 5.1.2, 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, $\Lambda_i = |q| + 1$ [33]. 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 t 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.