Revealing an inherently limiting factor in human mobility prediction

Licia Amichi, Aline Carneiro Viana, Mark Crovella, Antonio A. F Loureiro

Revealing an inherently limiting factor in human mobility prediction

Licia Amichi[†⋆], Aline Viana Carneiro[⋆], Mark Crovella[‡], and Antonio Loureiro[§]
[†] Ecole Polytechnique (IPP), France [⋆] Inria, France [‡]Boston University, USA [§] Federal University of Minas Gerais, Brazil
Email: licia.amichi@inria.fr, aline.viana@inria.fr, crovella@bu.edu, loureiro@dcc.ufmg.br

***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 areas. Yet, neglecting novelty-seeking activities at first glance appears to be inconsequential on the ability to understand and predict individuals' trajectories. We claim and show the opposite in this work: exploration-like visits strongly impact mobility understanding and anticipation. We start by proposing a new approach to identifying exploration visits. 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., 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, in the next-place prediction task, from 40% to 90% of predicted locations are wrong, notably with *Scouters*.**

***Index Terms*—Individual Mobility Analytic, Mobility Understanding and Profiling, Predictability, Prediction**

I. Introduction

The understanding and modeling of daily mobility of individuals became an accessible domain of study given the ubiquity of mobile devices, Internet connectivity, and positioning systems such as the Global Positioning System (GPS). In this context, several representative models are proposed to reproduce individuals' trajectories and various robust predictors to forecast their future locations. Indeed, accurate mobility understanding and predictors are crucial for epidemic prevention (e.g., the COVID-19 pandemics) [1], disaster response, and traffic management [2, 3]. Besides, such accuracy improves the services offered by pervasive computing applications [4] and provides energy-efficient and cost-effective network infrastructures [5].

Previous studies [6, 7] show that 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 a 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.

Given the difficulties involved in anticipating location visits in mobility-related behavior, a frequently 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, 8–10] of the mobility traces or the prediction accuracy achieved by different predictors (practical predictability) [11–13]. Nevertheless, the empirical results suggest that the predictability takes variable values ranging from under 40% to higher than 90% [13]. 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 predictability?*

Prior investigations demonstrate that the quality of the data considerably affects predictability, namely the temporal and spatial resolutions [9, 13–15]. 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 widespread versions are the *next-cell* and the *next-place* tasks, in 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 challenging to analyze due to data acquisition and privacy concerns [16].

Withal, a non-negligible impacting factor and focus of this paper is the tendency of individuals to explore and discover new places. Novelty-seeking is highly present in our daily lives since we are continuously hunting for new places to go to [13]. 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 [17, 18].

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 novelty-seeking activities on the predictability extent of individuals' mobility traces. This paper substantially extends the work detailed in [19] regarding novelty-seeking capture and the impact/evaluation of novelty-seeking tendencies on prediction. Here, we build on this prior

effort by presenting much more comprehensive investigation and offerings:

- We first improve our novelty-seeking identification methodology presented in [19]. We propose a per-user approach based on the average visitation frequency of the distinct locations 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. Following, we exploit it to identify moments of novelty-seeking. We endorse our proposal by a thorough experimental validation and a performance comparison with a state-of-the-art approach (cf. Section IV-B).
- We split individuals' visits into two categories based upon the two captured types of locations: explorations and returns (cf. Section IV-C). Only first occurrences of EV locations in the mobility traces are viewed as moments of novelty-seeking. This observation reduces the overestimation of novelty-seeking events present in state-of-the-art methods.
- Finally, we are the first (to the best of our knowledge) to measure and quantify the impacts of exploration-like visits on the potential predictability of individuals of each mobility profile. Using the two most widespread prediction formulations, i.e., the next-cell and next-place prediction and different sources of data. We corroborate that:
  - *Scouters* are the least predictable users due to their high exploration activities. When considering the next-place prediction, on average, 75% of the predictions are incorrect for the *Scouters* vs. less than 40% for the *Routiners*.
  - Exploration events are one of the principal origins behind the low predictive performance. The higher the exploration tendency of an individual, the less predictable she is (*Scouters* vs. *Routiners*). Additionally, the removal and substitution of exploration visits enhance the predictive performance. This emphasizes the role of novelty-seeking in making human mobility less foreseeable and compels the need to thoroughly understand the exploration phenomenon, to allow the design of representative and accurate predictors and models.
  - Considering exploration-like visits when developing predictors is critical for some categories of the population, namely, *Scouters* that are highly impacted by such events.

  Additionally, we also evaluate the effects of the most reviewed impacting factors, i.e., temporal resolution and spatial resolution, on the predictability extent of each profile.

The remainder of this paper is organized as follows. We start with an overview of the related work in 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 factors impacting the potential predictability of the mobility traces of each profile. Finally, we discuss 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. 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.

II. Related Work

Human mobility is extensively scrutinized to understand the mechanisms ruling an individual's movements. Several studies demonstrate that human movements are far from random and have a high degree of predictability [20].

Song et al. [8] propose, 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 Call Detail Records (CDR) dataset of 50,000 users, their study reveals a 93% potential predictability in an individual's mobility trace. Several subsequent studies refine the predictability upper bound estimation $\Pi^{\max}$. For instance, Lu et al. [2] determine that in a CDR dataset containing the mobility trace of 2.9 million individuals, the upper limit of their predictability is about 85%.

Building upon the above findings, many advanced *predicting algorithms* are designed to approach the theoretical predictability, such as Markovian predictors [11], Bayesian network models [12], or advanced deep learning approaches [21]. Lu et al. [11] seek to approach the theoretical limits of predictability and utilize a Markov Chain (MC) based predictor with a varying order and show that the practical predictability (denoted by s in this paper) reaches 91%. Moreover, they show that higher-order MC models do not significantly improve practical predictability. Gao et al. [12] propose a novel predictor based on Bayes Networks and find that using the Nokia Mobile Data Challenge that contains the mobility traces of 80 users, the practical predictability is about 50%.

Subsequent studies employ the same approach as in [8] 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. [14] examine 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 report high potential predictability for the data. Additionally, they show 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. Similarly, Lin et al. [9] use a high spatial and temporal resolution GPS dataset of 40 individuals. They show 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. [15] show that predictability is correlated with the temporal resolution and has an inverse correlation with the spatial resolution.

Type of prediction: Ikanovic et al. [22] emphasize the origins of the high potential predictability of individuals' mobility

obtained in earlier studies [2, 8]. They focus on the next-place prediction that considers moments of transitions only, i.e., moving from a place to a distinct one, then estimate the upper bound limit of the predictability and obtain significantly lower performance of approximately 71%. Thereby, they validate 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. [13] analyze 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 show the importance of considering individuals' tendencies to explore and discover new locations when modeling their mobility [7]. Notably, Cuttone et al. [13] highlight the importance of considering the exploration phenomenon when designing mobility predictors. 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 leads 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. [17] discern two categories of people: explorers and returners. They base their classification on the number of regularly visited places: explorers visit many locations regularly, whereas returners limit their mobility between a few places.

Besides, Scherrer et al. [18] use an unsupervised approach to classify individuals into travelers and locals. Travelers have 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 tendencies. Therefore, in our previous work [19], we propose a mobility profiling based on individuals' tendency to explore that we further improve in this paper. We reveal the existence of three main categories of individuals: (1) *Scouters:* whose proclivity for novelty-seeking is the most eminent all over the week and have a more spread spatial mobility; (2) *Routiners:* 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. For the first time in literature, we present a newly-tailored method to recognize novelty-seeking moments. By the mean of this, we deeply improve our previously proposed mobility profiling [19]. On the other hand, we study predictability extents, which is the main focus of this paper, and evaluate how each of the prediction formulations, the quality of the data, and the proclivity for novelty-seeking influence the predictability.

Dataset	Number of users	Duration	Sampling Frequency
Macaco [23]	132	34 months	5 min
Privamov [24]	100	15 months	few seconds
Geolife [25–27]	182	64 months	1 to 5 seconds
ChineseDB [28]	642K	2 weeks	1 hour

TABLE I: Datasets description.

III. Data Description

We use two types of data sources; three GPS and one 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 hereafter.

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 [23]. 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 over 34 months. Each tuple has a unique ID, which relates to a specific user along with her GPS coordinates (latitude and longitude) and a timestamp.

Privamov: It contains mobility traces collected by the Privamov sensing campaign [24], capturing the spatio-temporal footprints of 100 volunteers over 15 months around a city in Europe. The sampling frequency is of the order of seconds.

Geolife: The last GPS data source is the Geolife public dataset collected by Microsoft Research Asia [25–27]. The dataset stores 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 timestamped and geo-referenced records of voice phone calls and SMS of mobile network subscribers, called 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]. It provides

[1]The collection was initiated by Shanghai University [28].

aggregated human footprints with a frequency of one location per hour for two weeks. The locations 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

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 $t_i = t_0 + i \times \delta$, where i identifies a particular timestamp.

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 [13], 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, we re-sample 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. [28] and complete them as follows,

- 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}), and *workplace B* (id_{wp_B}) locations are identified and if existing,

- if a record is missing at $t_x \in [10\text{ am}, 11\text{ am}]$, we create a new record with the timestamp t_x and the cell identifier of the *workplace A* id_{wp_A}. Namely, we add the record (id_{wp_A}, t_x) to the mobility trajectory $\mathcal{T}_{u,c}$.
- 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 $\mathcal{T}_{u,c}$,
- if a record is missing at $t_x \in [2\text{ am}, 6\text{ am}]$, we add to $\mathcal{T}_{u,c}$ the record (id_H, t_x).

D. Experimental settings

In what follows, we briefly describe the parameter settings we use in this study. Unlike in our previous works [19], 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 with at least one 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 discoveries of new places daily, for instance, going to a new restaurant or a new shop. Considering the imprecision and uncertainty of GPS systems, we claim cells of size 200 m $\times$ 200 m roughly correspond to daily regions of interest and can still capture discovery moments. Second, the higher the temporal resolution, the better 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 understand our methods' effectiveness and validate our work extensively.

GPS data aggregation: Due to the small number of individuals in the GPS data sources, we aggregate the filtered and manipulated GPS traces and label this new dataset as *Agg_gps*. The aggregation consists of the simple concatenation of the GPS datasets. Namely, after uniformizing the frequency of sampling, i.e., 5 min, and the duration of data collection, i.e., 1 month, all the filtered GPS traces of the different datasets are added the new *Agg_gps* dataset. Starting from Section V, 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, 17], 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 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 initially presented in our work [19] and as in [7], we divide human movements into two primary states: *explorations and returns*. We define an (1) **exploration** (**E**) as *a discovery of a new place* and (2) a **return** (**R**) as *a visit to a previously visited locality. Note that a central point in the exploration investigation is to settle when a novelty-seeking moment occurs, the focus of this section. Hereafter, we describe our proposed strategy for this*

identification as well as our profiling methodology (cf. [19] for more details).

A. Formalization

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: (i) going back to historically known places and (ii) discovering new ones. In the exploring state **E**, discovering new places keeps the individual in state **E**. On the other hand, moving back to a known place, though recently explored, shifts the state from **E** to **R**. In the returning state **R**, visits to usual known places do not change the state, but a discovery of a new one shifts the state back to **E**. Hence, timestamped visited places in trajectories of users will be composed of records belonging to the returning state (**R**) or exploring state **E** –, $\mathcal{T}_{u,c} = \langle (id_0, t_0), (id_1, t_1), \dots (id_N, t_N) \rangle$ – and according to the sequence of visits, transitions between these two states will occur.

B. Novelty-seeking identification

Strictly speaking, for an individual, an exploration is the discovery of a new geographical place, i.e., a place never seen before. Nonetheless, existing works tackling the exploration problem consider the *first occurrence* of a place in the users' trajectories as an exploration [7, 13], which leads to an overestimation of exploration events. This means that the first appearance of the *home* or the *workplace* in the sequence is interpreted as a moment of novelty-seeking. Yet, overvaluing the frequency of exploration events might twist the understanding of the exploration problem. Hence, the question we focus on here is: *how can we distinguish users' novelty-seeking from routine-like visits?*

We propose a newly tailored per-user approach to distinguish between locations used for exploration visits and familiar regularly visited locations referred to as *visitation-frequency-based identification*. To verify and validate our approach, we conduct a performance comparison with the state-of-the-art location classification algorithm (*baseline*) proposed by Papandrea et al. [29].

1) Baseline identification: We use the widespread framework proposed by Papandrea et al. [29]. It is a seminal per-user scheme allowing the assessment of the importance of a place in a user's daily mobility. The baseline 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 [29], 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 Relevance_identification** ($\mathcal{T}_{u,c}$)
2: $T_{Relevance,u}, T_{MVP_u}, T_{OVP_u}, T_{EVP_u} \leftarrow \emptyset$
3: $F_u \leftarrow$ **Unique**($\mathcal{T}_{u,c}$) ▷ Extract the distinct visited locations
4: **for** j **in** F_u **do**
5: $T_{Relevance,u}[j] \leftarrow$ **Compute_relevance**(j) ▷ (1)
6: $T_{MVP_u}, T_{OVP_u}, T_{EVP_u} \leftarrow k$-means($T_{Relevance_u}, 3$)
7: **return** $T_{MVP_u}, T_{OVP_u}, T_{EVP_u}$

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 $F_u = \{id_1, id_2, \dots, id_n\}$ be the set of locations 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 Visitation_frequency_identification** ($\mathcal{T}_{u,c}$, $level$)
2: $w_u, T_{RV_u}, T_{EV_u} \leftarrow \emptyset$
3: $F_u \leftarrow$ **Unique**($\mathcal{T}_{u,c}$) ▷ Extract the distinct visited locations
4: **for** j **in** F_u **do**
5: $w_u[j] \leftarrow$ **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: **if** $w_u[j] \geq \overline{w}_u \times level$ **then**
10: T_{RV_u}.**Add**(j)
11: **else**
12: T_{EV_u}.**Add**(j)
13: **end if**
14: **end for**
15: **return** T_{RV_u}, T_{EV_u}

The parameter $level$ is critical for capturing moments of novelty-seeking. High values for $level$ can induce an overestimation of explorations, while small values lead to a

neglect of novelty-seeking moments. We quantify its impact, in Appendix A, 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).

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 (cf. Appendix A).

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, $ret_u = \{r_0, r_1, \ldots, r_m\}$, 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, $exp_u = \{e_0, e_1, \ldots, e_m\}$, 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 it repeatedly, before switching to revisits and vice versa.

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

The *degree of return* describes the exploration conducts of an individual compared to her returns. 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 in a row or just after a few discoveries, she goes back to a familiar location.

Following, we investigate whether the exploratory habit is the same among the population or if it is a distinctive property.

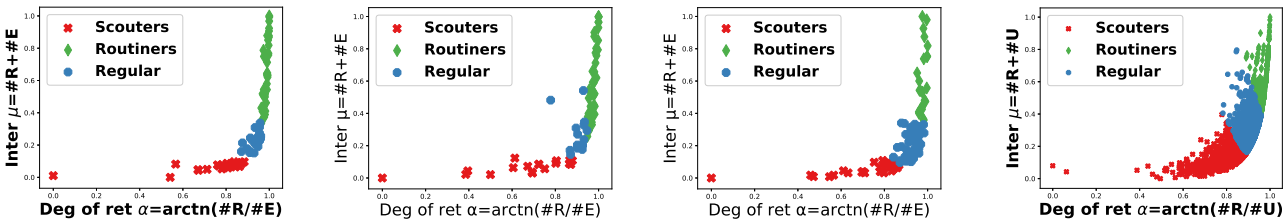
D. Mobility Profiling

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 A).

We 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. 1 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 observe that our metrics 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. They tend to be steady in the different states of the automaton M –i.e., they rarely break their routine. Hence, we deduce that these users rarely explore and prefer to stick among their 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*, they constantly alternate between explorations and revisits. Yet, their explorations are less important than *Scouters*'.

The proposed approach captures two major mobility features that fully describe the exploration phenomenon, i.e., *intermittency between returns and explorations, and the ratio*



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

Fig. 1: Mobility Profiling.

of explorations compared to returners, and allows a natural clustering of the individuals.

V. Revealing novelty-seeking impacts

In this section, we aim to evaluate explorations' effects (the first literature evaluation to the best of our knowledge), the quality of the data, and the prediction formulation, i.e., *next-cell* and *next-place* on the ability to forecast future visits of each of the *Scouters*, the *Regulars*, and the *Routiners*. We start by presenting the evaluation procedure followed in each prediction task. Next, we evaluate the performance reached in the next-cell task and present the impacts of each of (1) the quality of the data and (2) the proclivity to explore. Finally, we examine the attainable accuracy of prediction in the next-place prediction and show the effects of the impacting factor, in this case, novelty-seeking.

Due to the small number of individuals in each mobility profile for the GPS data sources, in what follows, we use the *Agg_gps* which is the concatenation of the GPS traces (cf. Section III-D).

A. Evaluation methodology

We measure per-profile the impacts of both the quality of data and individuals' tendency to explore on the two widespread prediction tasks. In view of the gaps in the leveraged GPS datasets, the limited number of users, and our filtering strategy in selecting only days having on average one record each 15 min, we set δ_{Agg_gps} to 15 min for the prediction tasks, and we keep 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.

The three pillars of our evaluation methodology are Prediction tasks formulation, Theoretical and Practical predictability, and Impacting factors. The evaluations in Sections V-B and V-C follow such steps for the next-cell and next-place prediction tasks, respectively. For the next-place task (cf. Section V-C), an additional step is required to remove stationary movements from the original sequence of visited locations. Figure 2 gives a general overview of the applied methodology.

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 (cf. Fig. 2–I):

- **Next-cell:** Given the mobility trace of an individual and considering a time window Δt, the next-cell prediction attempts to answer the following 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 [13, 15].
- **Next-place:** This formulation is independent of the temporal dimension [22]. 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 user's transition from her current location [13, 15].

2) Theoretical predictability: For each prediction task, we start by measuring the theoretical predictability of the mobility behavior of each of the *Scouters*, *Regulars*, and *Routiners* (cf. Fig.2–II). This will provide insights into the capacity of correctly forecasting the traces 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 $\Pi^{\max}$.

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 $\Pi^{\max}$ of the x_1^N sequence as in [8].

3) Practical predictability: Afterward, we estimate the practical predictability of each of the *Scouters*, *Regulars*, and *Routiners*. We compare the predictive performance of four state-of-the-art predictors, namely, Markov Chain (MC) [30], Predicting by Partial Matching (PPM) [31], Sampled Pattern Matching (SPM) [32], and Active LeZi (ALZ) [33] (cf. Fig. 2–III).

For the predictive performance comparison between the predictors, we measure the accuracy of the prediction achieved by each one. 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 correct predictions (accuracy of prediction) given by, $s_u = \frac{1}{N-N_s} \sum_{t=N_s+1}^{N} \mathbb{1}(x_t = x_t^* | x_1^{t-1})$,

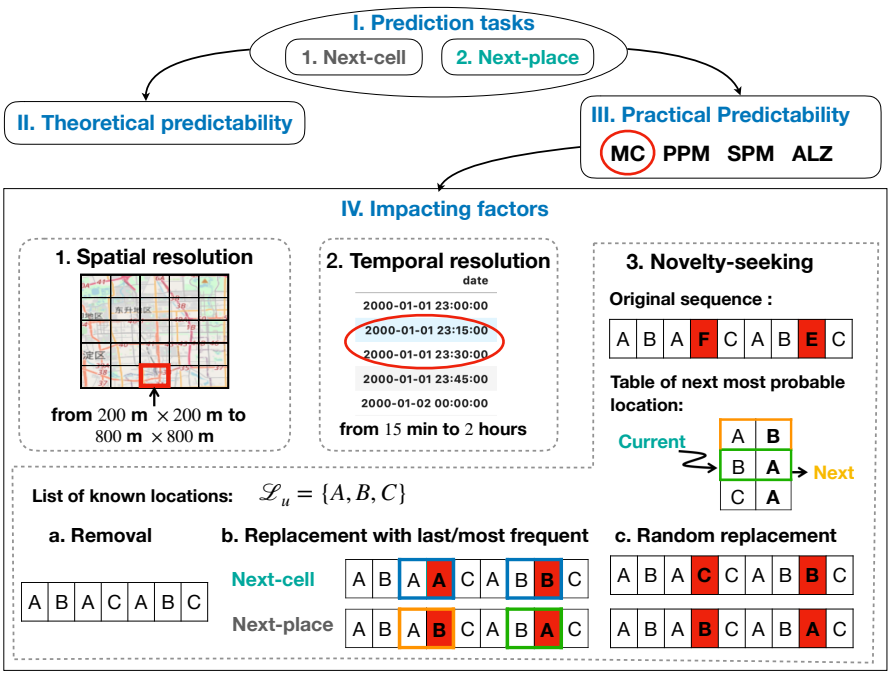


Fig. 2: General overview of the applied methodology.

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 $j \in [\![0, k]\!]$ [31]. 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 the quality of the data and individuals' tendency to explore *when relevant* on the predictive performance achieved by each prediction task (cf. Fig. 2–IV).

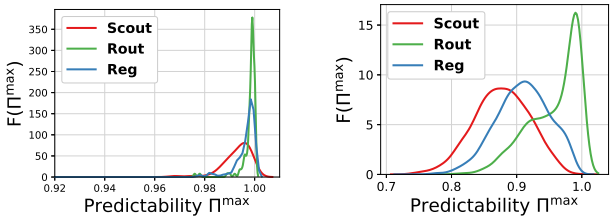
Spatial variation procedure: We investigate the effects of varying the spatial resolution on the accuracy of prediction s for users of each profile. We apply this variation with the next-cell task only, given that for the next-place task, actual points of interest identification are favored [13, 22] (cf. Fig. 2–IV.1).

Temporal variation procedure: In the case of next-cell prediction, we investigate the effects of varying temporal resolution on the accuracy of prediction. Provided that the next-place prediction task is independent of the temporal resolution, we do not investigate the impacts of the quality of the data factor on this formulation [22] (cf. Fig. 2–IV.2).

Exploration-like visits isolation procedure: We identify exploration-like visits 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 (cf. Fig. 2–IV.3). Quantifying impacts of exploration events on individuals of each profile allows to disclose what category of the population is more vulnerable and to what extent. Hence, it allows telling if special attention should be given to a fraction of the population that exhibits a high exploration tendency when developing models and predictors.

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 each profile. Next, we investigate the effects of varying the spatial and temporal resolutions on prediction accuracy. Finally, we identify exploration-like visits 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. 3: Upper bound of the theoretical predictability for each profile.

1) Theoretical predictability: Figure 3 portrays the distribution of the upper-bound predictability for each mobility profile for both the Agg_gps and the ChineseDB datasets. We observe the high inherent predictability of individuals of all profiles. Notably, individuals of the Agg_gps have a more eminent degree of potential predictability mainly due to the high frequency of sampling of the dataset $\delta_{Agg\_gps} = 15min$,

while $\delta_{ChineseDB}$ is set to 1 h. A higher frequency of sampling allows a more complete capture of the stationarity and, consequently, increases the degree of predictability [13]. More importantly, from Figure 3b, we note that the predictability $\Pi^{\max}$ 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. 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.

2) Practical predictability: The estimations of the predictability upper bound of individuals' trajectories reveal 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 [11] (see Section II). Hereafter, we evaluate the accuracy of prediction achieved by each of MC, PPM, SPM, and ALZ.

Figures 4 and 5 report the distribution of the practical predictability of the MC, PPM, SPM, and ALZ predictors concerning their possible parameters $k \in \{1, 2\}$ for MC and PPM and $\alpha \in \{0.1, 0.9\}$. We notice that the best performances are obtained with *Routiners* and the lowest ones with the *Scouters*. We emphasize that *Scouters* are the hardest category of people to predict. However, they still present moments of regularity and, thus, with high accurate prediction results (i.e., 80% of *Scouters* have an accuracy of prediction s above 80%). 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 *Scouters* and *Regulars*, the best performances are achieved by the MC models. In contrast, the SPM achieves the lowest performance, 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 only employ the MC(1) for our subsequent analyses.

For comparison simplification reasons, Figure 6 reports the distribution of the practical predictability of the MC(1) predictor for all of the *Scouters*, *Regulars*, and *Routiners*. We can notice that the best performances are obtained with *Routiners* and the lowest ones with the *Scouters*. We emphasize that *Scouters* are the hardest category of people to predict. However, 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%).

3) Impacting factors: We now investigate the impacts of spatial resolution, temporal frequency of sampling, and exploration-like visits on the next-cell prediction formulation.

Spatial resolution variation: In Figures 7a and 7b, we investigate the correlation between the size of the geographical cells and the accuracy of prediction s 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 [11] [13], the smaller the locations are, the less stationary behavior ascertained in the mobility trajectories of the individuals is. Hence, the less predictable they are.

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

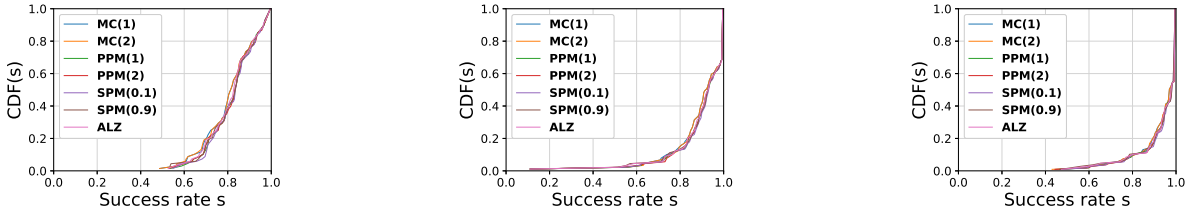
Temporal resolution variation: We now examine how the frequency of sampling affects the prediction. We reset the spatial resolution to $c = 200\,\text{m}$, and vary the temporal resolution $\delta_{Agg\_gps} \in \{15, 30, 60\}$ minutes for the Agg_gps and $\delta_{ChineseDB} \in \{1, 2\}$ hours for the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

Figures 8a and 8b show that the prediction accuracy decreases with the increase in the temporal resolution (when δ takes larger values). The larger the sampling frequency, the harder the capture of the stationary behavior of individuals' mobility.

Exploration-like visits isolation: We want to scrutinize the impacts of novelty-seeking on the predictability of users' trajectories. 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 $\mathcal{L}_u$. We adopt three novelty-seeking isolation strategies (cf. Fig. 2–IV.3):

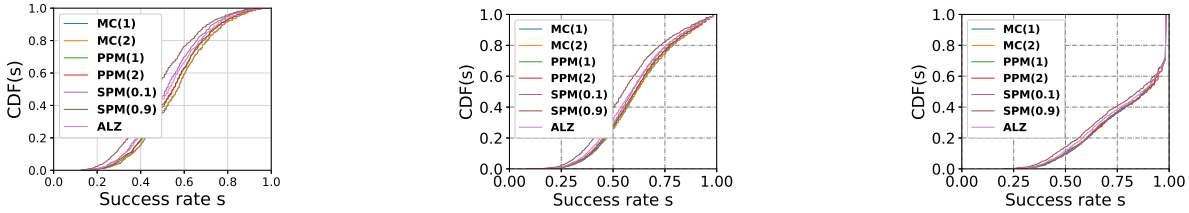
- ***1st proof-of-impact***: We remove exploration-like records for all profiles and measure the accuracy of prediction s achieved by MC with the new sequences (cf. Fig. 2–IV.3.a). This removal not only isolates exploratory visits but also decreases the size of the trajectories. Consequently, this strategy impacts the accuracy of prediction. The corresponding results are depicted in Figure 9.
- ***2nd proof-of-impact***: As a first countermeasure to avoiding this size-related impact, we replace the exploration-like records with the last symbol met in the sequence (cf. Fig. 2–IV.3.b). 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 predictor is only affected by the change in the length of the trajectories or if *the exploration-like visits play a role*. This substitution procedure favors the predictor once the stationary behavior is enhanced, as shown in Figure 10.
- ***3rd proof-of-impact***: As a second countermeasure to avoid both size-related impacts and stationarity increase, we identify exploration visits and substitute them with a random symbol found in the sequence (cf. Fig. 2–IV.3.c). This procedure allows tackling both size-related effects and attenuating stationarity betterment impacts. Figure 11 shows the obtained results.

The performance of the MC predictor indicates that, while the accuracy of prediction s 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 exploration-like records (see Figure 6), Figure 9 shows that the predictor is considerably enhanced and achieves an accuracy of prediction (on average) at least as high as 70% (resp. 95%)



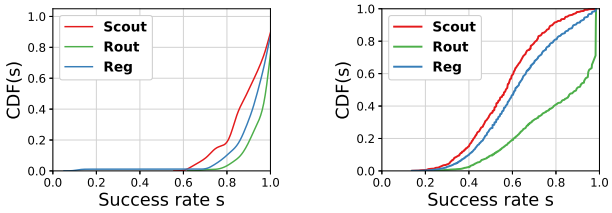
(a) Scout (b) Reg (c) Rout

Fig. 4: Distribution of the success rate score s_u of each predictor per profile for the Agg_gps dataset.



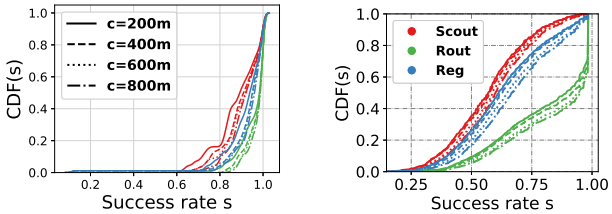
(a) Scout (b) Reg (c) Rout

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



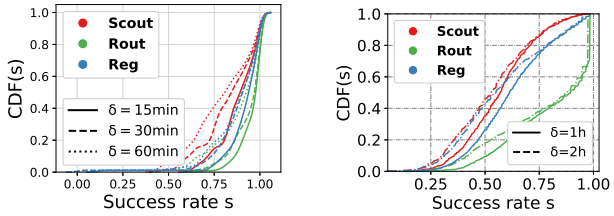
(a) Agg_gps (b) ChineseDB

Fig. 6: Success rate score s_u of the MC predictor per profile.



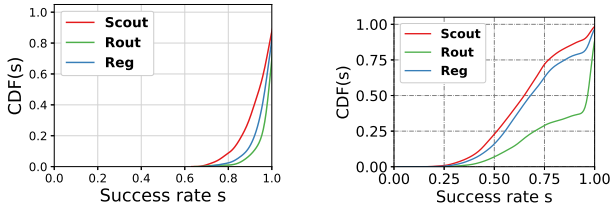
(a) Agg_gps (b) ChineseDB

Fig. 7: Effect of spatial granularity on the success rate score.



(a) Agg_gps (b) ChineseDB

Fig. 8: Effect of temporal granularity on the success rate score.



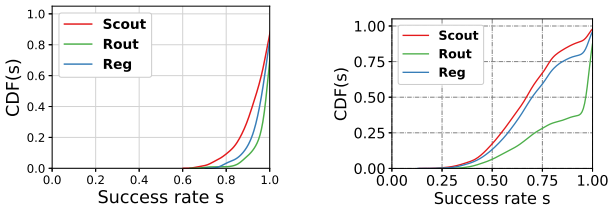
(a) Agg_gps (b) ChineseDB

Fig. 9: Effect of novelty-seeking removal on prediction.

after removing exploration-like 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.

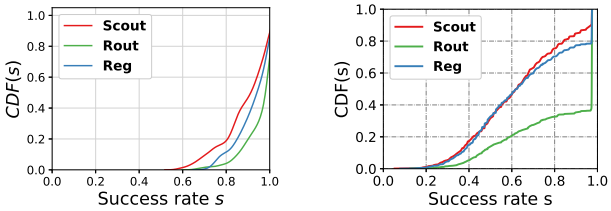
Replacing exploration-like visits allows us to assess one of the betterment's origins in the MC's predictive performance. Figures 10a and 10b show that when replacing exploration-like visits by adding stationarity, the accuracy of prediction is further improved compared to the removal approach. Whereas the replacement of exploration-like visits with random locations does not necessarily improve the performance compared to the removal approach, it still achieves comparatively higher performances concerning the original trace (see Figure 11). Particularly, *Scouters* represent the most vulnerable category to the exploration phenomenon (their average prediction accuracy s is above 60%.). These findings allow us to corroborate the harmful effects that exploration-like visits have on the predictive performance of the classical MC predictor. Moreover, *Scouters* are more affected by these events, as shown in Figures 10 and 11. The isolation of these events engendered substantial improvements in the practical predictability of the *Scouters* compared to the other profiles.

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



(a) Agg_gps (b) ChineseDB

Fig. 10: Effect of novelty-seeking replacement on prediction.



(a) Agg_gps (b) ChineseDB

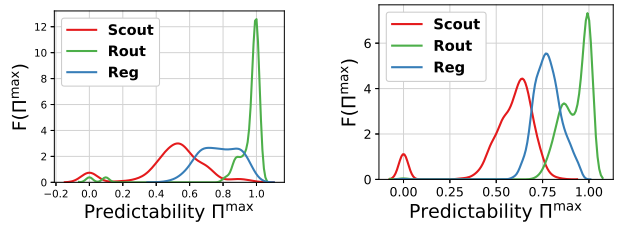
Fig. 11: Effect of novelty-seeking random replacement on prediction.

C. Next-place

We now tackle the next-place prediction task. We first reconstruct the mobility trajectories by removing stationarity records to fit the next-place prediction scenario. Next, we measure the theoretical $\Pi^{\max}$ and practical s predictability of the mobility traces. After that, since this prediction formulation is independent of the temporal resolution, we do not investigate the impacts of the data quality factor on this formulation of the prediction task. Finally, we measure the predictability of the three mobility profiles when isolating exploration-like visits.

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. Thereby, given the discrete mobility trajectory $\mathcal{T}_{u,c}$ of a user u, we identify consecutive tuples that have the same location id and keep only the first tuple. Note that the sampling frequency δ is not constant in this case, 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 $\Pi^{\max}$ of the stochastic sequence $x_1^N = \{x_1, \ldots, x_N\}$ extracted from her refurbished discrete mobility trajectory $\mathcal{T}_{u,c}$ (cf. Fig. 2–II).



(a) Agg_gps (b) ChineseDB

Fig. 12: Upper bound of the theoretical predictability for each profile.

The distributions of the upper bound of the theoretical predictability $\Pi^{\max}$ for individuals of each mobility profile are presented in Figure 12. We see that consistent with findings from previous studies [13], the predictability is markedly decreased for both Agg_gps and ChineseDB datasets. Additionally, Figure 12 reveals that *Scouters* are still the least predictable, even in this formulation of human mobility prediction, while *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 (cf. Fig. 2–III).

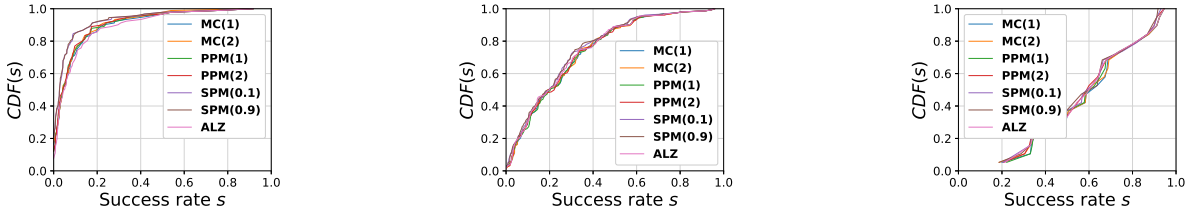
Figures 13 and 14 show the accuracy of prediction s achieved by each predictor with individuals of each profile. The accuracy of prediction s is markedly lower than in the next-cell prediction task. In particular, the SPM performs poorly with the next-place prediction, especially with *Scouters*. The 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 15a and 15b 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 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.

4) Impacting factors: Recall that we only evaluate the impacts of exploration-like visits on the prediction accuracy in this prediction formulation. The next-place prediction task is independent of the temporal resolution and varying the spatial resolution is not adequate [22].

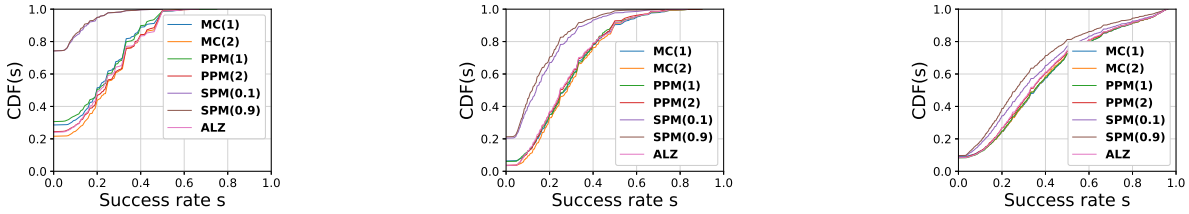
Exploration-like visits isolation: We now analyze the impacts of exploration events on the next-place prediction task. We start by identifying exploration-like visits per-user using the visitation frequency-based methodology Algorithm 2 with $level = 80\%$. Next, we employ three methods to emphasize the impacts of novelty-seeking, also exemplified in Fig. 2–IV.3:

- ***1st proof-of-impact***: As in the next-cell prediction analysis, we remove the exploration-like visits (cf. Fig. 2–IV.3.a).
- ***2nd proof-of-impact***: To avert size-related impacts, unlike in the previous prediction task, we do not replace exploration-like visits by adding stationary periods as it goes against the definition of the next-place formulation. Hence, given the last visited location "A" and the next visited one "C" if the current location "F" is assumed to be an exploration, we replace the exploration record "F" with the most frequent location that usually appears after "A" and different from "C" (which is "B" in Fig. 2–IV.3.b). The results are depicted in Figure 17.
- ***3rd proof-of-impact***: Slightly different from the 3rd proof of impacts of the previous prediction formulation, we replace exploration-like visits by a random symbol met in the



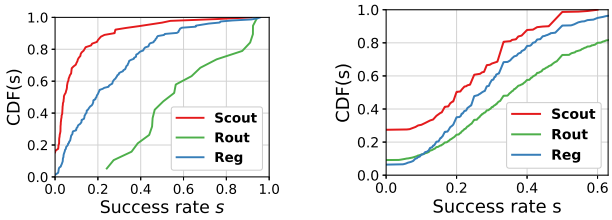
(a) Scout (b) Reg (c) Rout

Fig. 13: Distribution of the success rate score s_u of each predictor per profile for the Agg_gps dataset.



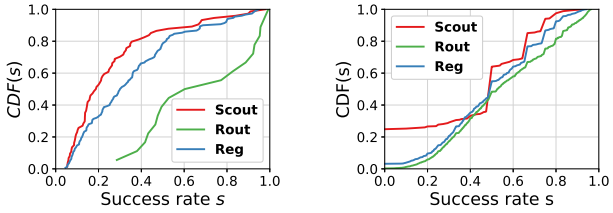
(a) Scout (b) Reg (c) Rout

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



(a) Agg_gps (b) ChineseDB

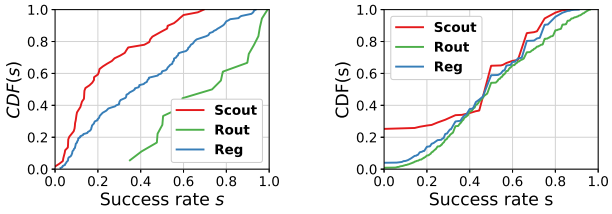
Fig. 15: Success rate score s_u of the MC predictor per profile.



(a) Agg_gps (b) ChineseDB

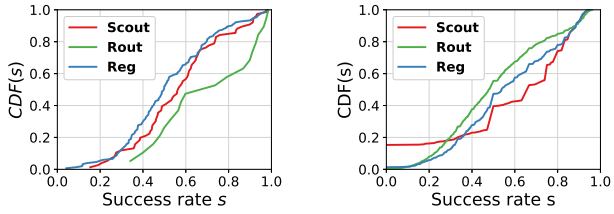
Fig. 18: Effect of novelty-seeking random replacement on prediction.

sequence that is different from the last and next visited locations (cf. Fig. 2–IV.3.c). Figure 11 shows the obtained results.



(a) Agg_gps (b) ChineseDB

Fig. 16: Effect of novelty-seeking removal on prediction.



(a) Agg_gps (b) ChineseDB

Fig. 17: Effect of novelty-seeking replacement on prediction.

Figure 16 displays the accuracy for the MC while keeping only familiar visits in the mobility traces. The prediction accuracy 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).

Figures 17 and 18 show that the replacement of novelty-seeking places improves the overall performance, with a further enhancement using the most probable known location substitution. This emphasizes the role of exploration in making individuals less predictable, particularly *Scouters*.

Further, we discern the substantial harmful effects of exploration visits on the predictability in the next-place prediction compared to the next-cell prediction. More importantly, the results show that the isolation of exploration more impacts *Scouters*. The original median accuracy for *Scouters* is approximately less than 20% (see Figure 15)), which is significantly lower than the performance of other profiles. Therefore, removing or replacing explorations events makes *Scouters* roughly as predictable as the other profiles.

Summarizing remarks: 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, in particular with *Scouters*. Understanding individuals' tendencies to explore can benefit next-place-based predictors. Indeed, quantifying and anticipating individuals' inclinations for novelty-seeking helps predictors to enhance their performance by looking at further contextual data or collective mobility behavior.

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, 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. This will help assess new novelty-seeking-related scaling laws per profile and provide more consistent and generative models. 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 to such individuals.

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. This is mainly due to the high stationarity present 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, 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 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. Although the widespread technological devices allow the collection of individuals' mobility traces, their acquisition is a nontrivial process and is getting more complex. Several sensitive professional and personal information can be inferred solely through mobility traces. Our future work can be divided into two directions: 1) investigate how our proposed profiling can be adapted 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 considers 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] H. S. Badr, H. Du, M. Marshall, E. Dong, M. M. Squire, and L. M. Gardner, "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] Lu Xin, Bengtsson Linus, and Holme Petter, "Predictability of population displacement after the 2010 haiti earthquake," *Proceedings of the National Academy of Sciences*, vol. 109, no. 29, pp. 11 576–11 581, 2012.

[3] L. Bengtsson, X. Lu, A. Thorson, R. Garfield, and J. von Schreeb, "Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: A post-earthquake geospatial study in haiti," *PLOS Medicine*, vol. 8, no. 8, pp. 1–9, 08 2011. [Online]. Available: https://doi.org/10.1371/journal.pmed.1001083

[4] L. Aalto, N. Göthlin, J. Korhonen, and T. Ojala, "Bluetooth and wap push based location-aware mobile advertising system," in *MobiSys '04: Proceedings of the 2nd international conference on Mobile systems, applications, and services*. New York, NY, USA: ACM Press, 2004, pp. 49–58.

[5] A. Nadembega, A. Hafid, and T. Taleb, "Mobility-prediction-aware bandwidth reservation scheme for mobile networks," *IEEE Transactions on Vehicular Technology*, 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, 2008.

[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, and Z. Q. Lee, "Predictability of individuals' mobility with high-resolution positioning data," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, ser. UbiComp '12. New York, NY, USA: Association for Computing Machinery, 2012, pp. 381–390.

[10] D. Teixeira, J. Almeida, and A. Carneiro Viana, "On estimating the predictability of human mobility: the role of routine," *EPJ Data Science*, vol. 10, no. 1, Sep. 2021.

[11] X. Lu, E. Wetter, N. Bharti, A. J. Tatem, and L. Bengtsson, "Approaching the limit of predictability in human mobility," *Scientific reports*, vol. 3, no. 1, pp. 1–9, 2013.

[12] H. Gao, J. Tang, and H. Liu, "Mobile location prediction in spatio-temporal context," in *Nokia mobile data challenge workshop*, vol. 41, no. 2. Citeseer, 2012, pp. 1–4.

[13] A. Cuttone, S. Lehmann, and M. C. González, "Understanding predictability and exploration in human mobility," *EPJ Data Science*, vol. 7, no. 1, p. 2, 2018.

[14] B. S. Jensen, J. E. Larsen, K. Jensen, J. Larsen, and L. K. Hansen, "Estimating human predictability from mobile sensor data," in *2010 IEEE International Workshop on Machine Learning for Signal Processing*, 2010, pp. 196–201.

[15] G. Smith, R. Wieser, J. Goulding, and D. Barrack, "A refined limit on the predictability of human mobility," in *2014 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, 2014, pp. 88–94.

[16] A. Y. Xue, R. Zhang, Y. Zheng, X. Xie, J. Huang, and Z. Xu, "Destination prediction by sub-trajectory synthesis and privacy protection against such prediction," in *Proceedings of the 2013 IEEE International Conference on Data Engineering (ICDE 2013)*, ser. ICDE '13. USA: IEEE Computer Society, 2013, p. 254–265.

[17] L. Pappalardo, F. Simini, S. Rinzivillo, D. Pedreschi, F. Giannotti, and A.-L. Barabási, "Returners and explorers dichotomy in human mobility," *Nature communications*, vol. 6, no. 1, pp. 1–8, 2015.

[18] 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.

[19] L. Amichi, A. C. Viana, M. Crovella, and A. A. Loureiro, "Understanding individuals' proclivity for novelty seeking," in *Proceedings of the 28th International Conference on Advances in Geographic Information Systems*, ser. SIGSPATIAL '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 314–324.

[20] H. Barbosa, M. Barthelemy, G. Ghoshal, C. R. James, M. Lenormand, T. Louail, R. Menezes, J. J. Ramasco, F. Simini, and M. Tomasini, "Human mobility: Models and applications," *Physics Reports*, vol. 734, p. 1–74, Mar 2018. [Online]. Available: http://dx.doi.org/10.1016/j.physrep.2018.01.001

[21] J. Wang, X. Kong, F. Xia, and L. Sun, "Urban human mobility: Data-driven modeling and prediction," *Acm Sigkdd Explorations Newsletter*, vol. 21, no. 1, pp. 1–19, 2019.

[22] E. L. Ikanovic and A. Mollgaard, "An alternative approach to the limits of predictability in human mobility," *EPJ Data Science*, vol. 6, no. 1, p. 12, 2017.
[23] K. Jaffrès-Runser, G. Jakllari, T. Peng, and V. Nitu, "Crowdsensing mobile content and context data: Lessons learned in the wild," in *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, 2017, pp. 311–315.
[24] S. B. Mokhtar, A. Boutet, L. Bouzouina, P. Bonnel, O. Brette, L. Brunie, M. Cunche, S. D'Alu, V. Primault, P. Raveneau *et al.*, "Priva'mov: Analysing human mobility through multi-sensor datasets," in *NetMob 2017*, 2017.
[25] Y. Zheng, Q. Li, Y. Chen, X. Xie, and W.-Y. Ma, "Understanding mobility based on gps data," in *Proceedings of the 10th international conference on Ubiquitous computing*, 2008, pp. 312–321.
[26] Y. Zheng, X. Xie, and W.-Y. Ma, "Geolife: A collaborative social networking service among user, location and trajectory," *IEEE Data(base) Engineering Bulletin*, June 2010.
[27] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma, "Mining interesting locations and travel sequences from gps trajectories," in *Proceedings of the 18th International Conference on World Wide Web*, ser. WWW '09. New York, NY, USA: Association for Computing Machinery, 2009, pp. 791–800.
[28] G. Chen, A. C. Viana, M. Fiore, and C. Sarraute, "Complete trajectory reconstruction from sparse mobile phone data," *EPJ Data Science*, vol. 8, pp. 1–24, 2019.
[29] 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.
[30] M. K. Cowles and B. P. Carlin, "Markov chain monte carlo convergence diagnostics: a comparative review," *Journal of the American Statistical Association*, vol. 91, no. 434, pp. 883–904, 1996.
[31] A. Moffat, "Implementing the ppm data compression scheme," *IEEE Transactions on Communications*, vol. 38, no. 11, pp. 1917–1921, 1990.
[32] P. Jacquet, W. Szpankowski, and I. Apostol, "A universal pattern matching predictor for mixing sources," in *Proceedings IEEE International Symposium on Information Theory,*, 2002, pp. 150–.
[33] K. Gopalratnam and D. J. Cook, "Active lezi: An incremental parsing algorithm for sequential prediction," *International Journal on Artificial Intelligence Tools*, vol. 13, no. 04, pp. 917–929, 2004.

ACKNOWLEDGEMENTS

We would like to thank the research agencies CAPES, CNPq, FAPEMIG, and FAPESP (grant 18/23064-8) and the support from INRIA, Sorbonne UPMC, LINCS, STIC AmSud LINT (code 22-STIC-07).

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 Sorbonne 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 Inria Senior Reseacher and TRiBE team's leader. She got her HDR (2011) and PhD (2005) degrees from UPMC, France. She was a visiting researcher at TKN/TU-Berlin, Germany (2009-2011). Her research interests are on human behavior understanding and mobility analytics with application in tactful, tactile, and smart networking. She is a recipient of the French Scientific Excellence award (2015-2022). She is author of more than one hundred papers on mobile networking. She has been involved in the organizing committee of major conferences (e.g. ACM SenSys, ACM Mobicom, IEEE Infocom, IEEE LCN).

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 LIP6-Paris, and in 2018-2019 at LIP6, INRIA Paris, and LINCS Paris. His research works to improve the understanding, design, and performance of 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) as well as over two hundred papers on networking and computer systems, and is a Fellow of the ACM and the IEEE.

Antonio A.F. Loureiro is a Professor in the Department of Computer Science at the Federal University of Minas Gerais (UFMG), Brazil. He received his PhD in Computer Science from The University of British Columbia, Canada. He was the recipient of the 2015 IEEE Ad Hoc and Sensor (AHSN) Technical Achievement Award. 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 has also presented keynotes and tutorials at international conferences.

APPENDIX

A. 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 19.

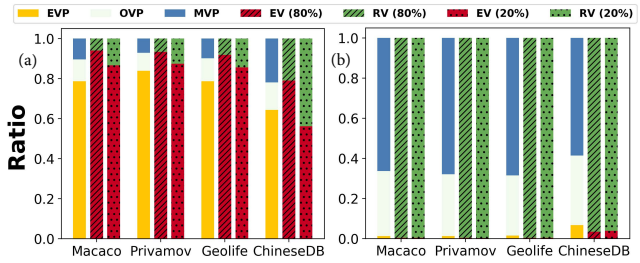


Fig. 19: (a-left) Percentage of visited places. (b-right) Average visitation frequency. EVP, OVP, and MVP are categorized according to Algo1. EV and RV are categorized according to Algo 2 for $level = 80\%$ and $level = 20\%$.

Figure 19 (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. 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 ∪ OVP– are not integrated into the daily routines of the individuals [2]. 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 $level = 20\%$. We notice when $level = 80\%$, the proportion of places classified as EV corresponds roughly to the percentage of places categorized as EVP ∪ OVP. 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.

Figure 19 (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. In 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. 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 only to 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 ∪ OVP.

[2]The CDR dataset describes visits in a smaller temporal resolution (i.e., per hour), this naturally impacts the precision in exploration inference of visits.

	EV (80%) ∈ EVP	EV (20%) ∈ EVP	EV (80%) ∈ EVP ∪ OVP	EV (20%) ∈ EVP ∪ 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 $level \in \{20, 80\}$.

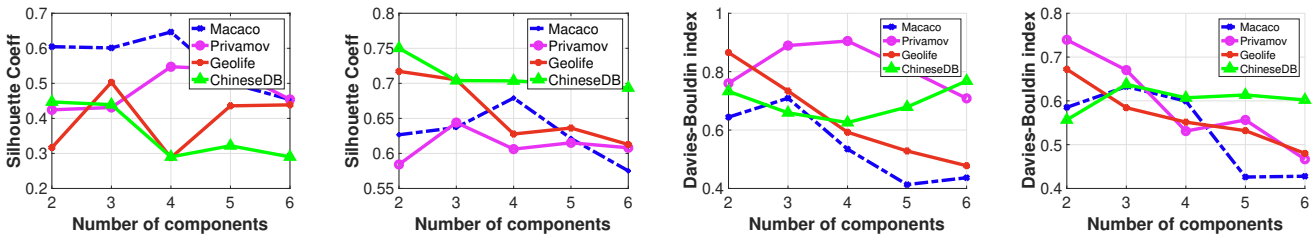
In Table II, we report the percentage of overlap between the locations categorized as EV with $level \in \{20, 80\}\%$ at first with EVP locations. Then, we quantify the similarity between EV and EVP ∪ OVP. We observe that the overlap between EV and EVP (EVP ∪ OVP) is higher when $level$ equals 80%. The degree of overlap between EVP ∪ OVP reaches up to 98.27% with the CDR dataset. Though the difference between our methodology and the baseline in quantifying the importance of a location, the resulting classifications are very similar.

From one side, setting $level$ to 80% allows EV to capture exceptionally and occasionally visited places as the baseline approach. From the other side, it allows capturing 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.

B. Clustering

Fig. 20 depicts the silhouette score and the Davies-Bouldin index obtained for the two clustering algorithms GMM Figs. 20a and 20c and k-mean Figs. 20b and 20d. Fig. 20a shows that the best achieved separability varies from a dataset to another. Though a clustering with three elements appears to be more equitable, as all datasets have a silhouette score above 0.4 and a Davies-Bouldin index above 0.58. Two, three, and four components are good candidates for the k-mean algorithm. Still, a clustering with three groups seems to be more balanced amid the datasets. Accordingly, we have two candidates for the best number of components. Nonetheless, we choose a clustering with three components as it maximizes the minimal score for both of the clustering algorithms, and appears to be more meaningful for all of our data sources taken together.



(a) GMM (b) k-mean (c) GMM (d) k-mean

Fig. 20: **(a)** Silhouette score for the GMM. **(b)** Silhouette score for the k-means. **(c)** Davies-Bouldin index for the GMM. **(d)** Davies-Bouldin index for k-means. .