Understanding individuals' proclivity for novelty seeking

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

Understanding individuals' proclivity for *novelty seeking*

Licia Amichi[†⋆], Aline Carneiro Viana[⋆], Mark Crovella[‡], and Antonio A.F. Loureiro [§]
[†] Ecole Polytechnique (IPP), [⋆] Inria, [‡] Boston University, [§] Federal University of Minas Gerais

Email: licia.amichi@inria.fr, aline.viana@inria.fr, crovella@bu.edu, loureiro@dcc.ufmg.br

***Abstract*—Human mobility literature is limited in their ability to capture the novelty-seeking or the exploratory tendency of individuals. Mainly, the vast majority of mobility prediction models rely uniquely on the history of visited locations (as captured in the input dataset) to predict future visits. This hinders the prediction of new unseen places and reduces prediction accuracy. In this paper, we show that a two-dimensional modeling of human mobility, which explicitly captures both regular and exploratory behaviors, yields a powerful characterization of users. Using such model, we identify the existence of three distinct mobility profiles with regard to the exploration phenomenon – *Scouters* (i.e., extreme explorers), *Routiners* (i.e., extreme returners), and *Regulars* (i.e., without extreme behavior). Further, we extract and analyze the mobility traits specific to each profile. We then investigate temporal and spatial patterns in each mobility profile and show the presence of recurrent visiting behavior of individuals even in their novelty-seeking moments. Our results unveil important novelty preferences of people, which are ignored by literature prediction models. Finally, we show that prediction accuracy is dramatically affected by exploration moments of individuals. We then discuss how our profiling methodology could be leveraged to improve prediction.**

***Index Terms*—Individual Mobility, Exploration, Mobility Profiling**

I. Introduction

Understanding human mobility and accurately predicting an individual's next location spans several disciplines, such as urban planning, public health, traffic management, and environmental management [1], [2]. In this context, human mobility can be studied at the individual level (i.e, individual mobility) or the group level (i.e, population flows). In this paper, *we focus on individual mobility* that is more exposed to irregular movements and routines rupturing effects, unlike population flows that consist of the aggregated mobility of individuals, and hence, uncertainties are less observable and have fewer impacts.

As an attempt to understand individual mobility dynamics, several models were developed [3], [4]. However, these models systematically fail in reproducing individuals' movements and substantially deviate from empirical results [3], [1]. Moreover, many prediction models have been proposed to forecast individuals' movements. Yet, they all show limited bounded predictive performance [5]. Regardless of the applied methods (e.g., Markov chains, Naive Bayes, neural networks), the type of prediction (i.e., next-cell or next place) or the used data sets (e.g., GPS, CDR, surveys), the accuracy of prediction never reaches the coveted 100%. The reasons are manyfold: the lack of ground truth data, human beings' complex nature and behavior, and the difficulty to forecast visits to non-routinary areas and discoveries of new places [6].

We focus on the exploration problem – i.e, the new-place discovery's tendency of individuals – that has rarely been tackled in the literature. We confirm such a problem represents a real issue and should be carefully addressed to propose realistic generative models and accurate predictors [5] (Section II). Most models addressing the exploration phenomenon assume it to be unfluctuating among the population. Besides, most existing predictors endeavor to forecast future locations from the set of known places only, which hinders predicting new unseen places and by consequence, reduces the predictive performance [5]. Fig. 1 emphasizes the harmful effects of explorations on the predictive performance of the classical Markov Chain predictor, when considering a CDR dataset (see Table III) with and without explorations (places visited only once). As shown, prediction using the no-exploration CDR trace achieves an average success prediction rate of 97%, which is approximately 24% higher than the total trace's score.

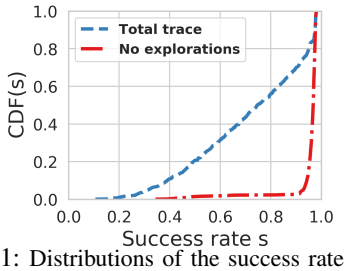


Fig. 1: Distributions of the success rate score

In this paper, we provide a better understanding of the exploration phenomenon and answer the following questions: *Is the proclivity for exploring new areas unfluctuating among the population? Is there any spatiotemporal pattern on the way people seek for novelty?* We propose to tackle the aforementioned questions by employing a novel approach to profile individuals and investigate their mobility traits according to their novelty-seeking tendency. In particular, our contributions are the following:

- We introduce a modeling approach that splits each individual visit into two states: *exploration* – i.e., the discovery of new places – and *return* – i.e., the revisit of known locations.

- We then define new metrics that capture the spatiotemporal properties of each individual visit – i.e., known/new and recurrent/intermittent visits. As such, we capture individuals' propensity to explore new places and their intermittency – i.e., the shift between the two types of visits.
- Using our newly designed metrics and the probabilistic Gaussian Mixture Model, we reveal the existence of three *visiting profiles*: *Scouters*, *Routiners*, and *Regulars* (Section IV). For this, we use four urban datasets, describing people mobility from 5 cities in 3 different continents around the world (Section III).
- We investigate the profiles according to 15 mobility features, providing a precise view of the mobility traits of each profile (Section V). Our analysis reveals that *Scouters* (i) are keener to explore and have larger sets of visited places, (ii) limit their routinary mobility to a small set of places, and (iii) walk longer distances. *Routiners* (i) rarely break their returning routine to discover new places, (ii) constantly visit their known locations, and (iii) have confined mobility. While *Regulars* have a medium behavior.
- We go deeper in our investigation by reporting our visiting profiles to the temporal and spatial use of the individuals (Section VI). We reveal that *Scouters*' proclivity for novelty-seeking is the most eminent all over the week and have a more spread spatial mobility. *Routiners* instead, rarely perform explorations and have confined mobility. More importantly, we show that, independently of the individual profile, spatiotemporal patterns can be clearly identified even during exploration moments.
- We show novelty-seeking effects on the predictive performance of two classical predictors (Section VII). We then discuss the benefits of our modeling and how it can be exploited to better capture individuals' dynamics and improve prediction accuracy.

Finally, we draw conclusions and comment on the perspectives of our work in Section VIII. Although a very short description of our profiling approach has been previously presented at the Student Workshop of ACM Conext 2019 [7], here we go deeper in our investigations by uncovering the mobility features and visiting patterns behind each profile, and the profiles' potential utility. For the best of our knowledge, we are the first to reveal spatiotemporal preferences present in exploration moments of people.

II. Related Work

Recent studies have shown the importance of distinguishing the exploration phenomenon from the revisits and revealed the existence of distinct classes of individuals in terms of mobility movements. Song et al. [3] have demonstrated that by considering the notions explorations and returns while analyzing the human movements helped to explain and to justify the origins of the scaling laws suggested by individual explanatory models (random walks, Lévy flight). Further, they proposed a more consistent statistical model of individual human mobility.

Following the work proposed by Song et al. [3], Pappalardo et al. [1] endeavored to explain the conflicting coexistence of heterogeneity and predictability characterizing human mobility by quantifying the impact of recurrent movements on the overall mobility. The authors reported the existence of two distinct mobility profiles: explorers and returners. Explorers are individuals who visit many spots on regular bases, whereas returners curb their mobility between few places. Besides, Pappalardo et al. [1] assumed that the probability of exploring new areas is correlated with the number of frequently visited places. Further, the authors adjusted the model proposed by Song et al. [3]. They suggested that an individual is attracted by popular locations at the group level when she discovers new places. And showed that their proposed model is more realistic when modeling human mobility. Nevertheless, this classification can be inconsistent; for instance, a person who regularly visits two different locations and usually explores many new areas is considered to be a returner, while a person who spends most of her time between eight different locations and rarely visits new ones can be viewed as an explorer.

Similar to Pappalardo et al. [1], Scherrer et al. [8] proposed a novel unsupervised mobility profiling approach. Their strategy showed the existence of two main classes of individuals: (i) travelers, who move around extensively, and (ii) locals, who move in a more constrained area and revisit many of their locations. Nevertheless, they do not bring any understanding of the exploration behavior of individuals. Although their approach does not classify all individuals and results in five groups of individuals, only two groups were interpreted and considered to be significant.

Contradicting the studies performed by Pappalardo et al. [1] and Scherrer et al. [8], Quadri et al. [9] assumed that given the significant number of visits to new places, all individuals are explorers. They showed that individuals' propensity to explore new areas is aroused by specific types of activities mainly shopping in particular fashion and clothing stores and usually happens during leisure time in distant areas far from frequently visited places.

Cuttone et al. [5] highlighted the importance of considering the exploration phenomenon when designing mobility predictors, showing it is a crucial factor behind the low accuracy of prediction models. Further, they proposed an exploration prediction model based on random guessing. Still, this model suggests that all individuals have the same probability to explore, which contradicts the studies performed by Pappalardo et al. [1] and Scherrer et al. [8].

The work in [10] is concerned with irregular activities, which differ from the explorations studied in our work in that they focus on the semantics of activities (eg, going to a gym) rather than the location alone.

Summarized remarks: The literature on human exploration is generally quite recent, and very few studies have investigated two basic types of human motions: exploration and returns (or

regular visits). Although some interesting observations have emerged, the few existing studies have several limitations [9], [5]. In particular, essential questions such as how to define a period of exploration, or how to identify exploration in an individual's movements, remain unsolved. Furthermore, in contrast to those assumptions [9], [5], one can observe a remarkable heterogeneity in human mobility behavior. Indeed, adopting such generalizations at the population level can be misleading while studying individual mobility. In summary, *understanding the exploration aspect of individual mobility is still in its infancy and deserves a deeper investigation*.

III. Data Description

In this section, we outline the data sources we leveraged in this study. We used four datasets capturing the spatio-temporal footprints of individuals' mobility with high spatial and temporal resolutions, three GPS data sources, and one CDR dataset. This last is collected by a major network operator in China from Shanghai, where each location represents the user's centroid of an hour with the precision of 200 meters. Our datasets are described in Table I.

Dataset	Type	Number of users	Duration	Sampling frequency
Macaco [11]	GPS	132	34 months	5 min
Privamov [12]	GPS	100	15 months	few seconds
Geolife [13], [14], [15]	GPS	182	64 months	1 to 5 seconds
ChineseDB*	CDR	642K	2 weeks	1 hour

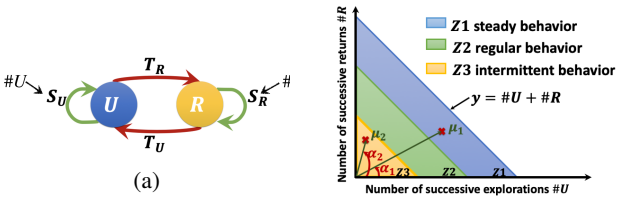
*The collection was initiated by Shanghai University [16].

TABLE I: Datasets description.

A. Data handling

For this study, we focus on the location data i.e. latitude and longitude. First, we reconstruct the mobility traces of the individuals by extracting the sequence of recorded locations along with the associated timestamps. Next, we discretize the geographical maps by placing uniform grids of c meters x c meters and draw out the grid cell IDs associated with the coordinates, by converting the tuple (lat, lon) into a tuple $(\lfloor \frac{lon}{c} \rfloor, \lfloor \frac{lat}{c} \rfloor)$ as in [5], where c meters is the cell-size in the grid. Hence, individuals' location history is converted into sequences of discrete symbols. 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. 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. [16] and complete them as follows,

- First, per individual, we identify the most frequent daily location and name it as *work location*. Intuitively it is the place that she usually visits and spends a large amount of time in it between 10 am and 11 am, and from 2 pm to 5 pm.



(a) (b)

Fig. 2: (a) Finite-State Automaton. (b) Avg successive visits.

- Next, we determine the most visited location by a user between 2 am and 6 am (night), which we refer to as *home location*.
- Once the home and work locations are identified, if a record is missing between 10 am and 11 am or from 2 pm and 5 pm, we add a new record with the grid ID associated to the workplace. If a record is missing from 2 am to 6 am, then a record is added with the grid ID associated to the home location.

B. Experimental settings

In what follows, we give a brief description of the datasets and the parameter settings we used in this study. We define a complete day for GPS datasets as a day in which an individual has a record at least each 15 min. And select only participants that have at least 10 complete days of data. We are left with 87 individuals in the Macaco database, 69 individuals in the Privamov dataset, and 101 in Geolife. For the CDR data, given the low frequency of sampling, we define a complete day as a day with at least one record every 2 hours, we are left with 3761 individuals.

We discretize locations to grid cells of size $c = 300$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, 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 300m $\times$ 300m 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. Proposed 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 [17], [1], [2],

both subject to a certain degree of uncertainty associated with free will and arbitrariness [18]. At each instant, an individual is confronted with an extensive list of choices with regard to *how* and *where* to spend her time, and has two alternatives: she either returns to a place 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. 2a, with two possible states: *exploring* (**U**) and *returning* (**R**). Initially the individual i is in the exploring state (**U**) if her current location $loc_i(t_0)$ is not present in the set of her known places $\mathcal{L}_i(t)$ at $t = t_0$, i.e. $loc_i(t_0) \notin \mathcal{L}_i(t_0)$ 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_u or S_u). In the exploring state **U**, discovering new areas (S_u) has no effect and keeps the individual in the state **U**. On the other hand, moving back to a known location (T_r), though recently explored, gives M an input and shifts the state from **U** 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_u), shifts the state back to the **U** state.

B. Mobility Profiling

Initially, all individuals have an empty set of visited locations $L_i(t = t_0) = \emptyset$. While analyzing an individual's mobility trace, we first identify the places she regularly visits, then, add them to her set of visited locations. Accordingly, the cold start problem is bypassed, alternatively stated, the first occurrences of familiar places in the trace of an individual are not considered as explorations. To this end, we examine the whole mobility trace of each individual and compute the visitation frequency of each location, let l_{max} be the place with the highest visitation frequency. Afterward, all locations that have a visitation frequency at least equal to 90% of the visitation frequency of l_{max} are added to her set of known places. After dissecting human transitions into explorations and returns, we assign to each individual two values: (1) $\#U$ the average number of her successive explorations– i.e., the average number of consecutive self-transitions she made in the U state, and (2) $\#R$ the average number of self-transitions she made in the R state.

Fig. 2b reports the average number of successive returns $\#R$ against the average number of successive explorations $\#U$ of the individuals. Intuitively, if we compare individual 1 (α_1, μ_1) and individual 2 (α_2, μ_2), we can state that the individual 1 spends more time exploring than the individual 2. Besides, the individual 2 performs more shifting between the U and R states than the individual 1 who is stationary with regard to the types of her visits (exploration or return). Hence, to characterize how individuals balance the tradeoff 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.

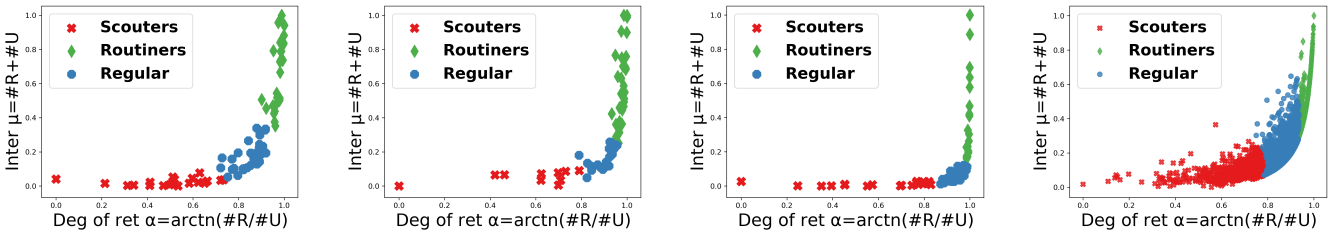
Definition 1 (***Intermittency*** μ)**.** *is the sum of the average number of successive explorations $\#U$ and the average number of successive returns $\#R$, $\mu = \#R + \#U$.*

When the average number of returns or explorations increases, the *intermittency* increases, indicating that fewer shifts occur between the exploring and returning states. Therefore, the intermittency 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 she starts a discovery she does is it repeatedly, before switching to revisits and vice versa. The second metric captures users' proclivity to make a 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 **U**,* $\alpha = \operatorname{arctg}\left(\frac{\#R}{\#U}\right)$.

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

Discovering similar users with regard to their mobility patterns has been broadly studied to address the issue of sparse mobility behavior among the population [19], [1], [8]. In what follows, we investigate whether the exploration 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 run one hundred fits for five different sets of clusters (two to six). Then, we consider the mean value when choosing the best score (For details see Appendix A). We choose a



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

Fig. 3: Mobility Profiling.

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 datasets.

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*'.

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. [1], where some individuals can be wrongly clustered as explorers or as returners. Contrary to [8] 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. Mobility Traits of Profiles

Here, we identify the specific mobility behavior traits of each profile: *Scouters*, *Routiners*, *Regulars*. Hence, we extract some of the fundamental features used to characterize human mobility from the spatiotemporal footprints of the individuals (see Table II). The derived features are divided into three groups: *Relocation Activities*, *Temporal Activities*, and *Spatial Activities*.

- *Relocation Activities*, this category aims at quantifying and characterizing individuals' visits, transitions habits, and capturing uniqueness and repetitiveness of visits.
- *Temporal Activities*, this category relates to the behavior of individuals in time and captures the amount of time spent by individuals exploring, returning, and visiting distinct locations.
- *Spatial Activities*, the last category gives an intuition on the distances walked by individuals when performing each type of visit and the covered distances.

In what follows, due to the small number of individuals in each mobility profile for the GPS data sources, and considering that the different GPS datasets are of the same nature with the same frequency of sampling and duration of analyses (10 days). We aggregate the mobility traces of individuals of the same mobility profile to perform a global characterization of each profile as well as a global comparison between them. We label this new dataset as *Agg_gps*. In view of its different nature, we separately analyze the profiles resulting from the CDR dataset.

For the sake of comparing and displaying the variations of the different features among individuals of each mobility profile, we report the box-plot[1] of each feature for *Scouters*, *Routiners*, and *Regulars* as shown in Figs 4, 5, 6, 7, 8,and 9.

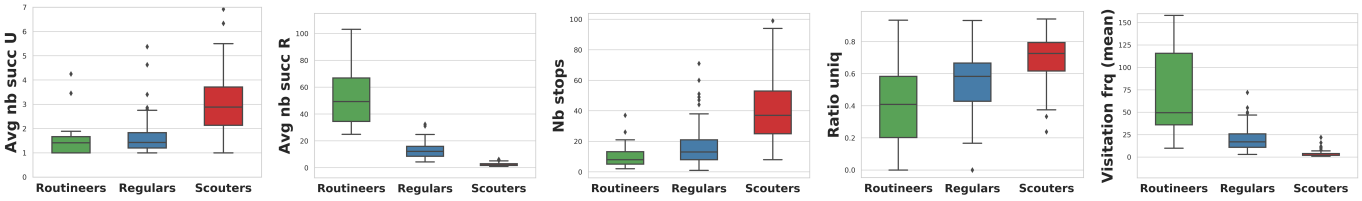
A. Scouters' mobility traits

Scouters are energetic, and dynamic when discovering new places. However, they become weary and flat while revisiting various areas they already know. Admittedly, when *Scouters* start exploring, they relish discovering many new other places

[1]Some overlaps between the box-plots of the different groups can be noticed, yet this is essentially due to the limitation in the number of users. Though, the tendency is clearly discernible among the mobility profiles, especially in the CDR figures where we leverage a larger number of users.

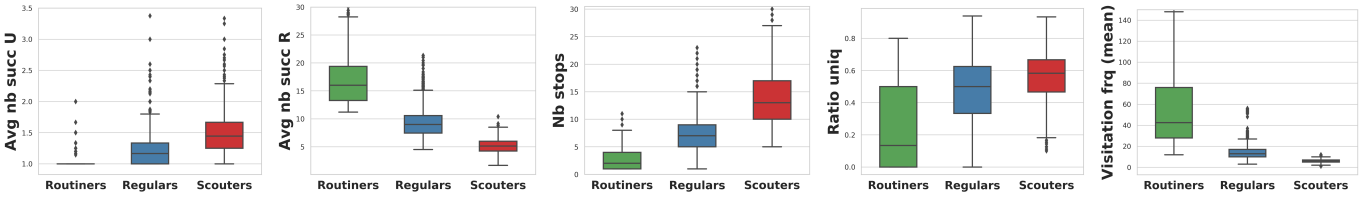
Category	N	Feature name	Description
Relocation Activities	1: (a)	Number of successive explorations	The average number of successive explorations performed by the individual
	2: (b)	Number of successive returns	The average number of successive returns performed by the individual
	3: (c)	Number of stops	The number of distinct areas visited by the individual
	4: (d)	Ratio of unique places	The ratio of places visited only once
	5: (e)	Visitation frequency	The frequency of visits to each area known by the individual
Temporal Activities	6: (a)	Total exploring time	The total amount of time spent by the individual when exploring new places (min)
	7: (b)	Total returning time	The total amount of time spent by the individual when revisiting her known places (min)
	8: (c)	Waiting time	The average amount of time spent by the individual before making a transition to another place (min)
	9: (d)	Duration of successive explorations	The average duration spent by the individual when exploring new places (min)
	10: (e)	Duration of successive returns	The average duration spend by the individual when revisiting her known places (min)
Spatial Activities	11: (a)	Total exploring distance	The total distance walked by the individual when exploring new places (km)
	12: (b)	Total returning distance	The total distance walked by the individual when revisiting her known places (km)
	13: (c)	Radius of gyration r_g	The total radius of gyration of the individual given by $r_g = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(r_i - r_0)^2}$ [1], where r_0 is the center of mass of the individual and N is her set of location history and r_i is a two-dimensional vector containing the geographical coordinates of the location i
	14: (d)	Ratio of distant explorations	The ratio of explorations located outside the circle of center r_0 and radius $R = r_g$ over the total number of visits
	15: (e)	Average displacement	The average distance an individual walks when transiting from a place to another (km)

TABLE II: Extracted features.



(a) Number of explorations (b) Number of returns (c) Number of stops (d) Ratio unique (e) visitation freq

Fig. 4: Relocation Activities in *Agg_gps* dataset (better seen in color).



(a) Number of explorations (b) Number of returns (c) Number of stops (d) Ratio unique (e) visitation freq

Fig. 5: Relocation Activities in *ChineseDB* dataset (better seen in color).

uninterruptedly compared to the rest of the population, as depicted in Figs. 4a and 5a. On the contrary, after a few revisits of familiar spots, they are keen to break their returning routine and chase for new areas to expand their sets of known places as shown in Figs. 4b and 5b. Figs. 4c and 5c depict that *Scouters* have remarkably large sets of of known places. Indeed, this class of individuals performs many explorations, and by consequence, they get to know diverse places. *Scouters* have a surprisingly high ratio of places visited only once. Manifestly, they relish discovering new places. Yet, sometimes they do not revisit or include them in their routinary patterns, as can be perceived in Fig. 4d and Fig. 5d. Moreover, from Fig. 4e and Fig. 5e, we can observe that *Scouters* do not revisit the same places several times, except for some specific ones, which indicates that their routinary patterns consist of a small set of areas.

Figs. 6a and 7a show that the total time amount of time spent by *Scouters* exploring is notably larger than the rest of the population, while their returning time is smaller as depicted in Figs. 6b and 7b. Besides, *Scouters* wait a shorter amount of time before transiting from a place to another as shown in Figs. 6c and 7c. Furthermore, the average duration of successive explorations is higher for this *Scouters* on average they spend more than $200min \approx 3h$ exploring as depicted in Figs. 6d and 7d. Hence, individuals of this class not only relish to discover many places successively but also do it for longer

periods. Conversely, their average returning time is shorter than the other profiles, approximately they spend less than $1000min \approx 16h$ returning as depicted in Figs. 6e and 7e.

Withal, *Scouters* are active, vivacious, and driven individuals. Figs. 8a, 9a, 8b, and 9b point out that they walk longer distances in general. Particularly, they cover longer distances as depicted by Figs. 8e, and 9e. Moreover, as shown in Figs. 8c and 9c, unlike the other groups *Scouters* are characterized by a larger radius of gyrations R_g better seen in the CDR dataset. Namely, they cover larger areas on a daily bases.

B. Routiners' mobility traits

Routiners are steady and rarely leave their zone of comfort. Unlike *Scouters*, they discover very few new places consecutively. Hence, once they explore, they either stay at the same place or go back to a familiar place as shown in Figs. 4a and 5a. Besides, they rarely interrupt their successive returns to discover new areas, this can be observed in the very high value of successive returns in Figs. 4b and 5b. Individuals of this profile have small sets of distinct visited places, meaning that they diversify less their visits and enjoy their routinary habits shifting between familiar locations as depicted by Figs. 4c and 5c. They are also characterized by a small ratio of places visited only once, and a large visitation frequency, as depicted by Figs. 4d, 5d, 4e and 5e. This indicates, that *Routiners* frequently revisit many places they know.

Figs. 6a, 7a, 6b, 7b suggest that *Routiners* spend shorter amount of time exploring. Additionally, they wait larger moments before making a transition to another place as shown by Figs. 6c and 7c. Likewise, Figs. 6d, 7d, 6e and 7e reveal that *Routiners* spend less than $300min \approx 5h$ exploring. Accordingly, they usually prefer to return to their comfort zone before performing another discovery and spend large amounts of time returning before aspiring to discover new spots. Consequently, the total time allocated by these individuals for discoveries is smaller than the rest of the population, and on the contrary, they spend a large amount of time returning.

Routiners do not walk long distances in general as depicted by Figs. 8a 9a, 8b 9b, 8e, and 9e, meaning that even when exploring they go to close areas. They are also characterized by a smaller radius of gyration R_g as depicted by Figs. 8c, and 9c.

C. Regulars' mobility traits

From Figs. 4a, 5a, 4b and 5b, we can observe that *Regulars* alternate between successive explorations and successive returns. In other words, they are constantly shifting between the exploring and the returning states. Besides, Figs. 4c, 5c show that they have a large sets of known places compared to *Routiners* but smaller than the *Scouters*'. From Figs. 4d and 5d, we can observe the same thing concerning the ratio of places visited only once. Further, unlike, *Routiners* they do not equally visit their known locations, but restrict their returns to a small set of places (see Figs. 4e, and 5e).

Regulars spend a larger amount of time exploring compared to the *Routiners* and a larger amount of time returning than *Scouters* as shown in Figs. 6a, 7a, 6b, and 7b. The same can be observed in terms of time spent in successive discoveries and revisits (see Figs. 6d, 7d, 6e, and 7e). Besides they usually wait a medium amount of time before performing transitions from a place to another (see Figs. 6c, and 7c). Additionally, they walk larger distances when exploring compared to *Routiners* as depicted in Figs. 8a, 9a, 8b, 9b, 8c, 9c, 8e, and 9e.

Furthermore, we can also notice from Figs. 9d, and 8d without exclusion all profiles have a high probability to go outside the circle of radius equal to their radius of gyrations $R = R_g$ when exploring.

VI. Spatiotemporal preferences

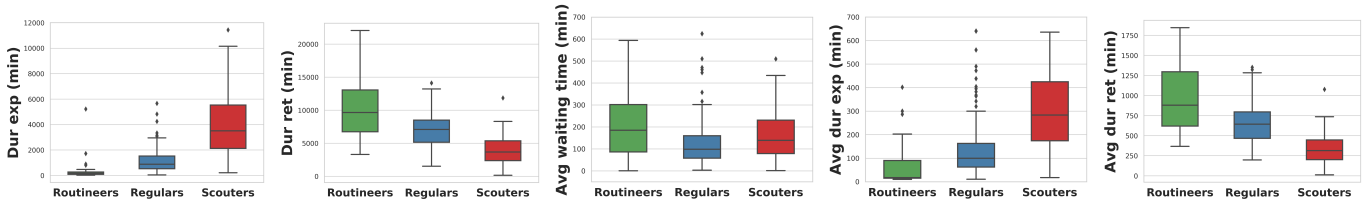
In this section, we verify if there exist temporal or spatial patterns followed by users of each profile when exploring. Admittedly, explorations are characterized by visits to new places (no fine-grained spatial regularity) that cannot be found in the past history of visited places of a user. However, such moments may present some patterns that can still be anticipated once the spatiotemporal features of a user's exploration behavior are well understood and modeled. This is motivated by the fact that such visits may have a temporal or a coarse-grained spatial regularity (e.g., users may like to visit different restaurants or bars but in the same neighborhood and usually on Saturday night) dictated by the user's motivations.

A. Temporal Patterns

We enrich our analysis with the *exploration of temporal semantics*, which refers to the interpretation of the occurring time of explorations, e.g., morning/evening weekday/weekend. This dimension is essential for a thorough understanding of exploratory behaviors, as some discovery events occurring only in certain periods may remain hidden from global patterns. We define *temporal exploration regularity* as repeated explorations over time. For instance, a user exploring at very similar times each week is considered to have a highly regular exploratory temporal pattern at that moment of the week. Hereafter, we use a week-by-week comparison to determine *temporal exploration regularity* of individuals. For this part, we only consider users with high temporal resolution (GPS) and who have at least 4 complete weeks of data. We are thus, left with 224 users.

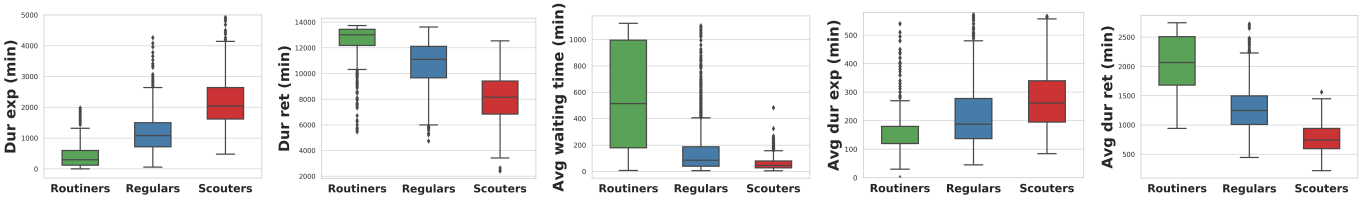
Let the **exploration timeline** denoted by $T_u^w = t_{u,1}^w, \ldots, t_{u,E_{w,u}}^w$ be the ordered sequence of times the user u performed explorations during the week w, where $E_{w,u}$ is the total number of explorations made by u during w and t the offset in minutes from the origin "Monday 00:00" of the considered week.

To quantify the *temporal exploration regularity* we adjust the ISI-Diversity [20] approach used in neural coding to our case of study. First, we define the *Inter-Exploration Interval (IEI)* as the time between two consecutive explorations. We divide each week into periods of one hour. Each week comprises then 24×7 periods $P = [0, 1, \ldots, 24, \ldots, 72, \ldots, 168]$,



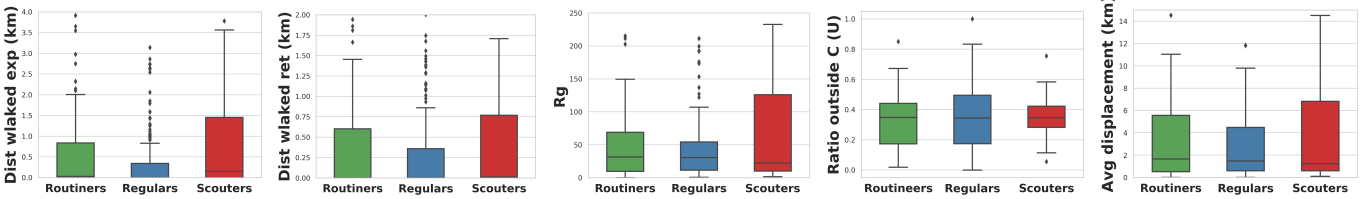
(a) Total U time (b) Total R time (c) Waiting time (d) Avg U duration (e) Avg R duration

Fig. 6: Temporal Activities in *Agg_gps* dataset (better seen in color).



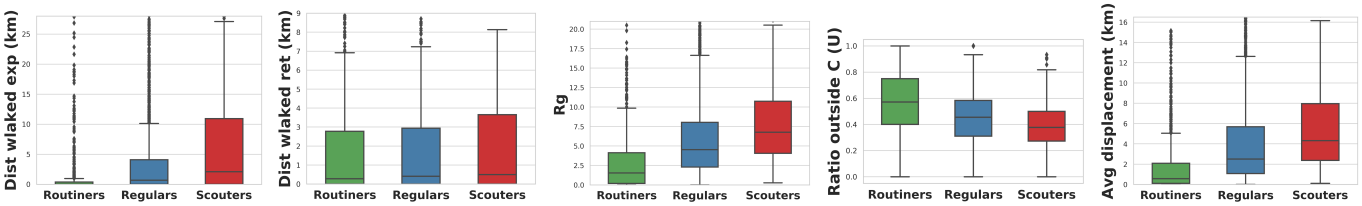
(a) Total U time (b) total R time (c) Waiting time (d) Avg U duration (e) Avg R duration

Fig. 7: Temporal Activities in *ChineseDB* (better seen in color).



(a) Total U distance (b) total R distance (c) Rg (d) Ratio outside C(U) (e) Avg displacement

Fig. 8: Spatial Activities in *Agg_gps* dataset (better seen in color).



(a) Total U distance (b) Total R distance (c) Rg (d) Ratio outside C(U) (e) Avg displacement

Fig. 9: Spatial Activities in *ChineseDB* (better seen in color).

and each period $p \in P$ has a starting time t^p_{min} and an ending time $t^p_{max} = t^p_{min} + 1$. Next, for each user u, we then measure the IEI function $I^w_u(t)$ that gives the IEI at time offset t of the week w, and is given by,

$$I^w_u(t) = min\Big(min(t^w_u | t^w_u > t), t^p_{max}\Big) - max\Big(max(t^w_u | t^w_u < t), t^p_{min}\Big), \quad (1)$$

if $t \in [t^p_{min}, t < t^p_{max}]$. If instead, there are no exploration events within the period p, the instantaneous IEI will take the maximum possible value of 1 hour, i.e, $I^w_u(t) = 60min$. Next, for each individual, we compute the average instantaneous IEI per period:

$$I^w_u(p) = avg(I^w_u(t) | t^p_{min} \leq t < t^p_{max}) = \frac{1}{M} \sum_{t \in p} I^w_u(t), \quad (2)$$

where $M = |I^w_u(t)|$ and $t^p_{min} \leq t < t^p_{max}$. Last, we compute the instantaneous means per period p for each user u, given by, $\mu_u(p) = \frac{1}{W} \sum_{w=1}^{W} I^w_u(p)$, where W is the total number of weeks (exploration timelines). Finally, we calculate the instantaneous mean per group as follows, $\mu(p) = \frac{1}{|U|} \sum_{u \in U} \mu_x(p)$, where U is the population of a mobility profile.

In Fig. 10, we report the influence of the time of the week on the IEI instantaneous mean per period $\mu(p)$ for each mobility profile. We observe that individuals' exploration activities over the week contribute to their mobility profiles:

- *Scouters*' proclivity to explore is the highest for all periods of the week: They have a smaller inter-exploration

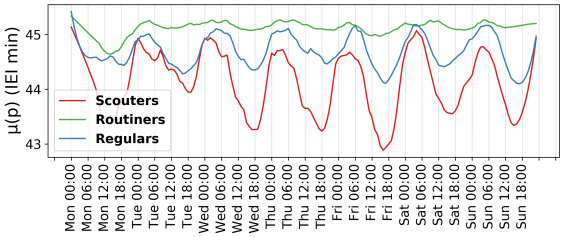
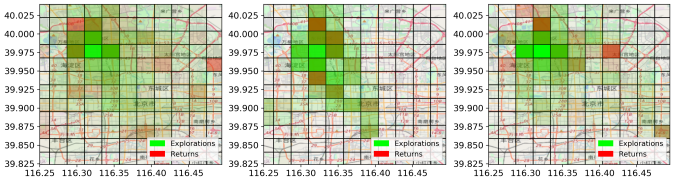


Fig. 10: IEI instantaneous mean per period of 1h



(a) Scouters (b) Routiners (c) Regulars

Fig. 11: Spatial use in Beijing (Downtown).

interval, which also means more exploration is performed. We can also notice that their exploration activities increases by the end of the week reaching its maximum on Friday. Besides *Scouters* tend to have a lower IEI from 4 pm to 8 pm during weekdays and hence explore more by the end of the day.

- *Routiners* have major discrepancies in exploration activities between Monday (cold start problem) and the other periods of the week. This reinforces our previous results on this group, as being stationary and having a higher inclination to stay in their zones of comfort.
- *Regulars*' average instantaneous IEI means are nearly stable over the week during daytime with slightly higher exploration activity on Friday and Sunday.

In summary, conversely to *Scouters*, *Routiners* and *Regulars* relish exploring all over the week mainly in the afternoon and evenings. *Regulars*' proclivity to explore remains stable over the week with a slight increase for the weekends. A larger variation between Monday and the other days of the week can be noticed for *Routiners*.

B. Spatial Coverage

We here analyze and compare the spatial exploitation of *Scouters*, *Routiners*, and *Regulars*. Our main idea is to identify the geographical areas where individuals of each profile prefer to explore and how predictable they are in terms of types of visits in a coarse-grained resolution. In this regard, we put additional grids of size 2 km^2 and we label each of these grids as a *Neighborhood*. Following, for each individual we compute the percentage of the explorations and returns she performed in each *Neighborhood*. Because the datasets (i) are collected independently in different cities and (ii) each city has its own attraction areas and social gathering particularities, hereafter we only present results for one city having the largest number of users, i.e., Beijing of the Geolife dataset.

In Fig. 11, we make a zoom on the most visited areas in Beijing (the city center) and report the spatial coverage of each group among 132 *Neighborhoods* (For entire city use see Appendix B). The intensity of green (cf. red) corresponds to the percentage of explorations (cf. returns) in a given *Neighborhood*: The lighter shades of color indicate a low probability, while darker shades designate a high probability to explore/return. In the following, we list our main observations.

- *Scouters* have a high proclivity to explore in most *Neighborhoods*: Their explorations activities (i.e., green cells) are spread all around the city center. In particular, 83 of the 132 city-center *Neighborhoods* (i.e., more than 62%) are visited for explorations. Besides, their return activities (i.e., red cells) are also dispersed: More than 67% of the *Neighborhoods* are used for returns (see Fig. 11a)[2].
- *Routiners* relish exploring in specific areas and have compact spatial use when visiting: They use around 18% of the city center for explorations and also less than 19% for return activities as shown in Fig. 11b.
- *Regulars* favor visiting *Neighborhoods* within their vicinity when returning, but tend to go to more distant ones when exploring: 34% of the territory is used for both explorations and returns as depicted in Fig. 11c.

Entropy rate	Predictability
Random entropy: $H_u^{rand}(\mathcal{V}) \equiv log(N)$	$\Pi_u^{rand} \equiv \Phi^{-1}(H_u^{rand}, N)$
Temporal-uncorrelated entropy: $H_u^{unc}(\mathcal{V}) \equiv -\sum_{v \in Unique(v_1^T(u))} \frac{N(v)}{T} log\left(\frac{N(v)}{T}\right)$	$\Pi_u^{unc} \equiv \Phi^{-1}(H_u^{unc}, N)$
Real entropy: $H_u(\mathcal{V}) \equiv -\sum_{v_1^T(u) \in \mathcal{V}} P(\mathcal{V} = v_1^T(u)) log\left(P(\mathcal{V} = v_1^T(u))\right)$	$\Pi_u^{max} \equiv \Phi^{-1}(H_u, N)$

TABLE III: Entropy and corresponding predictability, as in [16]

In what follows, we investigate the capacity of correctly forecasting exploring and returning activities with a coarse-grained spatial resolution. For each individual, we consider her sequence of visited *Neighborhoods* when exploring/returning as a stochastic process $\mathcal{V} = \{V_i\}$, where V_i is the i^{th} visited *Neighborhoods* during a period T. We denote by N the number of distinct *Neighborhoods* visited by the considered user u, $N(v)$ the number of appearance of the *Neighborhoods* v and $v_1^T(u)$ the time series of the user u.

For each user, u we assign three entropy measures [21] (see Table III) to capture the degree of predictability of the sequences of visited *Neighborhoods*: (i) the random entropy H_u^{rand} that assumes that all *Neighborhoods* have the same probability to be visited; (ii) the temporal-uncorrelated entropy $H_u^{unc}(\mathcal{V})$, which considers the visitation frequencies to the *Neighborhoods* but overlooks the temporal correlation; (iii) the actual entropy $H_u(\mathcal{V})$ that takes into account the visitation frequency of the *Neighborhoods* along with the order in which they were visited. Next, we evaluate the theoretical predictabil-

[2]Some *Neighborhoods* have light green shades, this implies that they were less visited compared to favorite ones, and are not revisited as regularly visited places.

ity Π, which refers to the maximum probability of correctly forecasting the current *Neighbourhood* from the sequence of previously visited ones. Let $\Phi \equiv xlogx + (1-x)log\frac{(1-x)}{N-1}$ be the function applied to compute the upper bound of the predictability as shown in Table III. Afterward, we compute the PDF of the three versions of the entropy and the corresponding predictability for the sequences of explorations (see Fig. 12) and the sequences of returns (see Fig. 16 in Appendix C) for each mobility profile.

Fig. 12 depicts the entropy rate distributions (left plots) and the equivalent predictability distributions (right plots) of individuals per profile (as shown in Table III), *when exploring new places only*. We can observe the important shift of H_u (green curve) in all groups compared with H_u^{rand} (blue curve) and H_u^{unc} (yellow curve). $f(H_u^{rand})$ picks at 5.8 for the *Scouters*, and around 5 for the *Routiners* and the *Regulars*. This indicates that, the next *Neighbourhood* where a *Scouter* is going to explore can be found among $2^{5.8} = 56$ *Neighbourhoods* and among $2^5 = 32$ *Neighbourhoods* for the others, if the individual chooses her next location to explore in a random way. Instead, $f(H_u)$ picks around 3 for the *Scouters*, 2 for *Routiners* and 2.5 for *Regulars*. In other words, the real uncertainty in terms of number of *Neighbourhoods* is about $2^3 = 8$ for *Scouters*, $2^2 = 4$ for *Routiners* and $2^{2.2} \approx 5$ for *Regulars*.

Additionally, $f(\Pi_u^{max})$ picks at $\Pi_u^{max} \approx 0.78$ for *Scouters* and at 0.8 for *Routiners* and *Regulars*. This means that *only* at least 22% (cf. 20%) of the time, a *Scouter* (cf. a *Routiner* or a *Regular*) chooses her location in a manner that appears to be random. This suggests that, though the apparent randomness of individuals' explorations, a historical record of an individual's discoveries hides an unexpectedly high degree of potential predictability on a *coarse-grained spatial resolution scope*.

VII. Exploration's *impact* on prediction

As previously introduced, any predictor that relies only on the past visiting history of individuals will systematically fail in predicting moments of explorations. As shown through our study, these are numerous and widely present in the daily lives of the *Scouters*. Hereafter, we show how our investigations allow to distinguish from (1) the rest of the population, individuals whose future location visits are hard to predict (essentially due to their high propensity to explore, i.e., the *Scouters*), and (2) the whole mobility of individuals, the moments with novelty-seeking connotation, which is also hard to predict even for *Routiners*.

As a way of illustration, we evaluate the success rate for right predictions using two classical Markovian predictors of order 1 from literature: Markov Chain (MC) and Prediction by Partial Matching (PPM) as in [22]. Such predictors forecast the current location from the set of previously visited locations. In what follows, we use the ChineseDB dataset, as it comprises the largest number of users.

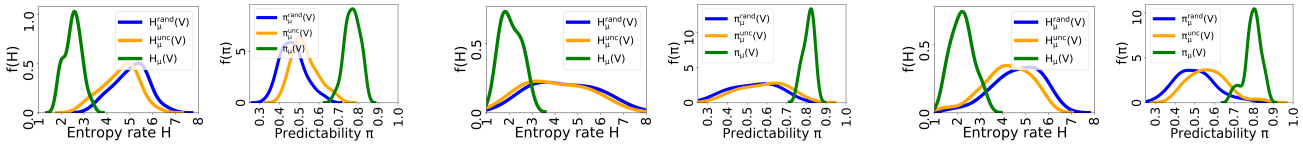
First, for each predictor, we assign to each individual a success rate score initially equal to $s = 0$. Second, we train the predictor using Q records (tuples), where Q is two-thirds of the size of the mobility trace. (we set aside the rest of the trace for testing). Third, we use the predictor to forecast the next location in the next time bin (within 1h). If the predictor correctly forecasts the next location, the score s is incremented ($s = s + 1$). Following, we retrain the predictor using Q plus the last predicted record. Next, we go back to the third step until Q equals the size of the mobility trace. Finally, we normalize the score by the total number of tests, i.e., one-third of the size of the mobility trace.

Fig. 13a depicts the cumulative distribution function of the success score for each predictor. We can see that both MC and PPM achieve their highest performances with the *Routiners* (green) and the lowest ones with the *Scouters* (red). While the success rate for right prediction is higher than 0.6 for 50% of the *Routiners*, the success score for 80% of the *Scouters* is under 0.25. Besides, *Regulars* hold low scores as well. Through, these two simple Markovian predictors, *we confirm the existence, of two main categories of people: those whose mobility is hard to predict,i.e., the Scouters, and those having a highly foreseeable mobility behavior Routiners.*

We now remove novelty-seeking records from the mobility traces, alternatively stated, we select only the records where the user performed a return to a routine location. Next, as earlier we use both predictors MC and PPM to predict the next location using these new mobility traces, i.e., traces with no exploration events, and compute the success score for right predictions.

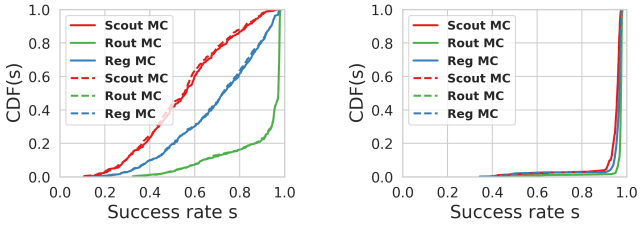
We have two main observations with regard to the results shown in Fig. 13b. First, the predictive performance of both prediction algorithms are no longer as distinguishable among the different mobility profiles as in Fig. 13a. Second, the success score is high for all the groups: 80% of the population has a success score above 0.75. From these notes, we shed light on one of the central origins of predictors low predictive performance, i.e., *explorations*. Indeed, *all groups become more predictable when overlooking novelty-seeking records.* Moreover, *Scouters* who are characterized by their high proclivity to explore become almost as predictable as the other group, whereas a significant difference can be observed among the groups when taking the exploration phenomenon into account.

Highlights: Through our study, we have shown the existence of a category of individuals whose mobility is very hard to predict that we labeled as *Scouters*, essentially due to their high proclivity to explore. While existing predictors can achieve high performance for the other groups, they exhibit weak and low scores for the *Scouters*. Hence, models and predictors considering individuals' tendencies for novelty-seeking are crucial for the *Scouters*. *Our mobility profiling can promptly and easily help to identify this category of people.* Besides, based on our temporal pattern analysis, we can draw out the probability to perform an exploration according to the period of the day and the day of the week. In moments of high exploration probability, a coarse-grained location could be inferred according to our spatial coverage analysis. The



(a) Scouters (b) Routiners (c) Regulars

Fig. 12: Entropy and predictability of profiles when considering only explorations (better seen in color).



(a) Whole trace (b) Sequence of returns

Fig. 13: Distributions of the success rate for each predictor.

intuition here is that services and applications leveraging people mobility could better take advantage of an accurate *Neighborhood*-scale exploration prediction, than of a wrong prediction to a previously visited fine-grained location.

VIII. Conclusions and Future Work

In this paper, we have accomplished four tasks. First, we proposed a new mobility profiling method, with the potential to capture individuals' propensity to explore new areas, namely, *Scouters* (adventurous and prone to explore); (ii) *Routiners*, (steady and routinary), and (iii) *Regulars* (with medium behavior). Second, we extracted the mobility traits of each group and strengthened the subsisting dissimilarity between them. Third, to sustain our profiling method we reported the profiles to the spatial and temporal use. We unveiled individuals' temporal patterns on a weekly basis and showed that *Scouters*' proclivity to explore is very significant throughout the week. Moreover, we showed that explorations in a coarse-grained spatial scenario are far from being random. Finally, we showed how our mobility profiling can help in pinpointing individuals who are hard to predict due to their high proclivity to explore. We then briefly discuss how our approach can help to improve the predictive performance of existing predictors.

For future work, we will apply our mobility profiling and spatiotemporal analysis to develop an adaptive factor, i.e., given a past history of an individual, her mobility profile, and her current context we can indicate her temporal proclivity to explore within the current moment. Further, we aspire to use the adaptive factor indicator to design a predictor. The latter will leverage our spatiotemporal analysis to yield an intuition on the next area where an individual is prone to be in case of an exploration and thus, to adjust simple classical predictions' results.

References

[1] 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.

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

[3] 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.

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

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

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

[7] L. Amichi, A. C. Viana, M. Crovella,and A. Loureiro, "Mobility profiling: Identifying scouters in the crowd," in *Student Workshop of ACM CoNEXT*, Dec. 2019.

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

[9] C. Quadri, M. Zignani, S. Gaito and G. Paolo Rossi, "On Non-Routine Places in Urban Human Mobility," in *IEEE DSAA*, Oct 2018.

[10] V. M. de Lira, S. Rinzivillo, C. Renso, V. C. Times, and P. C. Tedesco, "Investigating semantic regularity of human mobility lifestyle," in *Proceedings of the 18th International Database Engineering Applications Symposium*, ser. IDEAS '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 314–317. [Online]. Available: https://doi.org/10.1145/2628194.2628226

[11] 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.

[12] 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.

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

[14] 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.

[15] 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.

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

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

[18] 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.

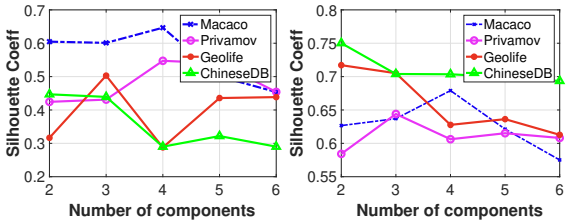
[19] H. Ma, H. Cao, Q. Yang, E. Chen, and J. Tian, "A Habit Mining Approach for Discovering Similar Mobile Users," in *Proceedings of WWW*, Apr. 2012.

[20] T. Kreuz, D. Chicharro, R. G. Andrzejak. J. S. Haas and H. D.I. Abarbanel, "Measuring multiple spike train synchrony," *J NEUROSCI METH*, 2012.

[21] C. Song, Z. Qu, N. Blumm and A.-L. Barabási, "Limits of Predictability in Human Mobility," *Science*, vol. 327, pp. 1018–1021, Feb 2010.
[22] G. Chen, A. C. Viana, M. Fiore., "Takeaways in Large-scale Human Mobility Data Mining," in *IEEE International Symposium LANMAN*, Jun. 2018.

APPENDIX

A. Clustering



(a) GMM (b) k-mean

Fig. 14: Silhouette score.

Fig. 14 depicts the silhouette score, obtained for the two clustering algorithms GMM Fig. 14a and k-mean Fig. 14b. Fig. 14a shows that the optimal number of components for the GMM method varies from a dataset to another. Though a clustering with three elements appears to be more equitable, as all datasets have a score above 0.4. Likewise, the clustering with two components is approximately just as effective. Fig. 14b depicts that 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.

B. Spatial Coverage

Figure 15 depicts the spatial use of *Scouters* in the city of Beijing.

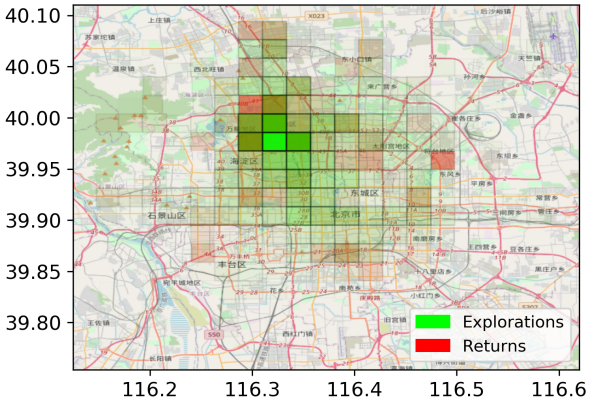
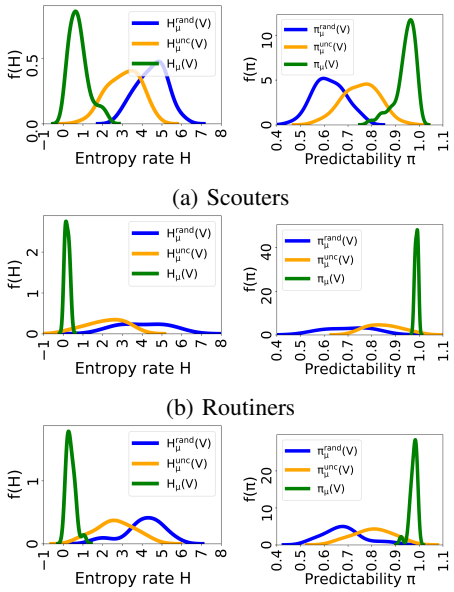


Fig. 15: Spatial use in Beijing for scouters.

C. Predictability of returns

Figure 16 depicts the entropy rate distributions of the three versions of entropy and the equivalent distributions of the upper bounds on the predictability distributions for returns. We can note that $f(\Pi_u)$ narrowly peaks around $\Pi_u \approx 0.98$ for *Routiners*, then comes *Regualrs* with a peack at $\Pi_u \approx 0.96$ than *Scouters* with a pick at $\Pi_u \approx 0.94$. Accordingly, we corroborate our mobility profiling through the spatial exploitation analysis.



(a) Scouters

(b) Routiners

(c) Regulars

Fig. 16: Returns