An inherent limiting factor of human mobility prediction

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

An inherent limiting factor of human mobility prediction

Licia Amichi[†], Aline Viana Carneiro[⋆], Mark Crovella[‡], and Antonio Loureiro[§]

[†] Ecole Polytechnique (IPP), France

[⋆] INRIA Saclay, Bâtiment Alan Turing Campus de l'École Polytechnique, 91120 Palaiseau, 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 we humans move within space and time is becoming a central topic in many scientific domains, ranging from epidemic propagation, urban planning to ride-sharing. However, current works neglect individuals' preferences for exploration and discovery of new places. Yet, novelty-seeking activities appear to have significant consequences on the ability to understand and predict individuals' trajectories. In this work, we propose a new approach for the identification of moments of novelty-seeking. Subsequently, we construct individuals' mobility profiles based on their exploration inclinations – *Scouters* (i.e., extreme explorers), *Routiners* (i.e., extreme returners), and *Regulars* (i.e., without extreme behavior).

Index Terms

Individual Mobility, Exploration, Mobility Profiling,

I. Introduction

With the ubiquity of mobile devices appointed with internet connectivity and positioning systems ranging from vehicles equipped with GPS receivers, mobile phones to fitness bracelets, understanding and modeling human mobility became an accessible domain of study. Over the last decades, the collection of large amounts of human-mobility data and individuals' whereabouts urged scientists from different disciplines to study the dynamics of human mobility behavior and develop representative models and accurate predictors able to reproduce an individual's

trajectories and forecast his/her future locations. Indeed, accurate mobility models and predictors are crucial for epidemic prevention [1], disaster response [2, 3], improving the services provided by pervasive computing applications [4–6], providing energy-efficient and cost-effective network infrastructures [7, 8], or traffic management [9].

Previous studies have shown that individual mobility exhibit high temporal and spatial regular patterns characterized by few locations where users return frequently and predictably, interrupted by irregular sporadic visits to unknown or rarely visited places [10, 11]. But, ***to what extent is human mobility predictable?*** In this regard, several works have been conducted, either by measuring the theoretical upper bound (theoretical predictability) [2, 12, 13] or by computing the accuracy of prediction (practical predictability) [14–16] of the advanced developed predictive algorithms. Nevertheless, the empirical results suggested the predictability takes variable values ranging from under 40% to higher than 90% [16]. So, ***what are the origins behind this large variation in the predictability measures?*** Alternatively stated, ***what are the significant factors influencing the predictability?***

A non-negligible impacting factor is the tendency of individuals to explore and discover new places. Indeed, novelty-seeking is highly present in our daily lives, we are continuously hunting for new places and spots to go [16]. Moreover, the susceptibility to break the returning routine to explore and discover new places is heterogeneous among the populations, in this vein several profiling according to the proclivity to explore were disclosed [17], and [18]. This indicates that the novelty-seeking factor can be a critical factor and should not be overlooked for certain categories of individuals who present a high exploration activity. A noteworthy question essential for the development of optimal predictors is, ***to what degree do novelty-seeking activities obstruct the predictability of human mobility trajectories?***

This paper is an extended version of the earlier work published in the *Proceedings of the 28th International Conference on Advances in Geographic Information Systems (SIGSPATIAL '20)* [19]. We provide a more thorough understanding of the exploration phenomenon and propose a mobility profiling accordingly. The key contributions of this paper are:

- We propose a novel per-user approach for the classification of the visited locations: (i) **RV** places visited for regular and routine activities, and (ii) **EV** places visited when being carried by the tendency to explore. Next, we validate our proposal by a thorough experimental validation and a comparison of the performance with a state-of-the-art approach. Based on

this classification, we highlight moments of novelty-seeking of the individuals.

- We introduce a modeling approach that splits each individual visit into two categories: exploration – i.e., the discovery of new places – and return – i.e., the revisit of known locations. Then we define new metrics that capture individuals' propensity to explore new places and their intermittency – i.e., the shift between the two types of visits. Next, using our newly designed metrics we reveal the existence of three visiting profiles: Scouters, Routiners, and Regulars. For this, we use four urban datasets, describing people's mobility from 5 cities in 3 different continents around the world (Section III).

The remainder of this paper is organized as follows. We start with an overview of the related works 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, we present our new method for the identification of exploratory moments, based upon we introduce our newly developed metrics able to capture the propensity of individuals for novelty-seeking and propose an exploration-based mobility profiling and reveal the existence of three main profiles – Scouters, Regulars, and Routiners– in Section IV. Finally, we provide a discussion on the future research directions and open issues and challenges in Section V.

II. Related Works

Over the last decade, human mobility has been extensively scrutinized to understand the mechanisms ruling an individual's movements. Several works have demonstrated that human movements are far from being random and have a high degree of predictability [20].

The seminal paper of Song et al. [12] proposed an approach based on the entropic level of a mobility trace to measure the upper bound of its maximum predictability Π^{max}. Analyzing a three-month-long CDR dataset of 50,000 users, their study revealed that there is a 93% potential predictability in an individual's mobility trace. Several subsequent works tried to refine the predictability upper bound Π^{max}. For instance, Lu et al. [2] find that, on a CDR dataset containing the mobility trace of 2.9 million individuals, the upper limit of the 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, Bayesian network models, neural network algorithms, and so on. Lu et al. [14] 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 reaches 91%. Moreover, they showed that higher-order Markov Chain models do not significantly improve the practical predictability. Gao et al. [15] proposed and 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 works employing the same approach as in [12] attempted to dig out the significant factors that affect the predictability of human mobility, and shed light on origins of the limitations in predicting the next location:

Novelty-seeking: Recent studies have shown the importance of considering individuals' proclivity to explore new locations when modeling their mobility. [21]. Cuttone et al. [16] 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 predictive he/she is as it is impossible to forecast the unknown. This led to an important question being raised, *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] discerned two categories of people: explorers and returners. They based their classification on the number of regularly visited places, explores are those who visit many locations on a regular basis, whereas returners limit their mobility between few places.

Besides, Scherrer et al. [18] 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.

Moreover, in our previous work [19] 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: (i) *Scouters or extreme explorers:* whose proclivity for novelty-seeking is the most eminent all over the week and have a more spread spatial mobility; (ii) *Routiners or extreme returners:* who rarely perform explorations and have confined mobility; (iii) *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.

Spatial and temporal resolutions: Jensen et al. [22] 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. [13] used a high spatial and temporal resolution GPS dataset of 40 individual. The authors showed that their finer-grained dataset produces higher upper bounds with a predictability exceeding 98% with a temporal scale of 20 minutes or less. Smith et al. [23] and Teixeira et al. [24], showed that the predictability is correlated with the temporal resolution and have an inverse correlation with the spatial resolution.

Type of prediction: Ikanovic et al. [25] emphasized the origins of the high potential predictability of individuals' mobility obtained in earlier works [2, 12]. 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 an accuracy of approximately 71%. Thereby, they validated that the high estimated values of predictability stem from the stationarity rather than movements. Cuttone et al. [16] analyzed the predictability of a GPS dataset with the two widespread formulations of prediction, namely, the next-time step prediction and the next-place prediction. While the next-time step prediction is shown 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%.

Position of our work: While the impacts of prediction formulation on the upper-limit of predictability 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 that arouses from the intrinsic uncertain nature of human mobility, namely, individuals' propensity to explore and perform a mobility profiling accordingly. On the other hand, we investigate how the prediction formulation and the spatial

and temporal qualities of the used data can impact the predictability of each mobility profile.

III. Data Description

In this work, we use two categories of data sources to investigate individuals' proclivity for novelty-seeking; 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.

Dataset	Category	Number of users	Duration	Frequency of sampling
Macaco [26]	GPS	132	34 months	5 min
Privamov [27]	GPS	100	15 months	few seconds
Geolife [28–30]	GPS	182	64 months	1 to 5 seconds
ChineseDB*	CDR	642K	2 weeks	1 hour

*The collection was initiated by Shanghai University [31].

TABLE I: Datasets description.

A. GPS datasets

GPS technology allows tracking individuals' movements with the highest level of accuracy and temporal frequency. Hereafter, we describe our three GPS data sources.

Macaco: it consists of the anonymized digital activities tracks of 132 volunteers from 6 different countries collected by the MACACO project [26]. The project 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 timestamp. Each tuple has a unique ID, which relates to a specific user.

Privamov: it contains mobility traces collected in the Privamov sensing compaign [27], 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: our last GPS data source was collected in (Microsoft Research Asia) Geolife [28, 29]. 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. 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 call originated.

ChineseDB: this dataset is collected from 642K anonymized mobile phone subscribers in Shanghai, China [1], and contains 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, $\mathcal{H}_u = \langle (lon_0, lat_0, t_0), (lon_1, lat_1, t_0 + \delta), \dots (lon_N, lat_N, t_0 + N\delta) \rangle$. 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 [16], 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-, $\mathcal{T}_{u,c} = \langle (id_0, t_0), (id_1, t_0 + \delta), \dots (id_N, t_0 + N\delta) \rangle$. 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,

[1]The collection was initiated by Shanghai University [31].

$\delta = 5min$. 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. [31] 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 location*.
- Once *home* (id_h), *workplace A* (id_{wp_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) .

D. Experimental settings

In what follows, we give a brief description of the parameter settings we used in this study. 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 264 users: 82 in Macaco, 77 in Privamov, and 103 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 15 days of complete data, we are left with 4860 individuals.

We discretize locations to grid cells of size $c =$ 200m, 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, in this paper we focus on the discoveries of new places on a daily basis, for instance, going to a new restaurant or a new shop. Therefore, a cell of size 200m $\times$ 200m along with the imprecision and uncertainty of GPS systems, roughly

corresponds to daily regions of interest. 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. A resolution of 5 min for the GPS datasets allows uniforming the frequency of sampling between the different sources while increasing the number of individuals and being reasonable for capturing most transitions. Moreover, having different datasets with the same resolutions allows us to test the effectiveness of our methods and to extensively validate our work.

IV. Profiling methodology

There exists a perplexity in understanding and predicting individuals' mobility patterns. Human beings' movements are a mixture of repetitive and regular transitions between known places and sporadic discoveries of new areas [11, 17, 32], both subject to a certain degree of uncertainty associated with free will and arbitrariness [33]. At each instant, an individual is confronted with an extensive list of choices with regard to *how* and *where* to spend his/her time, and has two alternatives: he/she either returns to a place he/she visited in the past or explores a new location. Here, *we intend to investigate whether there exist patterns when commuting from an exploration mode to a return mode and vice versa*. For this, we divide human movements into two primary states: *explorations and returns*. We define (i) the **exploration** as *a discovery of a new location*, i.e., a visit to a location that is not present in the visiting history of an individual and (ii) a **return** as *a visit to a previously seen locality*.

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**). Initially the individual u is in the exploring state (**E**) if his/her current location id_{t_0} is not present in the set of his/her known places $\mathcal{L}_u(t)$ at $t = t_0$, i.e. $id_{t_0} \notin \mathcal{L}_u(t)$ and in the returning state (**R**) otherwise. Two possible inputs 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, gives M an input and

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.

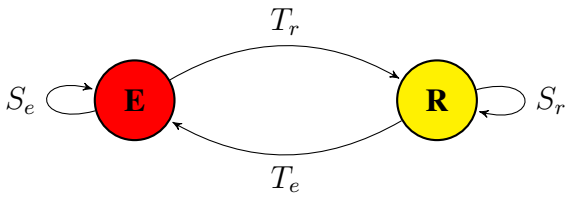


Fig. 1: Finite-State Automaton M.

B. Novelty-seeking identification

Strictly speaking, an exploration is a discovery of a new geographical location, i.e., a place where the concerned individual was never seen before. Given the mobility trace of an individual u, *how can we distinguish his/her novelty-seeking visits from his/her routine visits?* To this end we adopt two methods to characterize and classify the visited locations at the individual level:

1) Visitation-frequency-based identification: let $F_u = \{id_1, id_2, \ldots, id_n\}$ be the set of location visited by the user u. First, for each location $id_i \in F_u$, we assign a weight w, given by,

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

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 $\overline{w}_u = \frac{1}{|F|} \times \sum_{i=1}^{|F|} w_u(id_i)$. Following, we categorize the visited locations into locations used for: (i)Exploratory Visits (**EV**), (ii) 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}, otherwise it is assigned to the list of places used for **EV**, T_{EV} (see Algorithm 1).

2) Baseline identification: we compute the Relevance R_u of the location id_i visited by the user u as proposed in [34],

$$R_u(id_i) = \frac{d_{visit}(id_i, u)}{d_{total}(u)}, \tag{2}$$

Algorithm 1 Novelty-seeking identification A

function *location_classification_a ($\mathcal{T}_{u,c}$, $level$)*
$w_u, T_{RV_u}, T_{EV_u} \leftarrow \emptyset$
$F_u \leftarrow$ UNIQUE($\mathcal{T}_{u,c}$)
for j **in** F_u **do**
    $w_u[j] \leftarrow$ FREQUENCY_OF_APPEARANCE($j, \mathcal{T}_{u,c}$), (1)
end for
$\overline{w}_u \leftarrow$ MEAN(w_u)
for j **in** F_u **do**
    if $w_u[j] \geq \overline{w}_u \times level$ **then**
        T_{RV_u}.ADD(j)
    else
        T_{EV_u}.ADD(j)
    end if
end for
return T_{RV_u}, T_{EV_u}
end function

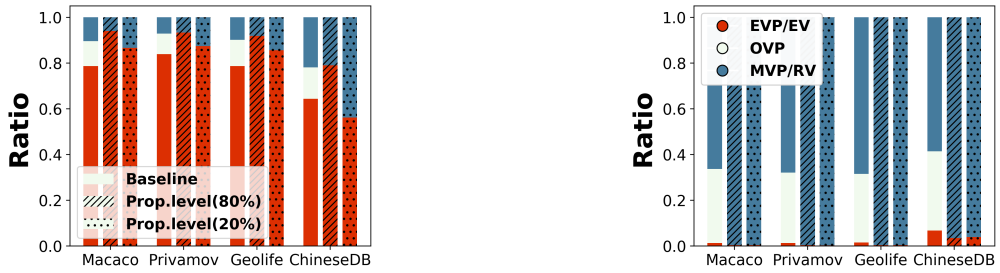
where $d_{visit}(id_i, u)$ is the number of days the individual u visited the location id_i, and $d_{total}(u)$ is the number of days the individual has been active. Following, as in [34] we use the k-mean unsupervised approach to classify the location into: (i) Mostly Visited Places (**MVP**), i.e, locations most frequently visited by the user; (ii) Occasionally Visited Places (**OVP**), i.e, locations of interest for the user, but visited just occasionally; (iii) Exceptionally Visited Places (**EVP**), i.e, rarely visited locations (see Algorithm 2).

For each individual u of our datasets, we classify his/her visited locations into **EV** or **RV** using our proposed Algorithm 1 at first with $level = 80\%$, then with $level = 20\%$. Following we use Algorithms 2 for the categorisation of the visited places into **EVP**, **OVP**, and **MVP**. Next, we compute the percentage of the places within each category (see Figure 2a). Afterward, we evaluate the average visitation frequency in each group as shown in Figure 2b.

Figure 2a reports the percentages of places classified in each category; **EV** and **RV** by Algorithm 1; **EVP**, **OVP**, and **MVP** by Algorithms 2. First, we observe the high ratio of **EVP** and **OVP** categorized by Algorithms 2, more than 78% of the places are not integrated in the

Algorithm 2 Novelty-seeking identification B

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



(a) Percentage of visited places

(b) Average visitation frequency

Fig. 2: As **EV** or **RV** according to our proposed algorithm with $level = 80\%$ and $level = 20\%$, and **EVP**, **OVP**, and **MVP** according to the Baseline algorithm (the legends are common for both Figures).

daily routines of the individuals. Likewise the proportion of locations used for **EV**s surpasses 78% while $level$ is set to 80%, and is higher that 60% with $level = 20\%$. Moreover, we can notice in the case where $level = 80\%$, the proportion of places classified as **EV** by Algorithm 1 corresponds roughly to the percentage of places categorized as **EVP** and **OVP** by Algorithms 2. In contrast, Algorithm 1 with $level = 20\%$ captures almost the same faction of **EV** as the number of locations classified as **EVP**. This indicates that setting $level$ to 80% may lead to an overestimation of the locations used for **EV**, while 20% seems a more reasonable setting.

Figure 2b illustrates the proportion of the average frequency of visits to each category of places (**EV** and **RV** by Algorithm 1; **EVP**, **OVP**, and **MVP** for Algorithms 2). Firstly, we see the markedly high proportion of visits to location used for **RV**, more than 90% of the visits

are towrds this category of places for $level \in \{20, 80\}\%$. Whereas the same score is obtained by Algorithms 2 while taking **MVP** and **OVP** together. Additionally, the average frequency of visits hold by **EV** for all datasets is lower than the scores obtained by **EVP**. Contrary to our assumption above, a setting with $leve = 80\%$ is not an over estimation of the locations used for **EV**, but it allows a more thorough capture of visits enhanced by the proclivity to explore of the individuals.

In summary, the proposed method Algorithm 1 allows a more satisfactory classification of the visited places compared to the baseline Algorithms 2. On the one hand, it allows the detection of a higher number of places used for **EV**, on the other hand, it guarantees that the visitation frequencies to these locations are notably lower compared to the **RV** as well as **EVP** of Algorithms 2. Withal, the performance of Algorithm 1 with $level = 80\%$ allows the identification of a higher number of places used for **EV**, and hence enables a better detection of moments of exploration compared to the setting with $level = 20\%$. Indeed, the first occurrence of a location present int the T_{EV_u} of a user u in his/her mobility trace is presumed to be a moment of exploration. In the remaining of the paper, we use Algorithm 1 and set $level$ to 80% for the categorization of the visited locations at the cost of sometimes overestimating moments of novelty-seeking.

C. Mobility Profiling

Initially, each user u has an empty set of known locations $\mathcal{L}_u(t = t_0) = \emptyset$. Using Algorithm 1 with $level$ set to 80%, for each user u we classify his/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 occurence of a location present the set of known locations $\mathcal{L}_u$ is a return, else it is an exploration. Note that after the discovery of new place, this latter is added to $\mathcal{L}_u$, i.e., its next occurence will be viewed as a return.

After dissecting human transitions into explorations and returns, we first extract two sets:

- **Returning set** ret_u**:** is a set containing the sets of consecutive returns

$$ret_u = \{r_0, r_1, \ldots, r_n\}, \tag{3}$$

  where each $r_i = \{id_0, id_1, \ldots, id_x\}$ is a set containing the ids of the cells where the individual u performed successive returns.

- **Exploring set** exp_u**:** is a set containing the sets of consecutive explorations

$$exp_u = \{e_0, e_1, \ldots, e_n\}, \tag{4}$$

where each $e_i = \{id_0, id_1, \ldots, id_x\}$ is a set containing the ids of the cells where the individual 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 his/her successive explorations– i.e., the average number of consecutive self-transitions he/she made in the E state, and (2) $\#R = avg(|r_i|), r_i \in ret_u$ the average number of self-transitions he/she made in the R state.

To characterize how individuals balance the trade-off between revisits of familiar locations and discoveries of new places, 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.

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. Namely, it helps to recognize if a user is constantly fluctuating between visits to familiar places and discoveries of new spots or once he/she starts a discovery he/she does is it repeatedly, before switching to revisits and vice versa. The second metric captures users' proclivity to make revisits rather than explore new places.

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 his/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 he/she visits many new places on a row, or just after a few discoveries he/she goes back to a familiar location.

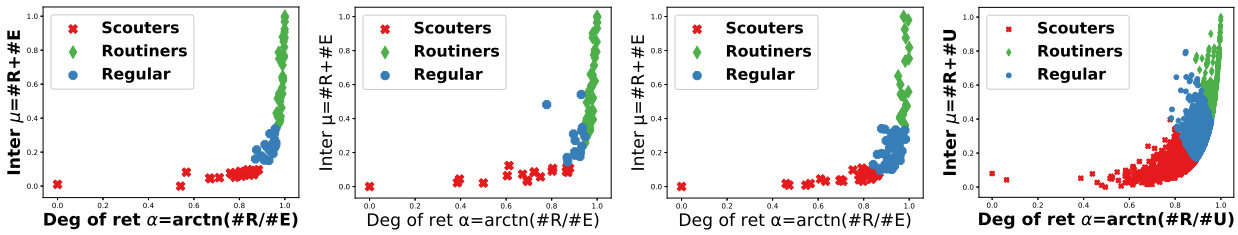
In what follows, we investigate whether the novelty-seeking habit is the same among the

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. 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 prob 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 and 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 [19]).

We now apply, the GMM and k-mean with three components on our data sources, we roughly obtain the same groups. Henceforth, hereafter 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 to our GPS and CDR data sets. 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 low compared to the others'. 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.

Our metrics allow a natural clustering of individuals. Although, having a different number of frequently visited locations, individuals who usually break their routines to explore are viewed as *Scouters*. This is unlike in the method suggested by Pappalardo et al. [17], where some individuals can be wrongly clustered as explorers or as returners. Contrary to [18] our 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*, as well as accordingly splits the populations.

V. 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.

REFERENCES

[1] V. Belik, T. Geisel, and D. Brockmann, "Natural human mobility patterns and spatial spread of infectious diseases," *Phys. Rev. X*, vol. 1, p. 011001, Aug 2011. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevX.1.011001

[2] X. Lu, L. Bengtsson, and P. Holme, "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. [Online]. Available: https://www.pnas.org/content/109/29/11576

[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] D. Quercia, N. Lathia, F. Calabrese, G. Di Lorenzo, and J. Crowcroft, "Recommending social events from mobile phone location data," in *2010 IEEE International Conference on Data Mining*, 2010, pp. 971–976.

[5] 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. [Online]. Available: http://dx.doi.org/10.1145/990064.990073

[6] D. Karamshuk, A. Noulas, S. Scellato, V. Nicosia, and C. Mascolo, "Geo-spotting: Mining online location-based services for optimal retail store placement," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '13. New York, NY, USA: Association for Computing Machinery, 2013, p. 793–801. [Online]. Available: https://doi.org/10.1145/2487575.2487616

[7] 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.

[8] F. De Rango, P. Fazio, and S. Marano, "Utility-based predictive services for adaptive wireless networks with mobile hosts," *IEEE Transactions on Vehicular Technology*, vol. 58, no. 3, pp. 1415–1428, 2009.

[9] Çolak, Serdar and Lima, Antonio and González, Marta C., "Understanding congested travel in urban areas ," *Nature Communications*, vol. 7, no. 1, p. 10793, 2016.

[10] D. Brockmann, L. Hunfnagel, and T. Geisel, "The scaling laws of human travel," *Nature*, vol. 439, pp. 462–465, Jan. 2006.

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

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

[13] Miao Lin,Wen-Jing Hsu, Zhuo Qi Lee, "Predictability of individuals' mobility with high-resolution positioning data," in *UbiComp '12: Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, Sep. 2012.

[14] Xin Lu, Erik Wetter, Nita Bharti, Andrew J. Tatem and Linus Bengtsson , "Approaching the Limit of Predictability in Human Mobility," *Scientific Reports*, vol. 3, no. 2923.

[15] H. Gao, J. Tang, and H. Liu, "Mobile location prediction in spatio-temporal context," 2012.

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

[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. 8166, Sep 2015.

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

[19] L. Amichi, A. C. Viana, M. Crovella, and A. A. Loureiro, "Understanding individuals' proclivity for novelty seeking," ser. SIGSPATIAL '20. New York, NY, USA: Association for Computing Machinery, 2020. [Online]. Available: https://doi.org/10.1145/3397536.3422248

[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, pp. 1 – 74, 2018, human mobility: Models and applications. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S037015731830022X

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

[22] Bjørn Sand Jensen, Jakob Eg Larsen, K. Jensen, J. Larsen, and Lars Kai Hansen, "Estimating human predictability from mobile sensor data," in *2010 IEEE International Workshop on Machine Learning for Signal Processing*, 2010, pp. 196–201.

[23] 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.

[24] D. de C. Teixeira, A. C. Viana, M. S. Alvim, J. M. Almeida, "Deciphering predictability limits in human mobility," in *ACM SIGSPATIAL*, Nov. 2019.

[25] Edin Lind Ikanovic and Anders Mollgaard, "An alternative approach to the limits of predictability in human mobility," *EPJ Data Science*, vol. 6, no. 1, 2017.

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

[27] 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*, Apr. 2017.

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

[29] 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.

[30] X. X. W. M. Y. Zheng, L. Zhang, "Mining interesting locations and travel sequences from gps trajectories," in *In Proceedings of International conference on World Wild Web, Madrid Spain.*, 2009, pp. 791–800.

[31] G. Chen, A. Carneiro Viana, M. Fiore, and C. Sarraute, "Complete Trajectory Reconstruction from Sparse Mobile Phone Data," *EPJ Data Science*, Oct. 2019.

[32] C. M. Schneider and V. Belik and T. Couronné and Z. Smoreda and M. C. González, "Unravelling daily human mobility motifs," *J R SOC INTERFACE*, vol. 10, no. 20130246, Jul. 2013.

[33] L. Alessandretti, P. Sapiezynski, V. Sekara, S. Lehmann and A. Baronchelli , "Evidence for a conserved quantity in human mobility," *Nature Human Behaviour volume*, vol. 2, pp. 485–491, May 2018.

[34] Michela Papandrea, Karim Keramat Jahromi, Matteo Zignani, Sabrina Gaito, Silvia Giordano, Gian Paolo Rossi, "On the properties of human mobility," *Computer Communications*, vol. 87, no. 1, pp. 19–36, Aug. 2016.