Impacts of novelty seeking on predictability in human mobility

Licia Amichi, Aline Carneiro Viana, Mark Crovella, Antonio Loureiro

Impacts of novelty seeking on predictability in human mobility

Licia Amichi, Aline Viana Carneiro, Mark Crovella, and Antonio Loureiro

***Abstract*—Predicting how humans move within space and time is a central topic in many scientific domains such as epidemic propagation, urban planning, and ride-sharing. However, current studies neglect individuals' preferences to explore and discover new places. Yet, neglecting novelty-seeking activities at first glance appears to be inconsequential on the ability to understand and predict individuals' trajectories. In this work, we claim and show the opposite: exploration moments strongly impact mobility understanding and anticipation. We start by proposing a new approach to identifying moments of novelty-seeking. Based on that, we construct individuals' mobility profiles using their exploration inclinations – *Scouters* (i.e., extreme explorers), *Routiners* (i.e., extreme returners), and *Regulars* (i.e., an individual with no extreme behavior). Finally, we evaluate the impacts of novelty-seeking, quality of the data, and the prediction task formulation on the theoretical and practical predictability extents. The results show the validity of our profiling and highlight the obstructive impacts of novelty-seeking activities on the predictability of human trajectories, in particular, on *Scouters*.**

***Index Terms*—Individual Mobility, Exploration (Novelty-seeking), Mobility Profiling, Predictability, Prediction**

I. Introduction

The understanding and modeling of human mobility became an accessible domain of study given the ubiquity of mobile devices, Internet connectivity, and positioning systems. The collection of large amounts of mobility data and individuals' whereabouts urged scientists from diverse disciplines to examine the dynamics of human mobility behavior. In the literature, there are several representative models proposed to reproduce individuals' trajectories and various robust predictors to forecast future locations. Indeed, accurate mobility models and predictors are crucial for epidemic prevention (e.g., the COVID-19 pandemics) [1], disaster response, and traffic management [2, 3]. Besides, such accuracies improve the services offered by pervasive computing applications [4], provide energy-efficient and cost-effective network infrastructures [5].

Previous studies [6, 7] showed that individual mobility is characterized by (i) *high temporal and spatial regular patterns* interrupted by (ii) *irregular sporadic visits to unknown or rarely visited places*. The pattern regularity is delineated by few visited locations, where users frequently return. On the other hand, irregularity and sporadic visits strongly impact predictability and are characterized as undetectable by predictors.

Licia Amichi is with École Polytechnique (IPP) and Inria, Palaiseau, France (email: licia.amichi@inria.fr)

Aline C. Viana is with Inria, Palaiseau, France (email: aline.viana@inria.fr).

Mark Crovella is with Boston University, Boston, United States (email: crovella@bu.edu).

Antonio Loureiro is with Federal University of Minas Gerais, Brazil (email: loureiro@dcc.ufmg.br).

Given the difficulties involved in anticipating location visits in mobility-related behavior, a frequent tacked question in the related literature is *to what extent is human mobility predictable?* In this regard, different predictive studies have been conducted, either to infer the theoretical upper bound (i.e., theoretical predictability) [2, 8, 9] or the prediction accuracy (practical predictability) [10–12]. Nevertheless, the empirical results suggest that the predictability takes variable values ranging from under 40% to higher than 90% [12]. Such varying results bring a new question: *what are the origins behind these significant variations in the predictability measures?* Alternatively stated, *what are the essential factors that influence the predictability?*

To answer this question, prior investigations demonstrated that the quality of the data considerably affects the predictability, namely the temporal and spatial resolutions [9, 12–14]. Indeed, human mobility is substantially more predictable when using finer-grained temporal resolution or when increasing the size of spatial units.

Another impacting factor is the prediction formulation. The literature reports a range of task formulations of the mobility prediction, namely, the next-cell, the next-place, the next-activity, or still the next-cell combined with contextual data. The most wide-spread versions are the *next-cell* and the *next-place* tasks, which formulations depend exclusively on the spatiotemporal specificity of the collected data. The other prediction formulations also require contextual information such as activity patterns, social ties, or semantic labels, making them less accessible and more challenging to analyze due to data acquisition and privacy concerns.

Withal, a non-negligible impacting factor, and focus of this paper, is the tendency of individuals to explore and discover new places. Admittedly, novelty-seeking is highly present in our daily lives, in fact, we are continuously hunting for new places and spots to go [12]. Moreover, the susceptibility to break the returning routine to explore and discover new places is heterogeneous among populations. In this vein, the literature reveals divergence in profiles according to the proclivity to explore [15, 16].

We claim that the high exploration susceptibility and related heterogeneous profiles of individuals indicate that the novelty-seeking factor is an essential element to consider and should not be overlooked, particularly for specific categories exhibiting high exploration activities. A resulting question is thus *to what degree do novelty-seeking activities obstruct the predictability of human mobility trajectories?*

In this paper, we answer this question. Toward this goal, we investigate the obstructive impacts of the novelty-seeking activities on the predictability extent of individuals' mobility

traces. In a preliminary version of this work [17], we profiled and evaluated the exploration phenomenon of individuals. Here, we build on this prior effort by presenting a much more comprehensive investigation and offerings:

- According to our investigations, we are the first to propose a novel per-user approach to distinguish between: (i) RV places visited for regular and routine activities, and (ii) EV places visited when being carried by the tendency to explore, and, thus, exploit it to identify moments of novelty-seeking. In Section IV-B, we endorse our proposal by a thorough experimental validation and a performance comparison with a state-of-the-art approach.
- In Section IV-C, based upon the two captured types of locations, we split individuals' visits into two categories: explorations and returns. Then, we define new profiling metrics that capture individuals' propensity to explore new places and their intermittent behavior– i.e., the shift between the two types of visits. Subsequently, using the newly designed metrics we reveal the existence of three visiting profiles: *Scouters*, *Regulars*, and *Routiners*. For this, we use four urban datasets, describing people's mobility from 5 cities in 3 different continents around the world (Section III).
- Finally, we are the first (to the best of our knowledge) to measure and quantify the impacts of novelty-seeking activities on the potential predictability of individuals' traces and corroborate that exploration events are one of the main origins of the decrease in predictive performance. In addition to that, we also evaluate the effects of the most reviewed impacting factors on the predictability, namely, temporal and spatial resolutions and prediction formulation (Section V).

The remainder of this paper is organized as follows. We start with an overview of the related work in the field of predictability and its impacting factors in Section II. Following, in Section III we describe the datasets used throughout the study and the experimental settings. Next, in Section IV we present our profiling methodology. Afterward, in Section V, we excerpt the factor impacting the potential predictability of the mobility traces of each profile. Finally, we provide a discussion on the future research directions and open issues and challenges in Section VI.

Summary of the main outcomes: The similar cohesive groups resulting from the diverse and heterogeneous datasets suggest the generality of our profiling approach. Additionally, with the variation of the spatial and temporal resolutions and the prediction formulation, the different profiles are still plainly distinguishable and support the stability of our clustering. Essentially, understanding the impacts of novelty-seeking on predictability and prediction extents per-profile offers the opportunity to gain control by adjusting the predictors to the profiles. Namely, the profiling method helps identify who can be trusted and who is uncertain and requires further analysis. Withal, we show that although being the hardest to predict category, *Scouters* do have a routine, and their prediction is acceptable even in metropolitan-scale analysis, i.e., in urban areas where Points of Interest usually span a few square meters (for cells = 200 m): 80% of *Scouters* have more than 80% of prediction accuracy, what indicates a prediction error only in 20% of cases.

II. Related Work

Over the last decade, human mobility was extensively scrutinized to understand the mechanisms ruling an individual's movements. Several studies demonstrated that human movements are far from being random and have a high degree of predictability [18].

Song et al. [8] proposed, in their seminal work, an approach to measure the upper bound of its *maximum predictability* $\Pi^{\max}$ based on the entropic level of a mobility trace. Analyzing a three-month-long CDR dataset of 50,000 users, their study revealed a 93% potential predictability in an individual's mobility trace. Several subsequent studies tried to refine the predictability upper bound $\Pi^{\max}$. For instance, Lu et al. [2] determined that in a CDR dataset containing the mobility trace of 2.9 million individuals, the upper limit of their predictability is estimated to be 85%.

Building upon the above findings, many advanced *predicting algorithms* were designed attempting to approach the theoretical predictability, such as Markovian predictors [10], Bayesian network models [11], neural network algorithms [19], or advanced deep learning approaches [20]. Lu et al. [10] sought to approach the theoretical limits of the predictability and utilized a Markov Chain based predictor with a varying order, and showed that the practical predictability (denoted by s in this paper) reaches 91%. Moreover, they showed that higher-order Markov Chain models do not significantly improve the practical predictability. Gao et al. [11] implemented a novel predictor based on Bayes Networks and found that, using the Nokia Mobile Data Challenge that contains the mobility traces of 80 users, the practical predictability is about 50%.

Subsequent studies employed the same approach as in [8] tried to dig out the significant factors that affect the predictability of human mobility and shed light on the origins of the limitations in predicting the next location:

Spatial and temporal resolutions: Jensen et al. [13] examined the upper bound predictability using various types of mobile sensor data, namely, GSM, WLAN, Bluetooth, and acceleration of 48 days' records for 14 individuals. Likewise, they reported high potential predictability for the data. Additionally, they showed that by varying the temporal resolution from a few minutes to a few hours, the highest predictive performance is obtained when the time scale is 4 to 5 minutes. Later, Lin et al. [9] used a high spatial and temporal resolution GPS dataset of 40 individuals. They showed that their finer-grained dataset produces higher upper bounds with predictability exceeding 98% with a temporal scale of 20 minutes or less. Likewise, Smith et al. [14] showed that the predictability is correlated with the temporal resolution and has an inverse correlation with the spatial resolution.

Type of prediction: Ikanovic et al. [21] emphasized the origins of the high potential predictability of individuals' mobility obtained in earlier studies [2, 8]. They focused on

the next-place prediction that considers moments of transitions only, i.e., moving from a place to a distinct one, then estimated the upper bound limit of the predictability, and obtained significantly lower performance of approximately 71%. Thereby, they validated that the high estimated values of predictability in previous studies stem from the stationarity captured by the prediction formulation rather than movements. Similarly, Cuttone et al. [12] analyzed the predictability of a GPS dataset with the two widespread formulations of prediction, namely, the next-cell prediction and the next-place prediction. While the next-cell prediction shows to have a very high upper bound $\Pi^{\max} = 95\%$ due to the stationarity in the human mobility, the next-place prediction appears to be more challenging with an upper bound lower than 68%.

Novelty-seeking: Recent studies showed the importance of considering individuals' tendencies to explore and discover new locations when modeling their mobility [7]. Notably, Cuttone et al. [12] highlighted the importance of considering the exploration phenomenon when designing mobility predictors. Indeed, the higher an individual is prone to discover new places, the less predictable he/she is as it is impossible to forecast the unknown. This point led to an important question, *do all individuals explore at the same rate?* Or, *is there a category of individuals who explore more and hence are less predictable?*

In this regard, Pappalardo et al. [15] discerned two categories of people: explorers and returners. They based their classification on the number of regularly visited places: explores visit many locations regularly, whereas returners limit their mobility between few places.

Besides, Scherrer et al. [16] using an unsupervised approached classified individuals into travelers and locals. Travelers have a spread mobility, whereas locals move in a more constrained area and revisit many of their locations.

We claim literature studies, although focusing on a very important mobility behavior, do not provide a precise understanding of individuals' exploration tendency. Therefore, in our previous work [17], we proposed a mobility profiling based on individuals' tendency to explore that we further improve in this paper. We revealed the existence of three main categories of individuals: (1) *Scouters or extreme explorers:* whose proclivity for novelty-seeking is the most eminent all over the week and have a more spread spatial mobility; (2) *Routiners or extreme returners:* who rarely perform explorations and have confined mobility; (3) *Regulars:* who have a medium behavior.

Accordingly, exploratory activities are not consistent among the population. While some groups depict a high propensity for discovering new areas and spots, others spend their time between familiar places. *Investigating how novelty-seeking inclinations of individuals affect the predictability of their mobility traces is a topic that has yet to be researched.*

Position of our work: While the impacts of prediction formulation and the quality of the data on predictability extents have been widely investigated, the limiting factors that arise from the intrinsic nature of human mobility have rarely been addressed. In this paper, on the one hand, we shed light on one of the main limiting factors of predictability, namely, individuals' propensity to explore, and for the first time in literature, we present a newly-tailored method to recognize moments of novelty-seeking, and by the mean of this, we deeply improve our previously proposed mobility profiling [17]. On the other hand, we study predictability extents, which is the main focus of this paper, and evaluates how each of the prediction formulation, the quality of the data, and the proclivity for novelty-seeking influences the predictability.

III. Data Description

In this work, we use two categories of data sources; three Global Positioning System (GPS) and one of Call Detail Records (CDR). These datasets capture spatio-temporal footprints of individuals' mobility with high spatial and temporal resolutions. We outline our datasets in Table I and discuss them hereinafter.

A. GPS datasets

GPS technology allows tracking individuals' movements with the highest level of accuracy and temporal frequency. We leverage three GPS data sources:

Macaco: it consists of anonymized digital activities tracks of 132 volunteers from 6 different countries collected in the context of the MACACO project [22]. For project-related privacy policies, this dataset is not publicly available. It provides a long-term and fine-grained sampling of individual behavior and network usage with a frequency of one sample every 5 minutes for a duration of 34 months. The data source contains about 900k tuples with raw GPS coordinates (latitude and longitude) and timestamps. Each tuple has a unique ID, which relates to a specific user.

Privamov: it contains mobility traces collected in the Privamov sensing campaign [23], capturing the spatio-temporal footprints of 100 unique volunteers over 15 months around a city in Europe. The data source was gathered over 156 million GPS records with a frequency of sampling roughly equal to a few seconds.

Geolife: the last GPS data source is the Geolife public dataset collected by Microsoft Research Asia [24–26]. The dataset stores information about the GPS trajectories of 182 individuals distributed in over 30 cities mainly in China, the USA, and Europe. The dataset includes time-stamped GPS tuples recorded every 1 to 5 seconds for more than 64 months.

B. CDR dataset

Mobile phone records consist of time-stamped and geo-referenced records of voice phone calls and SMS of mobile network subscribers, called Call Detail Records (CDR). Each record usually contains the hashed identifiers of the caller, the timestamp for the call time, and the location of the cell tower to which the caller's device is connected to when the phone activity is originated.

ChineseDB: this dataset is collected from 642K anonymized mobile phone subscribers in Shanghai, China [1], and contains

[1]The collection was initiated by Shanghai University [27].

Dataset	Category	Number of users	Duration	Frequency of sampling
Macaco [22]	GPS	132	34 months	5 min
Privamov [23]	GPS	100	15 months	few seconds
Geolife [24–26]	GPS	182	64 months	1 to 5 seconds
ChineseDB*	CDR	642K	2 weeks	1 hour

*The collection was initiated by Shanghai University [27].

TABLE I: Datasets description.

400k calls. It provides aggregated human footprints in the frequency of one location per hour during a period of 2 weeks. The locations in this dataset are gathered by merging the locations of the original CDR in each one-hour interval. Each location of an hour represents the user's centroid of the hour with the precision of 200 meters according to the instruction of the data provider. This accuracy of positioning is higher than that of the original CDR.

C. Data handling

Modeling and predicting individuals' mobility focus on the location data i.e. latitude and longitude. First, we *reconstruct the mobility trajectory* $\mathcal{H}_u$ of each individual u by extracting the sequence of recorded locations along with the associated timestamps at fixed time periods δ, $\mathcal{H}_u = \langle (lon_0, lat_0, t_0), (lon_1, lat_1, t_1), \dots (lon_N, lat_N, t_N) \rangle$, with $t_i = t_0 + i \times \delta$.

Next, we *discretize the geographical maps* by placing uniform grids of c meters $\times$ c meters and draw out the grid cell IDs associated with the coordinates, by converting the tuple (lat_i, lon_i) into a cell identifier $(id_i = \lfloor \frac{lon_i}{c} \rfloor, \lfloor \frac{lat_i}{c} \rfloor)$ as in [12], where c meters is the cell-size in the grid. Hence, the mobility trajectory of the individual u is converted into sequences of timestamped discrete symbols –a discrete mobility trajectory–, $\mathcal{T}_{u,c} = \langle (id_0, t_0), (id_1, t_1), \dots (id_N, t_N) \rangle$.

Afterward, given that the location of each individual is obtained at different uniform temporal rates in our GPS data sources – i.e., 5 min for the Macaco, few seconds for Privamov, and 5 seconds for Geolife –, we re-sampled all the GPS datasets to have an *equal frequency of one sample every 5 min*, i.e, $\delta = 5min$. However, some records can be missing due to delayed measurements produced by the sleeping phases of mobile devices collecting the data. Hence, to have a more *uniform and complete traces*, we comply with some steps proposed by Chen et al. [27] and complete them as follows [2],

- First, per individual u, we identify the most frequent daily location id_{wp_A} between 10 am and 11 am and name it *workplace A*.
- Second, we locate the most visited location id_{wp_B} between 2 pm and 5 pm and name it *workplace B*.
- Next, we determine the most prevalent place id_H between 2 am and 6 am (night), which we refer to as *home*.
- Once *home* (id_H), *workplace A* ($id_{wp_A A}$), and *workplace B* (id_{wp_B}) locations are identified,
  - if a record is missing at t_x between 10 am and 11 am we complete the mobility trajectory $\mathcal{T}_{u,c}$ with a new record (id_{wp_A}, t_x),
  - if a record is missing at $t_x \in [2\text{ pm}, 5\text{ pm}]$, we add the tuple (id_{wp_B}, t_x) to the mobility trajectory $\mathcal{T}_{u,c}$,
  - if a record is missing at $t_x \in [2\text{ am}, 6\text{ am}]$, we add to the mobility trajectory $\mathcal{T}_{u,c}$ the record (id_H, t_x).

[2]Note that if an individual does not have data allowing to detect her workplaces or home location is viewed as bad a user and is therefore filtered.

D. Experimental settings

In what follows, we give a brief description of the parameter settings we used in this study. Unlike in our previous works [17], we define a *complete day* for the GPS datasets as a day in which an individual has *on average* one record each 15 min. And select only participants that have at least 1 month of complete days of data. We are left with 266 users: 84 in Macaco, 77 in Privamov, and 105 in Geolife. For the CDR data, given the low frequency of sampling, we define *a complete* day as a day having *on average* one record every 2 hours and select only participants that have at least 14 days of complete data, we are left with 4860 individuals.

We discretize locations *to grid cells of size* $c = 200\,m$, with *a frequency of 1 record each 5 min* for the GPS datasets, and *1 record per hour* for the CDR dataset. There are two reasons to consider these spatial and temporal resolutions. First, we focus on the discoveries of new places on a daily basis, for instance, going to a new restaurant or a new shop. Considering the imprecision and uncertainty of GPS systems, we claim a cells of size 200 m $\times$ 200 m roughly corresponds to daily regions of interest and can still captures discovery moments. Second, the higher is the temporal resolution the better is the understanding of human movements. Nevertheless, there is a tradeoff between expanding the set of selected individuals and increasing the temporal resolution. Although corresponding to the highest sampling interval among the presented GPS datasets, a resolution of 5 min allows uniforming the frequency of sampling between the different sources while increasing the number of individuals and being reasonable for capturing most movements. Hence, having different datasets with the same resolutions allows us to test our methods' effectiveness and validate our work extensively.

GPS data aggregation: due to the small number of individuals in GPS data sources and considering that these sources are of the same nature (i.e., with the same frequency of sampling and duration of analyses (1 month)), we proceed as the following. We aggregate the filtered and manipulated GPS datasets and label this new dataset as *Agg_gps*. Starting from Section IV-E, we do not use the GPS datasets individually but employ the aggregated dataset *Agg_gps* to perform global characterizations and comparisons. In view of its different nature, the CDR dataset will be analyzed separately.

IV. Profiling methodology

Human beings' movements are a mixture of *repetitive and regular* visits between known places and *sporadic discoveries* of new areas [6, 15], both subject to a certain degree of uncertainty associated with free will and arbitrariness [6]. At each instant, an individual is confronted with an extensive list of choices concerning *where* and consequently, *how* to spend her time: she either returns to a place she visited in the past or explores a new location.

Contrary to the extensive literature investigations on mobility regularity patterns, we focus on the discoveries of new places. In particular, *we intend to investigate whether there exist patterns when commuting from an exploration mode to a return mode and vice versa*. For this, as in [7], we divide human movements into two primary states: *explorations and returns*. We define (1) the **exploration** as *a discovery of a new location* and (2) a **return** as *a visit to a previously seen locality*. Note that a central point in the exploration identification is to settle when a novelty-seeking moment happens. Hereafter, we describe our proposed strategy for this identification as well as our profiling methodology.

A. Formalization

Let M be the Finite-State Automaton (FSA) describing an individual's movements, as shown in Fig. 1, with two possible states: *exploring* (**E**) and *returning* (**R**). An individual u can either be in the exploring state (**E**) or the returning state (**R**). Two possible transitions can affect an individual's state: *return* (T_r or S_r) by going back to historically known locations, and *explore* by discovering new spots (T_e or S_e). In the exploring state **E**, discovering new areas (S_e) has no effect and keeps the individual in the state **E**. On the other hand, moving back to a known location (T_r), though recently explored, M shifts the state from **E** to **R**. In the returning **R** state visits to usual places (S_R) does not change the state, however, a discovery of a new spot (T_e), shifts the state back to the **E** state.

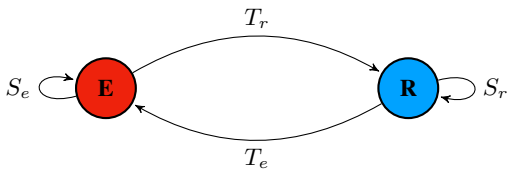


Fig. 1: Finite-State Automaton M.

B. Novelty-seeking identification

Strictly speaking, for an individual an exploration is the discovery of a new geographical location, i.e., a place where she was never seen before. Nonetheless, existing works tackling the exploration problem consider the first occurrence of a location in the mobility trace as an exploration [7, 12], which leads to an overestimation of exploration events. This means that the first appearance of the *home* location or the *workplace* in the sequence is viewed as a moment of novelty-seeking. Yet, overvaluing the frequency of exploration events might twist the understanding of the exploration problem. Hence, given the mobility trace of an individual u, *how can we distinguish her novelty-seeking visits from her routine visits?*

We propose a newly tailored per-user approach to distinguish between locations used for exploration visits and familiar regularly visited locations, the approach is described in Algorithm 2. Besides, to verify and validate our approach we conduct a performance comparison with the state-of-the-art location classification algorithm proposed by Papandrea et al. [28] and described in Algorithm 1:

1) Baseline identification: we use the widespread framework proposed by Papandrea et al [28] as a baseline strategy. Grounds for their seminal per-user scheme that allows the evaluation of the importance of a place in a user's daily mobility, meeting our case of study –i.e., individual mobility– . Moreover, it allows the classification of the locations according to their relevance from a single user viewpoint.

For each user u, we compute the Relevance $R_u(id_i)$ of each of her visited locations id_i (cf. Algorithm 1, lines 4–5),

$$R_u(id_i) = \frac{d_{visit}(id_i, u)}{d_{total}(u)}, \tag{1}$$

where $d_{visit}(id_i, u)$ is the number of days the individual u visited the location id_i, and $d_{total}(u)$ is the number of days the individual has been active.

Following, as in [28] we use the k-mean unsupervised approach with 3 components to classify the locations into: (1) *Mostly Visited Places* (MVP), i.e, locations most frequently visited by the user; (2) *Occasionally Visited Places* (OVP), i.e, locations of interest for the user, but visited just occasionally; (3) *Exceptionally Visited Places* (EVP), i.e, rarely visited locations (cf. Algorithm 1, line 7).

Algorithm 1 Baseline identification

1: **function** *location_classification_b* ($\mathcal{T}_{u,c}$)
2: $T_{Relevance,u}, T_{MVP_u}, T_{OVP_u}, T_{EVP_u} \leftarrow \emptyset$
3: $F_u \leftarrow$ UNIQUE($\mathcal{T}_{u,c}$)
4: **for** j **in** F_u **do**
5: $\quad T_{Relevance,u}[j] \leftarrow COMPUTE\_RELEVANCE(j)$ ▷ (1)
6: **end for**
7: $T_{MVP_u}, T_{OVP_u}, T_{EVP_u} \leftarrow k$-means($T_{Relevance_u}, 3$)
8: **return** $T_{MVP_u}, T_{OVP_u}, T_{EVP_u}$
9: **end function**

2) Visitation-frequency-based identification: likewise, we propose a per-user method for the classification of the locations. Yet, unlike the baseline approach, we evaluate the importance of a location for a user u according to the number of times she was seen in that location, i.e., the frequency of appearance of the location in her mobility trace.

Let $F_u = \{id_1, id_2, \ldots, id_n\}$ be the set of location visited by the user u and consider Algorithm 2, which details the steps of this method. First, for each location $id_i \in F_u$, we assign a weight w outlining the visiting importance of id_i among the whole set of trajectory's visited locations (cf. Algorithm 2, lines 4–5). It is given by,

$$w_u(id_i) = \frac{freq_u(id_i, \mathcal{T}_{u,c})}{\sum_{j=1}^{|F|} freq_u(id_j, \mathcal{T}_{u,c})}, \tag{2}$$

where $freq_u(id_i, \mathcal{T}_{u,c})$ is the number of occurrences of the location id_i in the discrete mobility trajectory $\mathcal{T}_{u,c}$ of the user u. Next, we compute the average value of the visitation frequency $\overline{w}_u = \frac{1}{|F|} \times \sum_{i=1}^{|F|} w_u(id_i)$, per-user u (cf. line 7). Following, we categorize the visited locations into locations used for: (1) Exploratory Visits (EV), (2) Return Visits (RV). Each location id_i that has a weight $w_u(id_i) \geq \overline{w}_u \times level$ is added to the set of locations used for RV, T_{RV} (cf. lines 9–10), otherwise it is assigned to the list of places used for EV, T_{EV} (cf. Algorithm 2, 11–12).

Algorithm 2 Visitation-frequency-based identification

1: **function** *location_classification_a* ($\mathcal{T}_{u,c}$, $level$)
2: $w_u, T_{RV_u}, T_{EV_u} \leftarrow \emptyset$
3: $F_u \leftarrow$ UNIQUE($\mathcal{T}_{u,c}$)
4: **for** j **in** F_u **do**
5: $\quad w_u[j] \leftarrow$ FREQUENCY_OF_APPEARANCE($j, \mathcal{T}_{u,c}$), (2)
6: **end for**
7: $\overline{w}_u \leftarrow$ MEAN(w_u)
8: **for** j **in** F_u **do**
9: $\quad$ **if** $w_u[j] \geq \overline{w}_u \times level$ **then**
10: $\quad\quad T_{RV_u}$.ADD(j)
11: $\quad$ **else**
12: $\quad\quad T_{EV_u}$.ADD(j)
13: $\quad$ **end if**
14: **end for**
15: **return** T_{RV_u}, T_{EV_u}
16: **end function**

The parameter $level$ is critical and should be carefully tuned to allow a thorough capture of moments of novelty-seeking. Indeed, high values for $level$ can induce an overestimation of explorations, while small values lead to a neglect of novelty-seeking moments. To quantify its impact, in Section IV-B3, we evaluate the $level$ parameter under two distinct values: $level = 80\%$, corresponding to a less conservative identification (i.e., more explorations), and $level = 20\%$, corresponding to a more conservative identification (i.e., more returns).

3) Level impact and baseline comparison: first, using the baseline identification approach (Algorithms 1), we categorize the visited places into EVP, OVP, and MVP. Next, we classify the visited locations into EV or RV using the proposed Algorithm 2 with $level \in \{20, 80\}\%$. Finally, we measure the fraction of places within each category of places and evaluate their average visitation frequency, as shown in Figure 2.

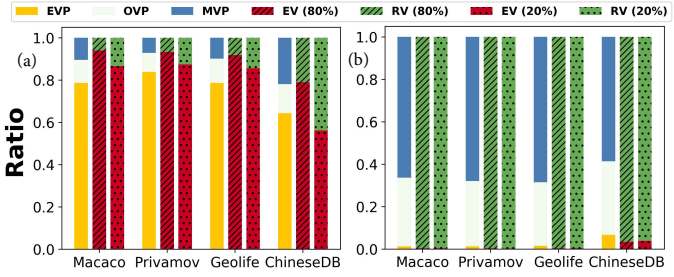


Fig. 2: (a-left) Percentage of visited places. (b-right) Average visitation frequency. EVP, OVP, and MVP are categorized according to Algo1. EV and RV are categorized according to Algo 2 for $level = 80\%$ and $level = 20\%$.

Figure 2 (a) reports the percentages of places classified within each category extracted from our datasets; EVP, OVP, and MVP by Algorithms 1, EV and RV by Algorithm 2. First, we observe the high ratio of EVP jointly with OVP categorized by Algorithms 1, for all GPS datasets, more than 78% of the places, –i.e., EVP $\cup$ OVP– are not integrated into the daily routines of the individuals. Note that the CDR dataset describes visits in a smaller temporal resolution (i.e., per hour), this naturally impacts the precision in exploration inference of visits. Likewise, in all datasets, the proportion of locations used for EV surpasses 78% when $level$ is set to 80%, and is higher than 50% with $level = 20\%$. Moreover, we can notice in the case where $level = 80\%$, the proportion of places classified as EV by Algorithm 2 corresponds roughly to the percentage of places categorized as EVP $\cup$ OVP by Algorithm 1. In contrast, in Algorithm 2 with $level = 20\%$ the fraction of places labeled EV is almost equal to the faction of locations classified as EVP by the baseline Algorithm 1.

Figure 2 (b) illustrates the proportion of the average frequency of visits towards each category of places. Firstly, we see the markedly high proportion of visits to locations used for RV, more than 90% of the visits are towards this category of places for $level \in \{20, 80\}\%$. Whereas the same score is obtained by Algorithms 1 when taking MVP and OVP together. Additionally, the average frequency of visits held by EV for all datasets with $level \in \{20, 80\}\%$ is lower than the scores obtained by EVP. Indeed, in the case of the baseline approach, the importance of a location is based on the number of days it was visited and not the amount of time she spent within it. This means, for an individual u, if she weekly visits the municipal library for 4 hours, this latter will have the same relevance score as the bakery where she goes once a week for a few minutes to only buy a baguette.

In addition to the rate of places categorized in each group, we measure the percentage of intersection between EV places and EVP, then between EV and EVP $\cup$ OVP.

	EV (80%) $\in$ EVP	EV (20%) $\in$ EVP	EV (80%) $\in$ EVP $\cup$ OVP	EV (20%) $\in$ EVP $\cup$ OVP
Macaco	60.1%	47.71%	78.33	68.38%
Privamov	50.75%	36.58%	76.92	65.38%
Geolife	41.19%	33.76%	67.82	59.18%
ChineseDB	88.78%	61.94%	98.27	84.23%

TABLE II: Percentage of EV places present in T_{EVP} and in $T_{EVP} \cup T_{OVP}$, with $level \in \{20, 80\}$.

In Table II, we report the percentage of overlap between the locations categorized as EV with $level \in \{20, 80\}\%$ at first with EVP locations only then with EVP $\cup$ OVP. Although the fraction of places categorized as EV with $level = 20\%$ is closer to the fraction of place categorized as EVP compared to when $level$ is set to 80%, the percentage of overlap between EV and EVP is higher when $level$ equals 80%. We can also observe that when measuring the degree of overlap of EV with EVP $\cup$ OVP the obtained scores increase for both $level = 20\%$ and $level = 80\%$, with a very high degree of overlap for CDR ChineseDB reaching 98.27% for $level = 80\%$. Succinctly, though the difference of our methodology in quantifying the importance of a location in the daily life

of an individual, we notice the significant overlap between the classifications of our proposed method and the baseline approach ones'.

Thereby, from one side setting $level$ to 80% allows EV capturing exceptionally and occasionally visited places as the baseline approach. From the other side, it allows capturing more precisely, the visits related to the individuals' proclivity to explore (i.e., locations that are rarely frequented). We claim thus a novelty-seeking identification method should consider both quantity and visitation frequency aspects of per-category locations.

In summary, *the proposed method, Algorithm 2 offers a satisfactory classification of the visited places*. First, it allows the detection of a higher number of places used for exploration visits (EV), on the other hand, it guarantees that the visitation frequencies to these locations are lower compared to the RV as well as EVP of Algorithms 1. Second, the performance of Algorithm 2 with $level = 80\%$ allows the identification of a higher number of places used for EV, and hence enables a more precise detection of moments of exploration compared to the setting with $level = 20\%$. Indeed, the first occurrence of a location in the set of a user's EV locations is presumed to be a moment of exploration. In the remaining of the paper, we use Algorithm 2 and set $level$ to 80%, which allows a more precise way to distinguish locations used for exploration.

C. Profiling rules

Initially, each user u has an empty set of known locations $\mathcal{L}_u(t_0) = \emptyset$. Using Algorithm 2 with $level = 80\%$ for each user u, we classify her visited locations into EV and RV. Subsequent, all locations classified as RV are added to the set of known locations $\mathcal{L}_u \leftarrow T_{RV_u}$. Therefore, each occurrence of a location present in the set of known locations $\mathcal{L}_u$ is a return, else it is an exploration. Note that after the discovery of a new place, this latter is added to $\mathcal{L}_u$, i.e., its next occurrence in the mobility trace will be viewed as a return.

After dissecting human visits into explorations and returns, for each user u we first extract two sets:

- **Returning set** ret_u**:** is a set containing the sets of consecutive returns, $ret_u = \{r_0, r_1, \ldots, r_n\}$, where each $r_i = \{id_0, id_1, \ldots, id_x\}$ is a set containing the ids of the cells where the user u performed successive returns.
- **Exploring set** exp_u**:** is a set containing the sets of consecutive explorations, $exp_u = \{e_0, e_1, \ldots, e_n\}$, where each $e_i = \{id_0, id_1, \ldots, id_x\}$ contains the ids of the cells where the user u performed successive explorations.

Next, we assign to each individual u two values: (1) $\#E = avg(|e_i|), e_i \in exp_u$, the average number of her successive explorations – the average number of consecutive self-transitions she made in the E state, and (2) $\#R = avg(|r_i|), r_i \in ret_u$ the average number of successive returns – the self-transitions she made in the R state.

To characterize how individuals balance the trade-off between revisits of familiar locations and new-places discoveries, we define the following metrics that utterly capture the exploration habits of an individual. The first metric captures the shifting habits between the exploration and the return modes. The second metric captures the susceptibility of users to remain in their routine rather than explore new places.

Definition 1 (*Intermittency* μ). *is the sum of the average number of successive explorations $\#E$ and the average number of successive returns $\#R$, $\mu = \#R + \#E$.*

The *intermittency* measure reveals whether an individual is versatile or prefers to remain steady with respect to a category of location (i.e., return or exploration). Namely, it helps to recognize if a user is constantly fluctuating between visits to familiar places and discoveries of new spots or once she starts a discovery, she does is it repeatedly, before switching to revisits and vice versa.

Definition 2 (*Degree of return* α). *is the angle whose tangent is the ratio between the average number of successive returns **R** over the average number of successive explorations **E**,$\alpha = \operatorname{arctg}\left(\frac{\#R}{\#E}\right)$.*

The *degree of return* describes the exploration conducts of an individual compared to her returns. Having a high degree of returns suggests that: the average number of successive returns is higher than the average number of successive explorations $\#R > \#E$. Hence, the *degree of return* reveals what kind of explorer an individual is: whether she visits many new places on a row, or just after a few discoveries she goes back to a familiar location.

In what follows, we investigate whether the novelty-seeking habit is the same among the population or if it is a distinctive property. Namely, if there exist patterns followed by individuals while shifting between the exploration mode and returning mode or if there are several groups of users sharing the same habits but distinct from the others.

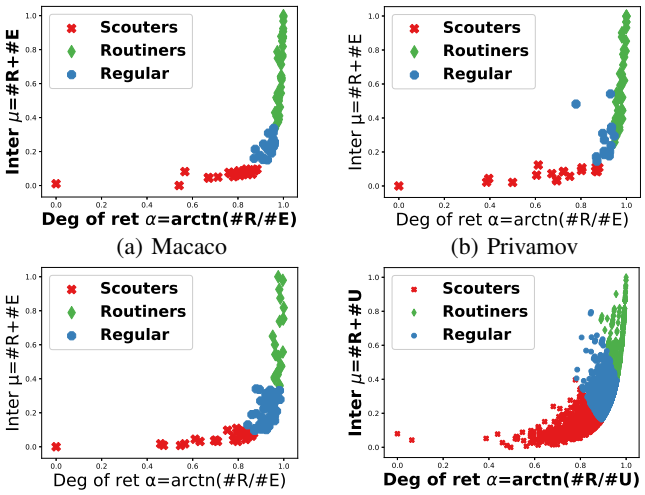
D. Mobility Profiling

After computing the intermittency μ and degree of return α for each individual, we use two clustering algorithms – the Gaussian Mixture probabilistic Model (GMM) and the k-means clustering method – to attest whether we can split the population into distinct cohesive and significant groups or not. To identify the best number of components of the clustering algorithms, and hence, the individuals' types, we use the silhouette score statistical test and the Davies-Bouldin Index as well as we run one hundred fits for five different sets of clusters (two to six). Then, we consider the mean value when choosing the best score. The results show that the best performance is obtained with a clustering with three components (see Appendix **??**).

We then apply the GMM and k-mean with three components on our data sources. We roughly obtain the same groups for both clustering algorithms. Thus, we only present the results obtained with the GMM algorithm. Fig. 3 depicts the normalized intermittency of individuals against their normalized degree of return and displays the clusters resulting from the application of the GMM algorithm on the GPS and CDR datasets. We can observe that our metrics can clearly capture the dissimilarity between the individuals in terms of human mobility dynamics. More importantly, the GMM

identifies three distinct groups that have identical *intermittency* and *degree of return* characteristics for all our data sources. We label the resulting groups as **Scouters** (red), **Routiners** (green), and **Regulars** (blue).

- Cluster 1: *Scouters or extreme explorers*, although holding varying degrees of return α, they are remarkably lower compared to others' scores. Moreover, they are notably intermittent –i.e., they are constantly shifting between the exploring and the returning states. These users are more prone to explore and discover new areas.
- Cluster 2: *Routiners or extreme-returners* have a surprisingly large degree of return. Besides, they tend to be steady in the different states of the automaton M –i.e., they rarely break their routine. Hence, we can deduce that these users rarely explore and prefer to stick among their common and known places.
- Cluster 3: *Regulars* adopt a medium behavior and have large degrees of return compared to the *Scouters*. Though, their intermittencies are distinctly smaller than those of *Routiners*. These users constantly alternate between explorations and revisits. Yet, their proclivity to explore is less important than *Scouters*'.



(a) Macaco (b) Privamov (c) Geolife (d) *ChineseDB* CDR.

Fig. 3: Mobility Profiling.

The proposed approach captures two major mobility features that fully describe the exploration phenomenon, i.e., *intermittency between returns and explorations, and the ratio of explorations compared to returners*, and allows a natural clustering of the individuals.

E. Profiles' Mobility Traits

We now analyze the mobility behavior of individuals of each profile according to three dimensions: *Relocation Activities*, *Temporal Activities*, and *Spatial Activities*. To perform a global characterization we use the aggregated GPS dataset *Agg_gps* along with the ChineseDB data source.

The *Relocation Activities* features aim at quantifying and characterizing individuals' visits, transition habits, and repetitiveness of visits. It involves four metrics:

- ***Number of successive explorations***: it measures the average number of successive explorations performed by an individual.
- ***Number of successive returns***: it estimates the average number of successive returns of an individual.
- ***Number of stops***: it is the total number of distinct areas visited by an individual.
- ***Visitation frequency***: it measures the frequency of visits to each area known by the individual.

The *Temporal Activities* features relate to the behavior of the individuals in time and captures the amount of time spent when discovering new places and when revisiting known locations. It comprises two measures:

- ***Duration of successive explorations***: it is the average duration spent by the individual exploring.
- ***Duration of successive returns***: it is the average duration spent by the individual revisiting known locations.

The *Spatial Activities* gives an intuition on the distances walked by the individuals when performing each type of visit. It consists in three features:

- ***Total exploring distance***: it measures the total distance walked by the individual when exploring new places.
- ***Total returning distance***: it is the total distance walked by the individual when returning to known places.
- ***Radius of gyration*** r_g: it estimates the total radius of gyration of the individual given by, $r_g = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(r_i - r_0)^2}$ [15], where r_0 is the center of mass of the individual, N is her set of location history, and r_i is a two-dimensional vector containing the geographical coordinates of the location i.

In Table III we provide a finer view of the mobility traits of each profile and sustain our profiling method by highlighting the subsisting dissimilarity between the profiles. Due to the lack of space, we only give a short description of each profile. A more detailed characterization of the profiles and statistical results are provided in our previous works [17].

In summary, *Scouters* exhibit a *high exploration activity* and are characterized by markedly large sets of known places, which make them *less regular* in terms of spatial and temporal visits.

V. Revealing Novelty-Seeking Impacts

In this Section, we aim to evaluate explorations' effects (the first literature evaluation to the best of our knowledge) and the quality of the data on the two most widespread next location prediction *next-cell* and *next-place* tasks individually. We start by presenting the evaluation procedure followed in each prediction formulation. Following, we evaluate the performance reached in the next-cell formulation, and present the impacts of each of (1) the quality of the data and (2) and the proclivity to explore. Finally, we examine the attainable accuracy of prediction in the next-place prediction task, and show the effects of the impacting factor in this case i.e., novelty-seeking.

Feature name	Scouters	Regulars	Routiners
Number of successive explorations vs number of successive returns	• Relish discovering many new other places uninterruptedly. • Keen to break their returning routine.	• Constantly shifting between the exploring and the returning states.	• Rarely leave their zone of comfort or interrupt their successive returns. • Once they explore, they either stay at the same place or go to a familiar place.
Number of stops	• Visit a remarkably large number of distinct places.	• Have large sets of known places compared to *Routiners* but smaller than the *Scouters*'.	• Diversify less their visits. • Enjoy routine visits and transitions between familiar locations.
Visitation frequency	• Do not revisit the same places several times, except for some specific ones (their routine patterns consist of a small set of areas).	• Do not equally visit known locations and restrict their returns to a small set of places.	• Frequently revisit know places.
Duration of successive explorations vs duration of successive returns	• Not only relish discovering many places successively but also do it for long periods. • Spend short amounts of time returning.	• Spend a larger amount of time exploring compared to the *Routiners* and a larger amount of time returning than *Scouters*.	• After performing an exploration they spend large amounts of time returning before aspiring to discover a new spot.
Total exploring / returning distance	• Walk long distances in general, i.e., when exploring or returning.	• Walk larger distances when exploring compared to *Routiners*.	• Do not walk long distances in general.
Radius of gyration r_g	• Have larger radius of gyrations.	• Hold medium values.	• Characterized by a small radius of gyration.

TABLE III: Mobility traits of the profiles.

A. Evaluation methodology

Hereafter, we describe the evaluation methodology to measure the impacts of each of the quality of data and individuals' tendency to explore on the two widespread prediction tasks. Note that the frequency of sampling for the Agg_gps is set to 15 min, i.e., $\delta_{Agg\_gps} = 15min$, and 1 h for the ChineseDB, i.e., $\delta_{ChineseDB} = 1h$, besides, we use a squared tessellation with cells of size 200 m $\times$ 200 m, i.e., $c = 200$ m.

1) Prediction tasks: there exist several ways to define the mobility prediction task depending on the quality of the available data and the objectives of the forecast. In this paper, we utilize the two most common prediction task formulations relying on location data only:

- **Next-cell:** given the mobility trace of an individual and considering a time window Δt, the next-cell prediction attempts to answer the subsequent question, *where will the individual be at time $t + \Delta t$?* The triggering element in this formulation is the *time*, after each period Δt the system tries to forecast the future location of the individual. This type of prediction can result in the current location as a future location for an individual, alternatively stated, the stationary nature of human trajectories is contained [12, 14].
- **Next-place:** this formulation is independent of the temporal dimension, it seeks to answer the following question, *where will the individual go next?* The next-place prediction aims at forecasting transitions between places. Hence, the triggering element is the transition of the user from her current location [12, 14].

2) Theoretical predictability: for each prediction formulation, we start by measuring the theoretical predictability of the mobility behavior of each of the *Scouters*, *Regulars*, and *Routiners*. This will provide insights about the capacity of correctly forecasting the mobility trajectories with an ideal and utter predictor. In this regard, we employ the state-of-the-art entropic-based approach proposed by Song et al. [8] to estimate the upper bound of the theoretical predictability Π^{max}.

For each user u of each profile, given her discrete mobility trajectory $\mathcal{T}_{u,c}$, we consider the stochastic sequence $x_1^N = \{x_1, \ldots, x_N\}$ where x_t is the cell id of her location at time t. Then, we estimate the upper bound of the theoretical predictability Π^{max} of the x_1^N sequence as in [8].

3) Practical predictability: afterwards, we estimate the practical predictability of each of the *Scouters*, *Regulars*,

and *Routiners*. We compare the predictive performance of four predictors, namely, Markov Chain (MC) [29], Predicting by Partial Matching (PPM) [30], Sampled Pattern Matching (SPM) [31] , and Active LeZi (ALZ) [32].

For the predictive performance comparison between the predictors, we measure the accuracy of the prediction achieved by each predictor. Given a stochastic sequence $x_1^N = \{x_1, \ldots, x_N\}$ of N observations capturing the trajectory of an individual u. For each predictor and each user u, we initialize (i.e. "warm-up") the considered predictor using the $N_s = \frac{2}{3} \times N$ first elements $x_1^{N_s}$ (i.e, 20 days for the Agg_gps and 10 days for the ChineseDB). Second, we use the predictor to forecast the next location x_{N_s+1}. After this forecast, we update the predictor by considering $N_s \leftarrow N_s+1$ first elements of the stochastic sequence x_1^N. We then repeat the second step while $N_s \neq N$. Finally, when $N_s = N$, we stop the iterations and compute the success rate score s_u for right predictions (accuracy of prediction) given by,

$$s_u = \frac{1}{N - N_s} \sum_{t=N_s+1}^{N} \mathbb{1}(x_t = x_t^* | x_1^{t-1}), \tag{3}$$

where x_t is the actual location and x_t^* is the predicted value.

Experimental settings: for the MC(k) and PPM(k) predictors, we choose a $k \in [\![1, 2]\!]$. A k-th order MC predictor bases its forecast solely on the k previous observations. Whereas, a k-th order PPM model employs a combination of MC(j) models with $j \in [\![0, k]\!]$ [30]. For the SPM(α), we choose $\alpha \in \{0.1, 0.9\}$. α represents the fraction of the maximal suffix employed to predict the future location. Note that the *maximal suffix* is the immediately longest foregoing set of locations whose copy appeared in the previous location history.

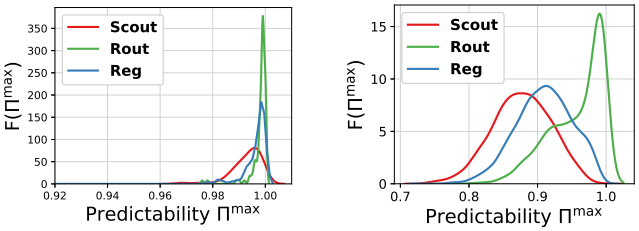
4) Impacting factors: finally, we evaluate the impacts of each of (1) the quality of the data and (2) individuals' tendency to explore when relevant on the predictive performance achieved by each prediction task.

Temporal variation procedure: in the case of next-cell prediction, we investigate the effects of varying spatial and temporal resolutions on the accuracy of prediction s for each mobility profile. Provided that the next-cell prediction task is independent of the temporal resolution, we do not investigate the impacts of the quality of the data factor on this formulation.

Exploration-isolation procedure: we identify moments of exploration using Algorithm 2 and remove them from the mobility trajectories or replace them and observe how they affect the predictors' performances. These manipulations are performed for both prediction tasks but in different ways for the replacement procedures.

B. Next-cell

We first tackle the next-cell prediction task. We measure and analyze the theoretical and practical predictability of the mobility traces of individuals of each profile. Next, we investigate the effects of varying the spatial and the temporal resolutions on the accuracy of prediction. Finally, we identify moments of exploration and remove/replace them from/in the mobility trajectories, to probe the impacts of novelty-seeking on the predictive performance.



(a) Agg_gps (b) ChineseDB

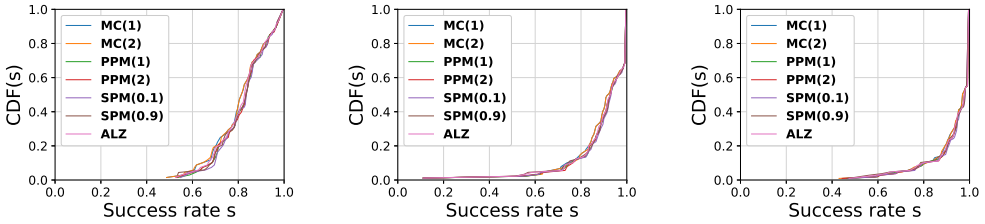
Fig. 4: Distributions of the upper bound of the theoretical predictability Π^{max} for individuals of each mobility profile.

1) Theoretical predictability: Figure 4 portrays the distribution of the upper-bound predictability for each mobility profile for both the Agg_gps and the ChineseDB datasets. We can observe the high inherent predictability of the mobility trajectories of individuals of all profiles. Particularly, individuals of the Agg_gps have a more eminent degree of potential predictability principally due to the high frequency of sampling of the dataset $\delta_{Agg\_gps} = 15min$, while $\delta_{ChineseDB}$ is set to 1 h. Admittedly, a higher frequency of sampling allows a more thorough capture of the stationarity and consequently, increases the degree of predictability [12]. More importantly, from Figure 4b, we note that the predictability Π^{max} picks around 0.97 for the *Routiners*, 0.91 for the *Regulars*, and 0.87 for the *Scouters*. Taken together, these results indicate that *Routiners* are characterized by a very high degree of predictability while the *Scouters* are the least predictable individuals. Still, although presenting the lower predictability among the three mobility profiles, the *Scouters* predictability is surprisingly high, mainly if considering the intuitive impossibility of predicting the uncertainties in *Scouters* mobility. Indeed, as reported in Table III, *Scouters* do have routines that consist in small sets of locations that they frequently visit.

2) Practical predictability: the estimations of the predictability upper bound of individuals' trajectories reveals the high potential of predictability for all the profiles, with a lower score for *Scouters* (i.e., at most 0.87 in the ChineseDB dataset). Nevertheless, the prediction accuracy does not always reach the score provided by the theoretical measure [10] (see Section II). Hereafter, we evaluate the accuracy of prediction achieved by each of: MC, PPM, SPM, and ALZ.

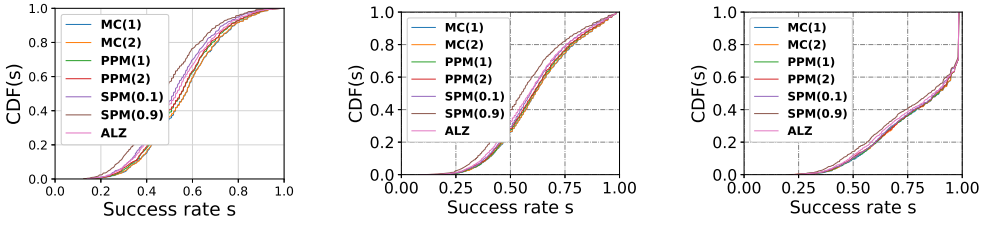
In Figures 5 and 6 we plot the CDF of success score s_u of the MC, PPM, SPM, and ALZ predictors with respect to their possible parameters $k \in \{1, 2\}$ for MC and PPM and $\alpha \in \{0.1, 0.9\}$. There is however little difference between the performance of the predictors. In the ChineseDB dataset, where we leverage a large number of users, for both of the *Scouters* and the *Rgulars*, the best performances are achieved by the MC models, whereas the lowest performances are achieved by the SPM particularly with $\alpha = 0.9$. For the *Routiners*, we observe that the performance of these predictors varies slightly with different settings. In general, the achieved performances by the distinct predictors are substantially comparable. Therefore, we employ the MC(1) for our subsequent analyses only.

For comparison simplification reasons, Figure 7 reports the distribution of the practical predictability of the MC(1) predictor for all of the *Scouters*, *Regulars*, and *Routiners*.



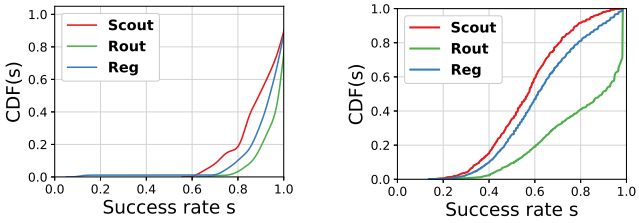
(a) Scout (b) Reg (c) Rout

Fig. 5: Distribution of the success rate score s_u of each predictor per mobility profile for the Agg_gps dataset.



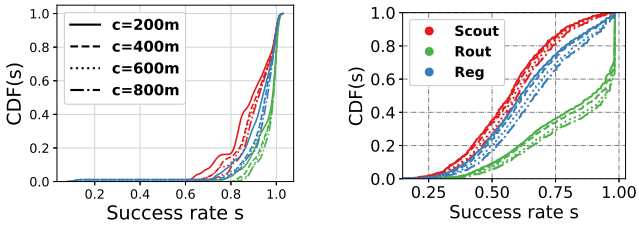
(a) Scout (b) Reg (c) Rout

Fig. 6: Distribution of the success rate score s_u of each predictor per mobility profile for the ChineseDB dataset.



(a) Agg_gps (b) ChineseDB

Fig. 7: Distribution of the success rate score s_u of the MC(1) predictor per mobility profile.



(a) Agg_gps (b) ChineseDB

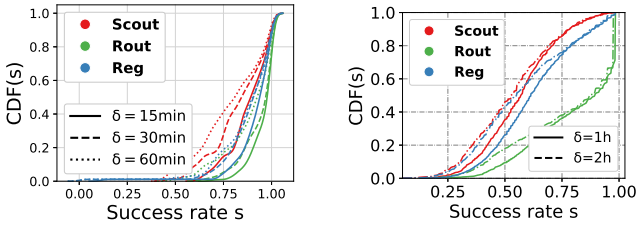
Fig. 8: Effect of spatial granularity on of the success rate score s_u of the MC(1) predictor per mobility profile.

We can notice that the best performances are obtained with *Routiners* and the lowest ones with the *Scouters*. Emphasizing that the *Scouters* are the hardest to predict category of people, though they still present moments of regularity and thus, with high accurate prediction results (i.e., 80% of *Scouters* have an accuracy of prediction s_u above 80%).

Spatial resolution variation: in Figures 8a and 8b, we investigate the correlation between the size of the geographical cells and the accuracy of prediction s_u per mobility profile. For this purpose, we vary the size of the squared tessellations $c \in \{200, 400, 600, 800\}$ meters. Intuitively and according to previous studies [10] [12] the smaller are the locations, the less stationary behavior is ascertained in the mobility trajectories of the individuals, and hence, the less predictable they are.

Not surprisingly and in agreement with previous works, the accuracy of prediction improves substantially with the increase in the size of the geographical cells. This is observed with individuals of all the profiles without any distinction.

Temporal resolution variation: we now examine how the frequency of sampling affects the ability to predict the mobility trajectories of each profile. We reset the spatial resolution to $c = 200\,\text{m}$, and vary the frequency of sampling $\delta_{Agg\_gps} \in \{15, 30, 60\}$ minutes for the Agg_gps dataset and $\delta_{ChineseDB} \in \{1, 2\}$ hours for the ChineseDB dataset.



(a) Agg_gps (b) ChineseDB

Fig. 9: Effect of temporal granularity on of the success rate score s_u of the MC(1) predictor per mobility profile.

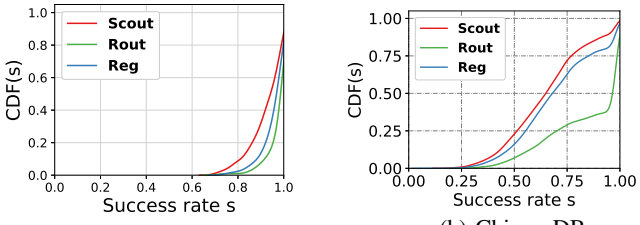
Figures 9a and 9b show that the accuracy of prediction decreases with the increase in the temporal resolution (when δ takes larger values). Indeed, the larger the frequency of sampling, the harder the capture of the stationary behavior of individuals' mobility.

Isolating explorations: we want to scrutinize the impacts of novelty-seeking on the predictability of users' trajectories. In the following, we reset the spatial resolution to $c = 200\,\text{m}$ and the temporal resolution to $\delta_{Agg\_gps} = 15min$ and

$\delta_{ChineseDB} = 1h$. For each user u we use the proposed methodology presented in Algorithm 2 with $level = 80\%$ to classify her locations into EV and RV. The places classified as RV are added to the set of known places $\mathcal{L}_u$.

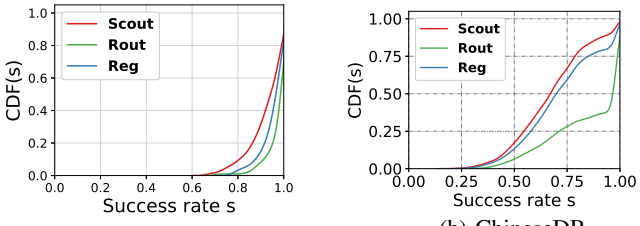
To evaluate the impacts of novelty-seeking on the accuracy of prediction s_u achieved by MC(1) we adopt three approaches:

- ***1st proof-of-impact case***: we remove the novelty-seeking records for all profiles and measure the accuracy of prediction s_u achieved by MC(1) with the new sequences. Clearly, this removal decreases the size of the trajectories and consequently can increase the accuracy of prediction. The corresponding results are depicted in Figure 10.
- ***2st proof-of-impact case***: as a first countermeasure to avoiding this size-related impact, we replace the novelty-seeking records with the last symbol met in the sequence. This action has the effect of adding a stationary period (equal to the size of each novelty-seeking period + 1). This approach is operated to assess whether the performance of the MC(1) predictor is only affected by the change in the length of the trajectories, or if *the exploration events play a role*. This substitution procedure is in favor of the predictor given that the stationary behavior is enhanced, the results are described in Figure 11.
- ***3st proof-of-impact case***: as a second countermeasure to avoid both size-related impacts and stationarity increase, we identify moments of novelty-seeking and substitute them with a random symbol met in the sequence. This procedure allows tackling both size-related effects and attenuating stationarity betterment impacts. Figure 12 shows the obtained results.



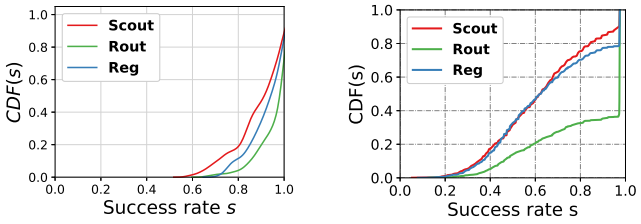
(a) Agg_gps (b) ChineseDB

Fig. 10: Effect of novelty-seeking records removal on the success rate score s_u of the MC(1) predictor per mobility profile.



(a) Agg_gps (b) ChineseDB

Fig. 11: Effect of novelty-seeking records replacement with stationarity stuffing on the success rate score s_u of the MC(1) predictor per mobility profile.



(a) Agg_gps (b) ChineseDB

Fig. 12: Effect of novelty-seeking records random replacement on the success rate score s_u of the MC(1) predictor per mobility profile.

The performance of MC(1) predictor indicates that, while the accuracy of prediction s_u is on average less than 60% (resp. 90%) for the least predictable class of users – i.e., *Scouters* – in the ChineseDB (resp. Agg_gps) dataset when considering novelty-seeking records (see Figure 7), Figure 10 shows that the predictor is considerably enhanced and can achieve an accuracy of prediction (on average) at least as high as 70% (resp. 95%) after removing exploration records. We have two hypotheses to explain this enhancement in the prediction accuracy: **H1:** the more irregular visits are omitted from the discrete mobility trajectory $\mathcal{T}_u$ of a user u, the more predictable she is. **H2:** decreasing the lengths of a discrete mobility trajectory $\mathcal{T}$, allows the predictor to achieve better performance.

Replacing novelty-seeking records allows us to assess one of the origins of the betterments in the predictive performance of the MC(1) predictor. Figures 11a, and 11b show that when replacing novelty-seeking records by adding stationarity, the accuracy of prediction is in fact further improved compared to the removal approach. Whereas the replacement of novelty-seeking records with random locations does not necessarily improve the performance compared to the removal approach, it still achieves comparatively higher performances with regard to the original trace (see Figure 12). Particularly, *Scouters* who represent the most vulnerable category to the exploration phenomenon (their average prediction accuracy s_u is above 60%.). These findings allow us to corroborate the harmful effects that exploration events have on the predictive performance of the classical MC predictor. Moreover, *Scouters* are more affected by these events as it could be seen within Figures 11 and 12, isolating these events engendered substantial improvements in the practical predictability of the *Scouters* compared to the other profiles.

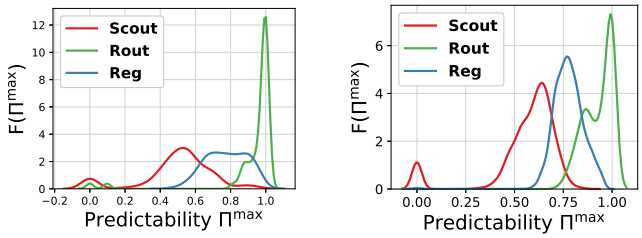
Summarizing remarks: in a nutshell, in the next-cell prediction task individuals of all profiles are impacted by both the quality of the data and novelty-seeking. Increasing the temporal resolution of the data or enlarging the size of the spatial cells allows achieving higher accuracies of prediction s_u. Moreover, although the high performances that are usually achieved with this prediction task mainly due to stationarity effects, moments of novelty-seeking do alter the predictive performance.

C. Next-place

We now tackle the next-place prediction formulation. We first reconstruct the discrete mobility trajectories $\mathcal{T}_{u,c}$ of the individuals by removing stationarity records to fit the next-place prediction scenario. Next, we measure the theoretical Π^{max} and practical s_u predictability of the discrete mobility trajectories $\mathcal{T}_{u,c}$. After that, since this formulation of prediction is independent of the temporal resolution, we do not investigate the impacts of the quality of the data factor on this formulation of the prediction task. Finally, we measure the predictability of the three mobility profiles when removing and replacing moments of novelty-seeking.

1) Discrete mobility trajectories refurbishment: the next-place prediction formulation refers to the prediction of transitions between places. This formulation is more exposed to uncertainty as the stationarity behavior is omitted. Namely, the next-place prediction is about forecasting the next location where an individual is going to be and that should be different from her current one. Thereby, given the discrete mobility trajectory $\mathcal{T}_{u,c} = \{(id_0, t_0), (id_1, t_0 + \delta), \ldots, (id_N, t_0 + N\delta)\}$ of a user u, we identify consecutive tuples that have the same location id and keep only the first tuple. Note that in this case the frequency of sampling δ is not constant and the size of the mobility trajectories is smaller.

2) Theoretical predictability: for each user u of each profile, as in Section V-B, we estimate the upper bound of the theoretical predictability Π^{max} of the stochastic sequence $x_1^N = \{x_1, \ldots, x_N\}$ extracted from her refurbished discrete mobility trajectory $\mathcal{T}_{u,c}$.



(a) Agg_gps (b) ChineseDB

Fig. 13: Distributions of the upper bound of the theoretical predictability Π^{max} for individuals of each mobility profile.

The distributions of the upper bound of the theoretical predictability Π^{max} for individuals of each mobility profile are presented in Figure 13. We can see that consistent with findings from previous studies [12], the predictability is markedly decreased for both of the Agg_gps and ChineseDB datasets. Additionally, the Figure reveals that *Scouters* are still the least predictable individuals, even in this formulation of human mobility prediction, whilst *Routiners* are the most predictable ones.

3) Practical predictability: we evaluate the predictive performance achieved by the four predictors MC, PPM, SPM, and ALZ with the next-place prediction task.

We apply the four predictors MC, PPM, SPM, and ALZ to the next-place prediction task.

Figures 14 and 15 show the accuracy of prediction s_u achieved by each predictor with individuals of each profile. Clearly, the accuracy of prediction s_u is markedly lower than in the next-cell prediction task. In particular, the SPM performs poorly with the next-place prediction especially with *Scouters*. The remaining predictors have comparable performances, with an average accuracy around 10%, 24%, and 60% (25%, 26%, 34%) for Agg_gps (ChineseDB) dataset for *Scouters*, *Regulars*, and *Routiners* respectively. The achieved performances by the distinct predictors are substantially comparable. Therefore, to homogenize with the next-cell evaluation in what follows we use MC(1).

For comparison simplification, Figures 16a and 16b display the accuracy of prediction of the MC(1) predictor in the next-place prediction scenario in CDF curves, one for each mobility profile: *Scouters*, *Regulars*, and *Routiners*. We can observe that the MC(1) predictor fares poorly, notably with the *Scouters*, where 85% of them have an accuracy of prediction below 20% in the Agg_gps dataset and below 40% for the ChineseDB. This conveys that the uncertainty in a typical individual's mobility trace is more significant than in the next-cell prediction.

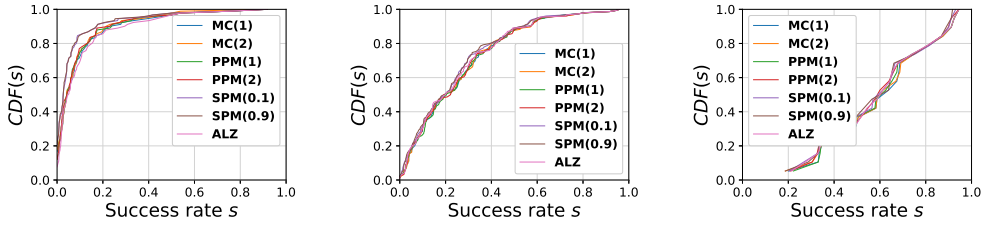
Isolating explorations: we now analyze the impacts of exploration events on the next-place prediction formulation. We start by identifying moments of novelty-seeking per-user using the visitation frequency-based methodology Algorithm 2 with $level = 80\%$. Next, we employ three methods to emphasize the impacts of novelty-seeking:

- ***1st proof-of-impact case***: as in the next-cell prediction analysis, we remove the novelty-seeking records (see Figure 17).
- ***2st proof-of-impact case***: to avert size-related impacts, unlike in the previous prediction task we do not replace novelty-seeking records by adding stationary periods as it goes against the definition of the next-place formulation. Hence, given the last visited location i if the current location j is assumed to be an exploration we replace the novelty-seeking records j with the most frequent location that usually appears after i. The results are depicted in Figure 18.
- ***3st proof-of-impact case***: slightly different from the 3 rd proof of impacts of the previous prediction formulation, we replace moments of novelty-seeking by a random symbol met in the sequence that is different from the last visited location. Figure 12 shows the obtained results.

Figures 17 displays the accuracy for the MC(1) predictor while keeping only familiar visits in the mobility traces. The accuracy of prediction is remarkably enhanced compared to the next-cell formulation case for all profiles notably for *Scouters* the average score is above 15% (above 50%) for the Agg_gps (ChineseDB).

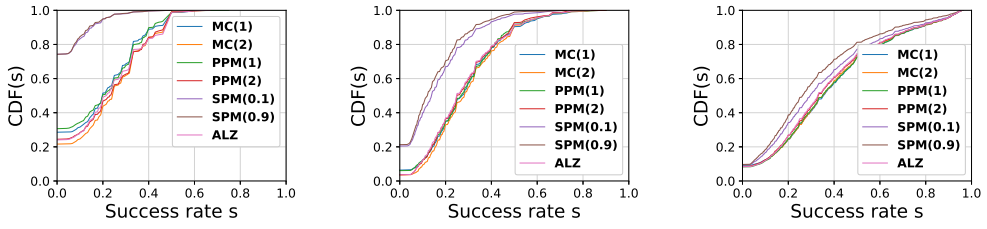
The replacement of novelty-seeking places by the most probable known location allows a further enhancement of the performance in particular for *Scouters* (see Figure 18).

Further, we can discern the substantial harmful effects of exploration envents on the predictability in the next-place prediction compared to the next-cell prediction. More importantly, the Figures show that *Scouters* are more impacted by the isolation of novelty-seeking records. The original median



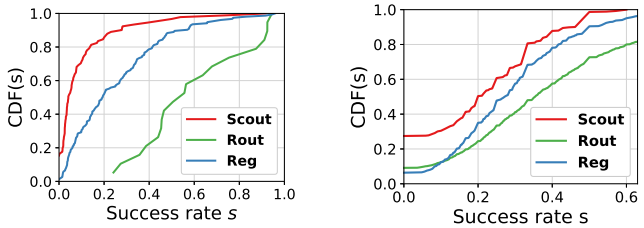
(a) Scout (b) Reg (c) Rout

Fig. 14: Distribution of the success rate score s_u of each predictor per mobility profile for the Agg_gps dataset.



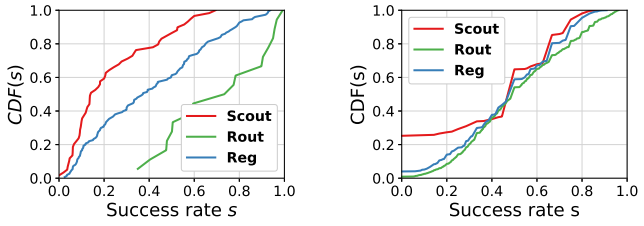
(a) Scout (b) Reg (c) Rout

Fig. 15: Distribution of the success rate score s_u of each predictor per mobility profile for the ChineseDB dataset.



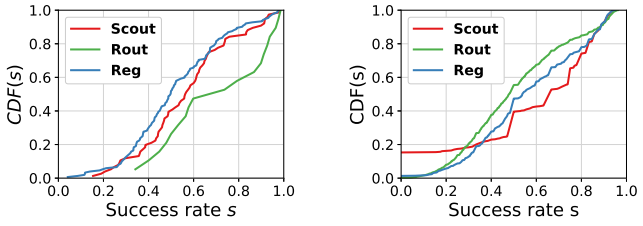
(a) Agg_gps (b) ChineseDB

Fig. 16: Distribution of the success rate score s_u of the MC(1) predictor per mobility profile.



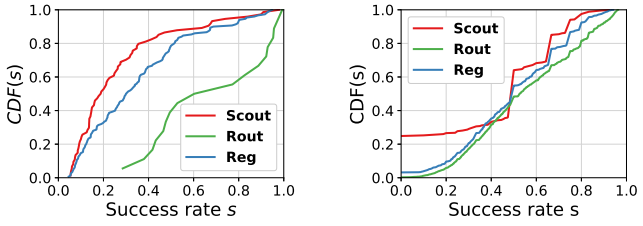
(a) Agg_gps (b) ChineseDB

Fig. 17: Effect of novelty-seeking records removal on of the success rate score s_u of the MC(1) predictor per mobility profile.



(a) Agg_gps (b) ChineseDB

Fig. 18: Effect of novelty-seeking records replacement with stationarity stuffing on of the success rate score s_u of the MC(1) predictor per mobility profile.



(a) Agg_gps (b) ChineseDB

Fig. 19: Effect of novelty-seeking records random replacement on of the success rate score s_u of the MC(1) predictor per mobility profile.

accuracy for *Scouters* is approximately less than 20% (see Figure 16), which is significantly lower than the performance of other profiles. Therefore, the removal or replacement of explorations events makes *Scouters* roughly as predictable as the other profiles.

Summarizing remarks: in summary, the next-place prediction task is a more challenging problem for individuals of all profiles. This formulation is more vulnerable to uncertainties as the stationarity behavior is overlooked. Therefore, the harmful effects of exploration activities are more discernible and have more impacts on the predictive performance.

VI. Final Remarks and Open Issues

Using real-world mobility traces, this paper proposes a new method for recognizing moments of novelty-seeking. Based on the exploratory tendencies of the population we revealed the existence of three groups of individuals with regard to their propensity to explore and discover new places, namely, *Scouters* (adventurous and prone to explore); (ii) *Routiners*, (steady and routinary), and (iii) *Regulars* (with medium behavior). This result has two major implications for the understanding of human mobility. First, in *mobility modeling*, individuals' propensity to explore i.e., degree of return metric, as well

as the elapsed time before the occurrence of an exploration event i.e., intermittency metric are substantial concepts that should be further investigated, to assess the existence of new novelty-seeking related scaling laws per mobility profile, and hence provide more consistent and generative models able to reproduce human trajectories. Second, in *mobility prediction* the proposed profiling allows distinguishing hard to predict individuals due to their exploration activity from the rest of the population, and therefore propose more adequate predictors.

Furthermore, we took a fresh look at the most significant factors affecting the predictability extent of individuals' mobility traces: (i) novelty-seeking, (ii) spatial and temporal resolutions, and (iii) prediction formulation. Utilizing our developed mobility profiling, we analyzed the effects of each factor on the predictability per profile. In accordance with previous studies, we showed that regardless of the mobility profiles, the next-cell prediction achieves higher degrees of practical and theoretical predictability compared to the next-place formulation. Particularly as a result, of the high stationarity presents in the next-cell prediction task. Besides, we asserted that increasing the size of the spatial cells leads to the increase of the stationarity and hence in the accuracy of prediction. Similarly, a finer-grained temporal resolution allows a higher capture of consecutive records with the same cell-id, and consequently a growth in stationary, which implies the achievement of higher prediction scores. More importantly, we shed light on the novelty-seeking phenomenon, as being a major factor impacting the predictability. Therefore, understanding the exploration phenomenon is fundamental to thoroughly model and predict human movements.

Meanwhile, further advances in understanding individual mobility are facing serious privacy issues. Indeed, human mobility trajectories containing geographical coordinates along with the date of collection are very sensitive data. Although the widespread of technological devices allowing the collection of individuals' mobility traces, their acquisition is a nontrivial process and is getting more and more complex. Moreover, a large number of sensitive professional and personal information can be inferred solely through an individual's mobility traces. Our future work can be divided into two directions: 1) investigate how our proposed mobility profiling can be adapted and used in a privacy-preserving environment. This means, given the mobility trace of a single individual, we aim at classifying her as a hard to predict individual or as a predictable one. 2) design a predictor that takes into account individuals' inclination to explore and that will leverage the spatiotemporal analysis presented in a previous work to yield an intuition on the next area where an individual is prone to be in case of an exploration.

REFERENCES

[1] D. H. M. M. D. E. S. M. M. G. L. M. Badr, Hamada S, "Association between mobility patterns and covid-19 transmission in the usa: a mathematical modelling study," *The Lancet Infectious Diseases*, vol. 20, no. 11, pp. 1247–1254, 2020.

[2] B. L. H. P. Lu, Xin, "Predictability of population displacement after the 2010 haiti earthquake," vol. 109, no. 29, pp. 11 576–11 581, 2012.

[3] L. X. T. A. G. R. v. S. J. Bengtsson, L., "Improved response to disasters and outbreaks by tracking population movements with mobile phone network data," *PLOS Medicine*, vol. 8, no. 8, pp. 1–9, 2011.

[4] L. Aalto, N. Göthlin, J. Korhonen, and T. Ojala, New York, NY, USA.

[5] T. T. A. Nadembega, A. Hafid, "Mobility-prediction-aware bandwidth reservation scheme for mobile networks," *IEEE Trans. Veh. Technol.*, vol. 64, no. 6, pp. 2561–2576, 2015.

[6] M. C. Gonzalez, C. A. Hidalgo, A. L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, pp. 779–782.

[7] C. Song, T. Koren, P. Wang, A. Barabási, "Modelling the scaling properties of human mobility," *Nature Physics*, vol. 6, p. 818–823, 2010.

[8] C. Song, Z. Qu, N. Blumm, A.-L. Barabási, "Limits of Predictability in Human Mobility," *Science*, vol. 327, pp. 1018–1021, 2010.

[9] M. Lin,W.-J. Hsu, Z. Qi Lee, "Predictability of individuals' mobility with high-resolution positioning data," in *UbiComp '12*, 2012.

[10] X. Lu, E. Wetter, N. Bharti, A. J. Tatem, L. Bengtsson , "Approaching the Limit of Predictability in Human Mobility," *Scientific Reports*, vol. 3, no. 2923.

[11] H. L. Huiji Gao, Jiliang Tang, "Mobile location prediction in spatio-temporal context," 2012.

[12] A. Cuttone, S. Lehmann, M. C. Gonzalez, "Understanding predictability and exploration in human mobility," *EPJ Data Science*, vol. 7, no. 1, 2018.

[13] K. J. J. L. L. B. Sand Jensen, J. Eg Larsen, "Estimating human predictability from mobile sensor data," in *2010 IEEE MLSP International Workshop*, 2010, pp. 196–201.

[14] J. G. D. B. G. Smith, R. Wieser, "A refined limit on the predictability of human mobility," in *PerCom*, 2014, pp. 88–94.

[15] L. Pappalardo, F. Simini, S. Rinzivillo, D. Pedreschi, F. Giannotti, A.-L. Barabási, "Returners and explorers dichotomy in human mobility," *Nature Communications*, vol. 6, no. 8166, 2015.

[16] L. Scherrer, M. Tomko, P. Ranacher, R. Weibel, "Travelers or locals? Identifying meaningful sub-populations from human movement data in the absence of ground truth," *EPJ Data Science*, vol. 7, 2018.

[17] V. A. C. Amichi, Licia, M. Crovella, and A. A. Loureiro, "Understanding individuals' proclivity for novelty seeking." Association for Computing Machinery, 2020.

[18] G. G. C. R. J. M. L. T. L. R. M. J. J. R. F. S. M. T. H. Barbosa, M. Barthelemy, "Human mobility: Models and applications," *Physics Reports*, vol. 734, pp. 1 – 74, 2018.

[19] M. A. G. B. Kulkarni, Vaibhav, "A mobility prediction system leveraging realtime location data streams: Poster." New York, NY, USA: Association for Computing Machinery, 2016.

[20] J. Wang, X. Kong, F. Xia, and L. Sun, "Urban human mobility: Data-driven modeling and prediction," *SIGKDD Explor. Newsl.*, vol. 21, no. 1, p. 1–19, 2019.

[21] E. L. Ikanovic, A. Mollgaard, "An alternative approach to the limits of predictability in human mobility," *EPJ Data Science*, vol. 6, no. 1, 2017.

[22] K. Jaffres-Runser, G. Jakllari, T. Peng, V. Nitu, "Crowdsensing Mobile Content and Context Data: Lessons Learned in the Wild," in *PerCom Workshops*, 2017.

[23] S. BenMokhtar, A. Boutet, L. Bouzouina, P. Bonnel, O. Brette, L. Brunie, M. Cunche, S. D'Alu, V. Primault, P. Raveneau, H. Rivano, R. Stanica, "PRIVA'MOV: Analysing Human Mobility Through Multi-Sensor Datasets," in *NetMob*, 2017.

[24] Y. C. X. X. W. M. Y. Zheng, Q. Li, "Understanding mobility based on gps data," in *UbiComp*, 2008, pp. 312–321.

[25] W. M. Y. Zheng, X. Xie, "Geolife: A collaborative social networking service among user, location and trajectory," in *Invited paper, in IEEE Data Engineering Bulletin*, vol. 33, 2010, pp. 32–40.

[26] X. X. W. M. Y. Zheng, L. Z., "Mining interesting locations and travel sequences from gps trajectories," in *ACM WWW*, 2009, pp. 791–800.

[27] C. V. A. F. M. S. C. Chen, G., "Complete Trajectory Reconstruction from Sparse Mobile Phone Data," *EPJ Data Science*, 2019.

[28] M. Papandrea, K. Ke. Jahromi, M. Zignani, S. Gaito, S. Giordano, G. P. Rossi, "On the properties of human mobility," *Computer Communications*, vol. 87, no. 1, pp. 19–36, 2016.

[29] M. K. Cowles and B. P. Carlin, "Markov chain monte carlo convergence diagnostics: A comparative review," *J. Am. Stat. Assoc.*, vol. 91, no. 434, pp. 883–904, 1996.

[30] A. Moffat, "Implementing the ppm data compression scheme," in *IEEE TCOM*, vol. 38, no. 11, pp. 1917 – 1921.

[31] I. A. P. Jacquet, W. Szpankowski, "A universal predictor based on pattern matching," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1462–1472, 2002.

[32] K. Gopalratnam, D. J. Cook, "Active LeZi: An Incremental Parsing Algorithm for Sequential Prediction," *Int. J.Artif. Intell. Tools*, vol. 4, p. 917–930, 2004.

Licia Amichi received the B.S. degree in computer science from Pierre and Marie Curie University, Paris, France, in 2016, and the M.S. degree in computer science option Smart Mobility and Internet of Things from Sorbonnes University, France, in 2018. From February to August 2018, she was an intern at National Institute of Informatics, Tokyo, Japan, where she worked on IoT wireless networks. Since October 2018, she is a PhD student at INRIA Saclay and École Polytechnique, France. Her current research interests include mobility modeling, prediction, and IoT wireless network.

Aline Carneiro Viana is a Senior Reseach Director (DR) at Inria, where she leads the team TRiBE team. After a 1-year sabbatical leave at the TKN Group of the TU-Berlin, Germany, she got her habilitation degree from UPMC - Sorbonne Universités, France in 2011. Dr. Viana got her PhD in Computer Science from the UPMC - Sorbonne Universités in 2005. Her research addresses the design of solutions for tactful networking, smart cities, mobile and self-organizing networks with the focus on human behavior analysis. She is a recipient of the French Scientific Excellence award since 2015 and for 6 years now and was nominated in 2016 as one of the ”10 women in networking/communications that you should Watch” (1st-year nomination of N2Women community). She has published more than 95 papers, presented in these fields in top-tier conferences and in important peer-reviewed journals. She has been involved in the organizing committee as well as been a TPC member of major conferences (ACM SenSys, ACM Mobicom, IEEE Infocom, IEEE SECON, IEEE LCN). She is Area Editor of ACM Computer Communication Review (ACM CCR) and member of the editorial Board of Urban Computing Spring book series, and Elsevier Ad Hoc Networks. She has coordinated French and International projects (ANR MITIK and EU CHIST-ERA MACACO, and STIC AmSud UCOOL).

Mark Crovella is a Professor in the Department of Computer Science at Boston University, where he has been since 1994. During 2003-2004 he was visiting faculty at Laboratoire d'Infomatique de Paris VI (LIP6), and in 2018-2019 he was visiting faculty at LIP6, INRIA Paris, and LINCS Paris. His research interests span both computer networking and network science. Much of his work has been on improving the understanding, design, and performance of parallel and networked computer systems, mainly through the application of data mining, statistics, and performance evaluation. Professor Crovella is co-author of Internet Measurement: Infrastructure, Traffic, and Applications (Wiley Press, 2006) and is the author of over two hundred papers on networking and computer systems. Between 2007 and 2009 he was Chair of ACM SIGCOMM. Professor Crovella is a Fellow of the ACM and the IEEE.

Antonio A.F. Loureiro Antonio A.F. Loureiro received the B.Sc. and M.Sc. degrees in computer science from Universidade Federal de Minas Gerais (UFMG), Brazil, and the Ph.D. degree in computer science from The University of British Columbia, Canada. He is currently a Full Professor with UFMG, where he leads the research group on mobile ad hoc networks. He was the recipient of the 2015 IEEE Ad Hoc and Sensor (AHSN) Technical Achievement Award and the Computer Networks and Distributed Systems Interest Group Technical Achievement Award of the Brazilian Computer Society. His main research areas include ad hoc networks, mobile computing, and distributed algorithms. In the last 20 years, he has published regularly in international conferences and journals related to those areas, and have also presented keynotes and tutorials at international conferences.