On the Regularity of Human Mobility

Eduardo Mucelli Rezende Oliveira[a,b,*], Aline Carneiro Viana[b], Carlos Sarraute[c], Jorge Brea[c], Ignacio Alvarez-Hamelin[d]

[a] *École Polytechnique, France*
[b] *INRIA, France*
[c] *Grandata Labs, Argentina*
[d] *Facultad de Ingeniería UBA - CONICET - ITBA, Argentina*

Abstract

Understanding human mobility patterns is crucial to fields such as urban mobility and mobile network planning. For this purpose, we make use of large-scale datasets recording individuals spatio-temporal locations, from eight major world cities: Beijing, Tokyo, New York, Paris, San Francisco, London, Moscow and Mexico City. Our contributions are two-fold: first, we show significant similarities in people's mobility habits regardless of the city and nature of the dataset. Second, we unveil three persistent traits present in an individual's urban mobility: repetitiveness, preference for shortest-paths, and confinement. These characteristics uncover people's tendency to revisit few favourite venues using the shortest-path available.

Keywords: human mobility, mobility, dataset, analysis

1. Introduction

The expansion of metropolitan areas increased the possibility of moving around [1]. This fact together with the increase of smartphone usage brings a very rich opportunity to collect and to investigate human mobility.

People habitually behave as semi-rational entities, routinely moving and interacting within a reduced and predictable geographic landscape, yet unexpected situations can interfere with their preferred direction of motion [2] thus altering their preferred mobility patterns, e.g.: an individual may have to alter his daily commute to work due to a traffic jam or problems with the public transportation. When choosing an itinerary, people try to follow the shortest-path to their destination, this path is also known as the *"desire line"* [3], that is, individuals try to follow the available path closest to the *"desire line"*. Furthermore, people's habitual set of itineraries is characterized by its *confinement*, i. e., people roam close to their main physical address [4].

Datasets are of enormous importance to the analysis of human mobility. They provide the convenience of a non real-time analysis, that is, one can analyze mobility after its parameters (e.g., timestamp, geographic coordinates) have been collected and logged. In the context of large-scale mobility and networks, where real-time analysis is arduous due to the enormous amount of individuals and parameters, logged datasets are widely used as primary source of information.

*Corresponding author
Email address: edumucelli@inria.fr (Eduardo Mucelli Rezende Oliveira)

Most work in the literature study human mobility predominantly from GPS datasets. Although this allows for a fine-grained mobility investigation, datasets collected in large urban scenarios are rarely publicly available. In particular, the experiments to collect human mobility data generally involve people carrying GPS-capable devices which regularly collect their precise positioning. Due to the complexity of those experiments, they tend to be limited in number of participants (e. g., up to 35), time duration (i. e., a few weeks), and space as in university campuses [5], or shopping malls [6]. Lausanne campaign [7] and GeoLife [8] represent some of the few relatively large experiments with around 200 participants, that attempt to collect fine-grained human mobility. The dataset collected from the former is not publicly available, while the one from the latter is. Furthermore, human mobility datasets covering large areas tend to rely *only* on automobile transportation, which is not in the scope of this paper [9, 10].

More recently, datasets collected from cellular networks are being considered by the networking research community. Such datasets, named Call Detail Records (CDR), constitute another source of human mobility. CDR is a metadata record that describes phone communication using a series of data fields, e. g., the identification of callee and caller, call type (voice call or SMS), starting time, ending time, duration of the call, and GPS location of the caller's cell tower [11]. CDR datasets are usually released by Telco operators to a limited number of partners under a non-disclosure agreement and with limited access. As both mobility and network traffic are susceptible of giving away private users' information, entities responsible for such data are careful on providing it to third-parties anonymizing sensitive information.

Besides, it is important to understand the limitations of such CDR datasets. For instance, when modeling mobility using CDRs, one has to know its two biggest limitations: *sparseness in time and coarseness in space*. Time sparseness occurs because records are generated only when a subscriber sends or receive a call or a SMS, which makes he *invisible* at all other periods of time. Space coarseness is due to the size of a cell tower sector, which leads to a location uncertainty of about 1 square mile [12]. It is important to consider that those two characteristics are not uniformly distributed in time due to the fact that subscribers tend to place their calls in bursts, then staying nonactive for long periods [13], around 70% of the total time [12]. Finally, sparseness and coarseness play a negative role on the understanding of some specific human mobility aspects, such as the usage of shortest-path. Although providing coarse-grained mobility information, CDR datasets, when available, allow the mobility investigation in large metropolitan areas.

Additionally, we note that mobility investigations in the literature relate to one or another category of dataset, usually dictated by the dataset availability (Section 2). We claim that human mobility investigation in the context of datasets providing different scales and mobility information brings much stronger certitudes on identified features. Therefore, our contributions to this are is two-fold. We first present an *extensive human mobility analysis from several fine-grained and one additional large-scale coarse-grained urban datasets*. Our datasets represent human mobility from 8 cities in 3 different continents around the world, namely London, Moscow, New York, Paris, San Francisco, Tokyo, Mexico, and Beijing. For each of them, we first model urban scenario with GPS- or CDR-based trajectories and points of interest (Section 4). Our points of interest represent real venues. We have collected information regarding more than 1.5 million unique venues distributed among the studied cities. Our human mobility evaluation comprises visit, temporal and spatiotemporal aspects (Section 5). From our analysis we show that human mobility presents three main characteristics: *tendency to use shortest-path, confinement and a strong repetitive behavior relative to few locations*. From that, our second contribution is to *unveil consistent human mobility characteristics regardless of the dataset representation*. It is not evident that human mobility characteristics seen in one type of dataset will also be found in a different one, which, to the best of our knowledge, no previous work in the literature has assessed.

We conclude this paper by discussing future research directions (Section 6) and providing last remarks (Section 7).

2. Related work

The understanding of mobility and its modeling started with animals such as monkeys [14], jackals [15], and albatrosses [16]. Such works indicated that animal mobility follows a random walk for which their displacement is power-law distributed, i. e., Lévy flight [17]. Early human mobility studies used tracking methods such as bank notes dispersion [4]. Latterly, the lower cost of GPS devices increased the possibility of collecting mobility datasets. In [18], the authors evaluate GPS traces of 44 volunteers in various outdoor scenarios including two different college campuses, a metropolitan area, a theme park and a state fair. The analysis shows that human mobility resembles Lévy flight within a scale of less than 10 km, which corroborates the findings from [4]. Authors then create a Lévy flight model that captures the mobility from those individuals. More recently, easier methods for collecting human mobility in large scale such as mobile phones open new horizons for deeper human mobility investigations.

Through extensive analysis, [19] presents a seminal study on human mobility using a CDR dataset of 100,000 subscribers. Authors show that human trajectories show a high degree of temporal and spatial regularity, in disagreement with the aforementioned random trajectories predicted by the prevailing Lévy flight random walk models. Besides, each individual is characterized by a specific travel distance that is time-independent and a significant probability to return to a few highly frequented locations. The return to a previously visited location occurs with a frequency proportional to the ranking in popularity of the location with respect to other locations. This means, that humans have a strong tendency to return to locations that they visited before, due to the recurrence and temporal periodicity inherent to human mobility. [20] presents an extension of this work using two CDR datasets totalizing 3 million subscribers focusing on the visiting time, i. e., the period of time spent at one location. The resulting curve shows a truncated power-law with a cutoff of 17 hours, which authors link to the typical awake period of humans.

In [21] authors analyze a CDR dataset of 97,000 subscribers in Los Angeles and 71,000 in New York aiming to identify important locations in peoples' lives. Using ground-truth data of home and work location from 19 subscribers, authors were able to identify home and work locations with about 1 and 21 miles of error, respectively. In [22] authors evaluate a dataset of CDRs with information for about 450,000 subscribers to capture city dynamics. More specifically, authors want to discover two main groups in a city, one active during the day (*laborish*) and another during the night (*partyish*). Their grouping strategy relies on a set of fixed rules, e. g., a subscriber is set to laborish group if he makes 4 calls (or send 4 SMSs) during business hours using city cell towers, at least, twice per week. This algorithm correctly matched 81% of the individuals to their corresponding groups using the US Census dataset as ground-truth.

In [23] authors analyze a subset of Lausanne mobile phones dataset [7] with 38 participants in order to understand how temporal and personal factors, e. g., occupation and age, affect individual mobility patterns. Temporal analysis indicate that people are less active during workdays and night than during weekends and daytime. Occupational analysis shows that among full-workers and students, the former are more prone to shorter displacement during the day due to the stricter time rules imposed on companies compared to universities. Finally, age analysis shows higher nightly mobility of younger people compared to older counterparts, which is the result of nightlife attractions being more interesting for younger people. In [24], a study was made using a CDR dataset containing information of 180 subscribers, which presented similar temporal findings.

Due to the routinary behavior of individuals, human mobility is highly predictable. [12] presents a study using a CDR dataset with 50,000 subscribers aiming to measure how predictable

human mobility is. Authors measure the predictability of subscribers' next whereabouts by using three entropy measures: (1) uniform probability among all locations the subscriber visited, (2) probabilities given by the frequency of the visited locations, and (3) probabilities based on frequency, time spent and the order of the visits. As a result, for the typical subscriber, the uncertainty of the next location (i. e., the cell tower the subscriber will be connected to) resides, on average, in a set of less than two locations. Moreover, [19] shows that individuals are found at their first two preferred locations on 40% of the time.

In [25, 26] authors analyse *check-in*[1]-like datasets as source of mobility, wherein a check-in represents a location point, and one can infer user's mobility by a sequence of check-ins. The goal of both works is to create a model of human mobility. Nevertheless, [26] also investigates interpersonal contacts (by means of social-network links and phone calls) and show that social relationships can explain around 10% to 30% of all human displacement, while periodic behavior explains up to 70%.

Besides the efforts above, human mobility has been widely studied from several other points of view, specially with regards to the inter-contact and contact time between people, i. e., the time gap separating two contacts and its duration considering the same pair of people. The importance on those studies comes from a specific problem on intermittently connected networks: as messages are transmitted among nodes when they get in contact with each other, the contact time between pairs of nodes is a key factor on the end-to-end communication delay. In the context of human mobility, people carrying mobile phones are nodes and a contact between devices signify respective people getting closer to each other. The longer they stay close, i. e., the contact duration, the larger the amount of data that can be exchanged.

In [27, 28] authors show that empirical distributions of inter-contact times present two characteristics. First, they are well fitted by log-normal curves, with exponential curves also fitting a significant portion of the distributions. Second, they can be well approximated by a power law over some specific time ranges, from few minutes to 12 hours. [29–31] conducted experiments involving Bluetooth contacts between people carrying devices: [29] studies data from 41 participants at Infocom 2005 conference rooms, [30] analyses 9 participants in a campus scenario, and [31] assess data collected from 16 undergraduate students. Similar results are present in those works regarding contact duration: it is power-law distributed with variations in the scaling exponent k inherent to the specificities on the scenarios where the experiments were carried-out. For instance, the contact duration distribution curve presented in [31] decays slower when compared to the ones from [30, 32]. Authors associate this behavior to students that tend to stay longer periods of time in the vicinity of each other as they may attend the same classes.

The aforementioned works have mostly studied aspects of the human mobility unveiling characteristics on people's displacement such as distance, high probability to revisit certain locations, and dynamic of encounters. Their conclusions indicate that temporal and spatial factors recurrently influence human mobility. However, our intuition says that people's mobility presents other characteristics such as tendency to use shortest-paths. Besides, no large scale evaluation of fine-grained datasets was performed to verify this intuition nor the aspects previously assessed in the literature. We complement thus the literature by providing insights about people's mobility behavior (1) collected from several scenarios presenting different cultural habits, time and space granularities[2], physical infrastructure organization, and geographical size factors that may impact people's routine; (2) coming from different sources/natures of datasets (GPS and CDR).

[1] https://support.foursquare.com/hc/en-us/articles/201065340-Check-ins

[2] Time and space granularity refers, respectively, to time and distance interval between consecutive position samples.

3. Datasets

We analyze several different datasets aiming to find and measure a set of consistent characteristics present in the routine dynamics of urban individuals. In order to avoid biases of a particular data type, the datasets come from different sources, cities and periods of time. Table 1 describes the characteristics of each of the datasets. The mobility datasets come from OpenStreetMap[2], GeoLife[8] and a Telco operator in Mexico, thus presenting three distinct records of users mobility. Generally speaking, each of the datasets present their mobility in a similar way as input files, i.e., each user has a set of trajectory files. Besides, each trajectory file contains, in each line, the latitude and longitude at a given timestamp. Therefore, each line represents the exact location where the respective person was at a certain time. In order to recreate the mobility of certain user, it suffices to repetitively concatenate latitude and longitude points in a temporal order by their respective timestamp for all the user's trajectory files. Section 4 presents more details about the specificities of the trajectories in each of the datasets.

Before presenting our mobility analyzes, we describe hereafter the considered datasets. Due to space restrictions, in the next sections, the graphics mostly show results for GeoLife dataset. However, we will highlight results from different cities throughout the discussion. Since mobility in Beijing and Mexico are present in two datasets, we will refer to Beijing′ and Mexico′ as the ones from OpenStreetMap.

GeoLife. We use the latest version of GeoLife dataset [8]. GeoLife is considered to be unique in the literature. This is due to the fact that it provides a rich view of people mobility using 11 different transportation modes, in an urban area, for a long period of time. It provides geolocalized and timestamped points from 182 people during a 4 year span, from 2007 to 2011, mostly in Beijing. For each person, the dataset provides a set of geolocalized points ascendantly sorted by timestamp, i. e., a fine-grained GPS trajectory. All components are based on geolocalized information, i. e., latitude and longitude coordinates within a 2004 km^2 central area in Beijing. Moreover, to better understand specific behaviors inherent from different periods of the day, every day is divided into four periods of 6 hours, from 00:00 to 05:59, from 06:00 to 11:59, from 12:00 to 17:59, and from 18:00 to 23:59. Due to the routine behavior of people and the large time scale of the GeoLife dataset, it suffices to study a subset of the whole dataset in order to capture the daily behavior of subscribers. Therefore, we select the data of the two most active months in terms of number of users and trajectories. This subset spans from 1st November to 31st December of 2008 and contains 39 users and 2203 trajectories. The following results use this subset of data, unless stated otherwise.

OpenStreetMap. We have collected trajectories using the official OpenStreetMap API[3]. OpenStreetMap is a collaborative project with more than 1.9 million registered users. It has a feature in which users can upload their geolocalized trajectories in order to improve the mapping. We analyze about 14,800 public trajectories uploaded to OpenStreetMap from 8 cities, London, Moscow, New York, Paris, San Francisco, Tokyo, Mexico City and Beijing. As in GeoLife, each user's fine-grained GPS trajectory is a set of geolocalized points ascendingly sorted by timestamp. Besides, similarly to GeoLife, the days were divided in periods of 6 hours each.

Telco. Consists of a CDR dataset with about 6.8 million subscribers collected in a large urban area of Mexico city. It contains the geographic position of the antenna being used and the instant of time when the call was performed for each subscriber from July to October, 2013. As usual

[3] http://wiki.openstreetmap.org/wiki/API

to CDRs, only a few number of geographic points are present per user each day due to the time sparsity of the calls. Moreover, due to the routinary behavior, people tend to make calls using the same antenna. To remedy this sparseness, we have created a 1 week dataset from the original 4-month dataset as following: each day of the week we aggregate all the geographic positions for this respective day from the original dataset, e.g, Monday dataset has the GPS positions of all Mondays in the 4-month dataset for each user. Consequently, this dataset better represents the routine mobility of the subscribers.

Table 1: Characteristics of the mobility datasets

City	*Users*	*Period*	*Days*	*Source*
London	167	7th Nov., 2006 to 14th Dec., 2014	1073	OpenStreetMap
Moscow	197	4th Sep., 2005 to 17th May, 2014	1628	
New York	41	14th Feb., 2008 to 30th Oct., 2014	120	
Paris	182	19th Aug., 2007 to 8th Jan., 2015	556	
San Francisco	62	18th Apr., 2008 to 16th Sep., 2013	214	
Tokyo	87	10th Dec., 2007 to 13th Sep., 2013	513	
Mexico City	22	4th Aug., 2009 to 16th Jun., 2014	85	
Beijing	58	9th Jan., 2008 to 26th Jun., 2015	199	
Mexico City	6.8 M	1st Jul. to 31st Oct., 2013	123	Telco
Beijing	182	12th Apr., 2007 to 27th Oct., 2011	1603	GeoLife

4. System model

This section details the entities of our analysis and the methodology used to extract (from the datasets or external sources) and to represent them in our system model (see Figure 1(a)). Besides, we use data describing more than 1.5 million real points of interest spread in the cities we consider.

4.1. Background

Each of the considered mobility datasets represents a fairly real urban scenario composed by people and their mobility (i.e., a set of timestamped geographic coordinates). Although all the dataset files represent mobility in a similar way, i.e., through a set of trajectories (Section 3), there are conceptual differences between their mobility representation. For instance, GPS and CDR datasets differ in spatial and temporal information: mobility described by CDR dataset is sparse in time and coarse in space (Section 1). This makes the analysis of common human mobility behavior extracted from datasets with different natures and collected in different cities a challenging task: the main problem tackled in our work. Besides such datasets, we perform the collection of data describing more than 1.5 million real points of interest spread in the cities of the considered datasets.

We extract and use three major entities: people, trajectories, and points of interest. The first two are directly represented in the datasets of Table 1. As described in Section 3, recreating the user mobility from the dataset requires chaining user's trajectory points in a temporal order by their timestamp attribute. Doing so for every user in a specific dataset results in the mobility representation of all the users in a respective city. As discussed in the introduction, due to the nature of each of the datasets (i.e, CDR or GPS), GPS datasets allow us to more precisely recreate user mobility due to its high frequent positioning sampling rate (Figure 1(c)), while CDR datasets

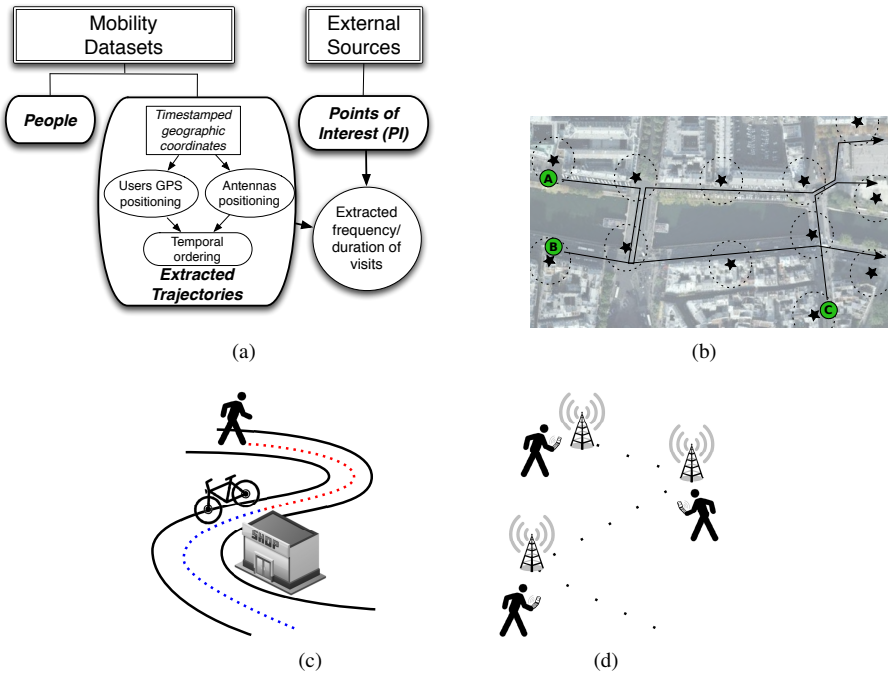


Figure 1: (a) Methodology diagram. (b) People A, B, and C move on the map and visit (circular interaction ranges) PIs encountered on their trajectories. (c)-(d) Illustration of trajectories from different natures: (c) GPS (d) CDR. Each dot corresponds to a latitude and longitude pair. In (c), the shop venue is a PI, extracted for all cities in GPS datasets.

do not possess precise user positioning and the sampling rate depends on the user's call frequency (Figure 1(d)). Thus, analysis related to the shortest path comparison could not be performed for the Telco datasets, the only exception in terms of mobility analysis in this paper.

Points of interest (or PI) are the third entity of our system model. They represent geolocalized physical venues spread in a real urban scenario. Those PIs describe more than mere locations in the map, they reflect routinary aspects (e.g., repetitiveness, confinement) of human's mobility behavior involving people and PIs in a city: e.g., students are frequently going to meet their colleagues in a coffeehouse close to the university they attend. People move and build their trajectories. While moving around, they "interact" with many PIs (e.g., shops, touristic places, bus stations, etc.) and sometimes may stop by. We model this interaction as a *visit*: A visit is considered to happen when an user enters in the interaction range of a point of interest and lasts as long as the user is inside its interaction range. We consider an interaction range, i.e., a circular area, of 50m centered in the point of interest (Figure1(b)). Note that the size of the interaction range was configured to roughly represent an average area occupied by a PI. It is worth to stress that the selection of this value does not bias the performed analysis, since our focus is on the identification of common features present on human's mobility behavior and not on the precise social context interaction of users. We then analyse two properties of users visits: their frequency and duration. The first property counts the frequency in which a certain PI has people in its vicinity. The second takes into consideration the amount of time (i.e., the coverage time) the user stays in the interaction range of the PI . We present results for both in the next section.

Table 2: Average speed ranges by transportation mode

Transp. mode	Walk	Bus	Bike (99.7%) or Run (0.3%)	Taxi (39%), Moto (0.1%), or Car (60%)	Subway	Train	Airplane
Avg. speed range (m/s)	≤ 1.5	$(1.5, 4]$ [33]	$(4, 4.4]$ [34]	$(4.4, 11.5]$ [35]	$(11.5, 28]$ [36]	$(28, 250]$	> 250

4.2. Trajectories

A trajectory represents how people move around and it is described as a set of geolocalized points periodically collected. Regardless the dataset, each trajectory point has latitude, longitude, and timestamp, to indicate when the position was recorded. Hereafter, we detail trajectories described in the considered datasets.

GeoLife and OpenStreetMap. In the GeoLife dataset, trajectories are made up of ten different transportation modes: taxi, bike, run, bus, walk, train, subway, car, boat, and motorcycle. The transportation mode is given by labels set by the users being tracked. In order to improve the precision of our shortest path analysis when comparing users' path lengths with the benchmark's path lengths collected from Google Directions API (e.g., Figure 8(a)), we divided each trajectory into *legs*. A leg is a contiguous set of geolocalized points traveled using a unique transportation mode. Figure 1(c) illustrates a trajectory composed by "walk" (cf. red dots) and "bike" (cf. blue dots) legs. It may represent a situation in which a person went to a bicycle-sharing station, took a bicycle, and continued on it. Not all trajectories in the GeoLife dataset, however, have been labeled. That is due to the fact that labeling was not mandatory for people participating on the GeoLife experiment. In order to label all the trajectories, we have created a simple inference strategy that labels legs by their speed compared against known average speeds for transportation modes (Table 2). Consider we have an unlabeled leg traveled with average speed of 5 m/s. Such speed falls on a range describing average speeds of three transportation modes: "taxi", "motorcycle", and "car". In order to keep the proportion of legs that were originally labeled by the users in the experiment, we have calculated the percentage of legs (shown between parentheses) on each overlapping range. Therefore, the unlabeled leg will be labeled either as "taxi", "motorcycle", or "car" with 39%, 0.1%, and 60% of chance, respectively.

Contrarily to GeoLife, OpenStreetMap trajectories are not labeled at all. To overcome this limitation, we have used our inference strategy described here above: we infer the transportation mode for a given leg based on the average speed of the user along this leg (Table 2). Nevertheless, due to the lack of a sample with labeled trajectories from the original dataset, the probabilities are equally divided, e.g., "taxi", "motorcycle", or "car" have 33.3% of chance of being assigned to a leg.

Telco. As is usual for mobility traces based on CDRs, we use the antennas used by the subscriber as a proxy for identifying geographic locations along a users trajectory. For each subscriber, the identified geographic locations are ascendantly sorted by time of the day. In our dataset, about 70% of the antennas are inside Mexico City urban area and the median pairwise distance between sequential locations per subscriber is 1.5 *km*. It is much more coarse-grained than the 16 *m* from GeoLife or the 7 to 18 *m* granularity from OpenStreetMap. Furthermore, no transportation mode is available or could be inferred from the trajectory locations in this dataset.

Figure 2(a) shows the total number of users and the trajectories they have performed for each day over a three month period in the city of Beijing. As expected, *there is a strong temporal correlation between the daily number of users and trajectories.* The shape similarity of these two

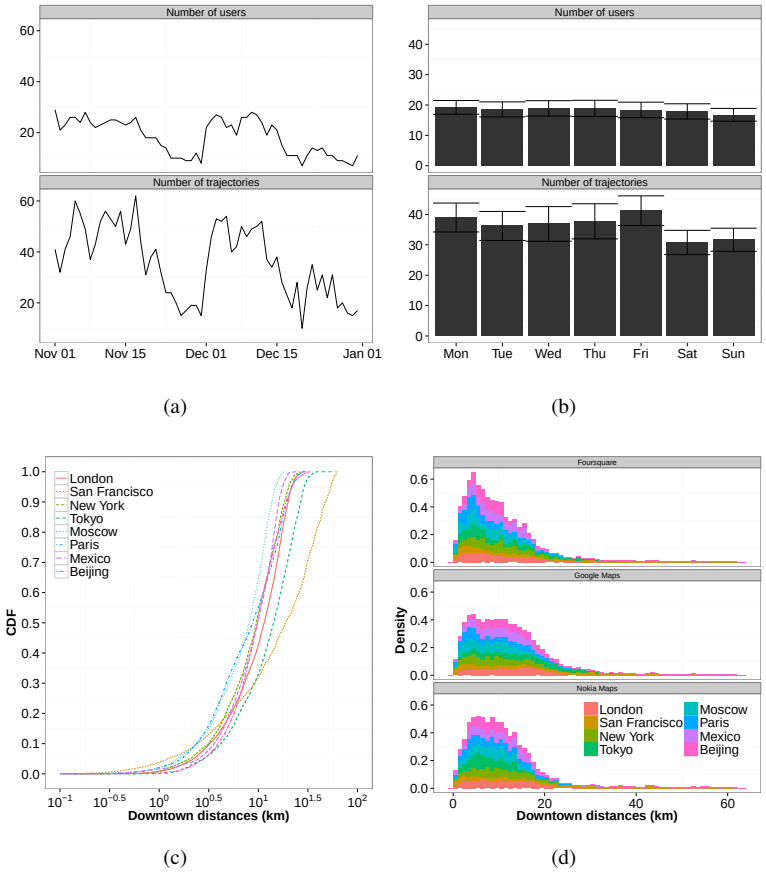


Figure 2: (Better seen in colors) (a) Number of users and trajectories per day in Beijing. (b) Number of trajectories per user grouped per day of the week in Beijing. (c) CDF of the distance from points of interest to the downtown using all sources. (d) Distance from points of interest to the downtown grouped by source.

time series is remarkable: the appearance (cf. disappearance) of a new user in a day of the dataset also implies the appearance (cf. disappearance) of trajectories that day, resulting in a proportional relationship among number of users and number of trajectories on each day. Indeed, Pearson's correlation between number of users and number of trajectories is 92%. Similarly, this correlation is 72% in Moscow, 70% in London, and 60% in Beijing′.

In Figure 2(b), we present the average number of users and trajectories per week day in Beijing. The day-wise difference for the number of users slightly decreases as the week progresses. On average, the highest difference is 12% more users on Monday than on Sunday. As expected, *week days present more people than weekend days.* The average difference over the seven days of the week is 7%. As for the number of trajectories, there is a decreasing trend from Mondays to Sundays, except for a noticeable peak on Fridays presenting 25% more trajectories than Saturdays which is the least active day. This peak is partly due to the fact that Fridays is not only a working day but also has high night activity adding trips to bars, night clubs, etc. All cities presented similar results, for instance, London, San Francisco, Tokyo, and Moscow have 29%, 19% 35%

46% more trajectories on Friday than on Saturday. As expected, *on average, the number of trajectories is higher during the weekdays than during the weekend.* The difference is, on average, 2%, 12%, 23%, 9%, 19%, and 10%, for London, San Francisco, Moscow, Paris, Beijing, Beijing′, and Mexico′ respectively.

4.3. People and Points of Interest (PI)

As people move along their trajectories being recorded via their mobile devices, they may also visit or stop by different points of interest (PIs), e. g., bar, bus station, supermarket, etc. In order to work with real PIs, we have collected information from different databases of places (e. g., Google Places[4]). Such databases are growing and are the most accurate source of public information about points of interest. To avoid possible biases given by a particular database, e. g., some types of interests might be over-represented in a given database, we have collected data from multiple sources, namely Google Places, Nokia Maps[5], and Foursquare[6], adding up to more than 1.5 million distinct points of interest with their respective IDs, latitudes and longitudes. While Google Places and Nokia Maps databases provide information about points of interest collected from city hall's records, by the respective owner of the venues, Foursquare provides only information from places where its users checked in, generally places related with leisure and social relationships. For each set of places collected from a source, repeated points of interest were removed by keeping an unique occurrence of each latitude and longitude pairs, i.e., there will be only one point of interest per geographical position throughout all the datasets. Table 3 describes the characteristics of the sets of collected PIs per source in 98 categories, e. g., market, library, school, etc.

Figure 2(c) shows the distribution of points of interest by their distance to each of the cities' downtown. For instance, we have taken the Big Ben as the center of London downtown, Market Street for San Francisco, Central Park for New York, Imperial Palace for Tokyo, the Red Square for Moscow, Île de la Cité for Paris and Forbidden city for Beijing. Regardless the city, there is a concentration of points of interest closer to the downtown, specially up to 10 km (Table 4). Figure 2(d) shows histograms of the distances from the points of interest to city downtown grouped by source, Foursquare, Nokia Maps, and Google Maps. These result shows that, regardless the source, *points of interest are more concentrated closer to downtown*. Indeed, considering all cities, the highest concentration for Foursquare, Nokia Maps, and Google Maps falls into (4.19, 6.18], (4.19, 6.28], and (6.25, 8.32] km, respectively. Moreover, the median distance from points of interest of Foursquare, Nokia Maps, and Google Maps to the downtown is 8.9, 9.4, and 11.3 km, respectively. Since Foursquare venues are mostly related to leisure, they tend to be, on average, closer to downtown than the ones from Nokia Maps and Google Maps, whose points of interest are distributed in a wider range of niches.

5. Mobility Dynamics

So far we have shown isolated characteristics of the dataset such as number of users, trajectories and how PIs are arranged within the studied cities. This section complements this analysis looking at people's space and time dynamics for different urban scenarios and PIs, e. g., time spent on shopping areas during weekdays and weekends.

[4] https://developers.google.com/maps/documentation/places
[5] https://developer.here.com
[6] https://developer.foursquare.com

Table 3: Area and number of points of interest per city

City	*Area* (km^2)	*Points of Interest*		
		Google Maps	*Nokia Maps*	*Foursquare*
London	1747	227757	56434	44469
Moscow	645	65712	34795	8659
New York	836	88608	61167	33690
Paris	1725	193237	41476	18767
San Francisco	2433	131470	36677	37901
Tokyo	2288	155696	7954	5415
Mexico City	5515	107787	30036	16365
Beijing	2004	77919	119346	5059

In order to understand some of the people's routinary mobility and interaction within their urban environment, we have aggregated the location data by time period of the day or by day of the week. For instance, the curve labeled *Monday* on Figure 3(a) shows the cumulative distribution for the average number of unique users per PI for all Mondays. This also applies to the periods of the day, whose data represents all occurrences of the respective measurement for each of the periods in all days of the week. Our description of the results per day of the week considers that the week *progresses* (or *passes by*) from Monday to Sunday, i. e., it follows the ISO 8601 [37]. Additionally, we refer to Saturday and Sunday together as *weekend*.

5.1. Visit Behavior

Table 4: Summary of some results discussed in Sec. 4.3, 5.1, and 5.2

City	*PI < 10 Km* *Figure 2(c)*	*100/1000 sec. per PI* *Figure 3(b)*	*PI WE/WK* *Figure 3(d)*	*Rep. PI WK 50%/85%* *Figure 5(a)/5(b)*	r_g *WK/WE* *Figure 6(a)*	*Len. Rat. (50%)* *Figure 8(a)*
London	-	49%/80%	8%	23%/47%	82%	65%
Moscow	65%	75%/92%	-	23%/56%	7%	67%
New York	54%	52%/87%	54%	30%/56%	24%	71%
Paris	56%	56%/88%	-	30%/55%	2%	70%
San Francisco	-	63%/87%	19%	30%/58%	-	58%
Tokyo	-	29%/50%	72%	13%/26%	97%	65%
Mexico City	50%	-	-	6%/90%	-	-
Beijing	52%	62%/87%	25%	23%/55%	200%	62%
Beijing′	52%	77%/94%	14%	31%/67%	-	75%
Mexico′	50%	74%/96%	64%	-	-	61%

This section assesses how people interact with their urban environment, exploring several aspects of their daily routines. Given that PIs are essential for this analysis, we must exclude Mexico City from this study.

Figure 3(a) shows the CDFs for the number of unique users per PI, i. e., number of users that visited a PI, in Beijing. Due to the large number of PIs in the city, most of the PIs are rarely visited by the users on a single day. Indeed, 78% of the PIs are visited by only one user per day. This holds for each of the days of the week, 92%, 92%, 91%, 92%, 91%, 94%, 96% of the PIs are visited by up to two users on Monday, Tuesday, Wednesday, Thursday, Friday, Saturday and Sunday, respectively. All the other cities present similar results. Moreover, PIs receive 9% less users on weekends than on weekdays. For Moscow, 93% of the points of interest are visited once. This percentage is even higher for the other cities. Although GeoLife has less users than OpenStreetMap in Moscow, the former tends to have a higher number of users per day, which increases the number of unique visits per PI.

The total number of visits to a given PI can give us information on how frequently people visit this place but it does not inform us on the time spent around this location. Most of the PIs present low coverage time, i. e., the amount of time that users have spent inside PI's interaction range for up to 100 and 1000 seconds (Table 4). Figure 3(b) shows the coverage time per PI per day of the week in Beijing. The time spent in the vicinity of a given PI increases as the week passes by. A plausible explanation for this is that people start the week at a fast pace, slowing down as the week ends. That is, at the beginning of the week, people simply disregard venues passing by in a hurry, but as the weekend approaches, they are willing to spend more time planning for leisure time on Saturdays and Sundays, e. g., looking at the showcases or visiting stores. On average, *people spent 63%, 87%, 33%, and 33% more time on the vicinity of places on Sunday than on Monday* for San Francisco, Beijing, Moscow, and Mexico′. Additionally, *PIs, on their vicinity, have people 12%, 20%, 17%, and 73% more time on weekend than on weekdays* for the same cities.

Figure 3(c) shows the CDFs of the total number of PIs seen, i. e., including repeated visits, per user per period of the day in Beijing. The earliest and the latest periods of the day, i. e., from 00:00 to 05:59 and from 18:00 to 23:59, present the least number of visited PIs. Briefly, the main reason for that is the shorter length of users' trajectories during those two periods compared with the other periods of the day. We further develop the idea of trajectory length in Sec. 5.2. On the other hand, *users visit the highest number of PIs from 12:00 to 17:59*, which is expected due to the daily activities. On average, users from 12:00 to 17:59 visit 82% and 79% more PIs than from 00:00 to 05:59 and from 18:00 to 23:59, respectively. In Tokyo, for the same periods this difference is 57% and 60%, respectively. Additionally, if we consider the majority of the users, for example, for up to 75% of them in Beijing, 94% more PIs are visited per user from 12:00 to 17:59 than from 00:00 to 05:59, which is the period with least number of PI visits per user. Still comparing in Beijing, for the same percentage of users, 52% more PI are visited per user from 12:00 to 17:59 than from 18:00 to 23:59. Regarding the period from 06:00 to 11:59, which is also almost as active as from 12:00 to 17:59, the former has 11% less visits per PI than the latter.

New York presents similar results, for up to 75% of users, the most active period of the day, which is from 12:00 to 17:59, has 27%, 72%, and 21% more visits to PIs than from 00:00 to 05:59, from 06:00 to 11:59, and from 18:00 to 23:59, respectively. It is important to note that this behavior slightly varies in certain cities, while it presents extremely similar results in others. Aside from New York, Tokyo presents very similar results. On the other hand, cities such as London and San Francisco still had high number of visits during daylight periods, but similar number of PI visits during the night. In such cases it is difficult to indicate a single reason. It might be due to the city context, i. e., people are as active during the day as during the night, or to dataset's characteristics, which contain a more balanced number of users during the day and night.

Figure 3(d) shows the CDFs of the number of PIs seen per user per day of the week in Beijing. *Similar to spending more time on the vicinity of PIs during the weekends than on weekdays (Figure 3(b)), on our data, people tend to pass by more PIs on weekends than on weekdays (Table 4).* Moreover, there is a growth on the number of PIs from Monday to Friday. Indeed, *on average 40% more PIs are visited on Friday than on Monday* considering all cities. Cumulative results show similar tendency, 30%, 21%, 40%, 65%, 25%, 15%, and 40% more visited PIs on weekends than on weekdays for up to 75% of the users in London, San Francisco, New York, Tokyo, Beijing, Beijing′, and Mexico′, respectively.

We further investigate the interaction between people and urban scenario by segmenting the city using *cells*. In our context, cells are square-shaped regions of $50m^2$ organized in a grid fashion on the city terrain. Figure 4(a) shows the *hexagonal bin plot* [38] of cell distance to downtown and number of PIs inside the cell in Beijing. The intensity of a bin represents the frequency of

cells that contain a number of PIs laying within the bin. There is a *densification, i. e., higher concentration of PIs closer to downtown, that decreases with the increase of the distance.* This is a common aspect of metropolitan areas, there is a strong negative correlation between distance to downtown and number of venues, -95%, -55%, -64%, -57%, -90%, -97%, -93%, and -95% for London, San Francisco, New York, Tokyo, Moscow, Paris, Beijing, and Mexico′. Although San Francisco and Tokyo are large metropoles, their concentration on the surface tend to be more truncated due the limitations of the bays present on both of them. That is probably the cause of their lower correlation compared to the other evaluated cities. There is also a high frequency of cells containing a low number of PIs irrespective of the distance to downtown. Bigger venues may explain this, e. g., a city hall could occupy the whole space of a single cell.

Figure 4(b) shows the total number of visited cells, i. e., including repeated visits, per user per day of the week in Beijing. The tendency of the CDF curves is similar to the Figure 3(c), but shifted to the left due to an expected lower number of cells than PIs. Besides, *on average the number of visited cells grows from Monday to Sunday with a peak on Friday.* For instance, people visited 65% more cells on Friday than on Monday and 6% more cells on weekends than on weekdays. Considering all cities, on average, those percentages are 55% and 12%, respectively.

In order to better understand the predictability of people's mobility, we calculate the L rank [19] of the visited PIs and cells. The rank is calculated per user and it takes into consideration the number of times he visits a cell/PIs, e. g., the most visited cell/PI by a user has rank $L = 1$. Figure 4(c) shows a Zipf plot of the visiting frequency for the cells and PIs ranked L in San Francisco and Beijing. The dashed straight line shows that the distribution can be approximated by $1/L$. Furthermore, *this plot shows that users concentrate most of their visits to few frequently visited cells and PIs, i. e., to a very restricted area.* For example in Beijing, 43% and 40% of the visits are made to 1% of the cells and PIs, respectively.

From the previous analysis, we see a strong repetitive tendency in human mobility, i. e., to repetitively visit the same areas. To quantitatively express how repetitively a PI is visited, a metric called *Repetitiveness* was conceived. The repetitiveness of a PI v_i is based on the number of unique users (NUU) and total users (NTU) that visited it as following:

$$re(v_i) = \frac{NTU_{v_i} - NUU_{v_i}}{NTU_{v_i}} * 100 \quad (1)$$

Figure 5(a) shows the CDF of the average repetitiveness per week day in Beijing. This result shows that the majority of the PIs present low repetitiveness and a minority has high repetitiveness. Table 4 shows the percentage of PIs in each of the cities for up to 50% and 85% of repetitiveness. Besides, for all cities, $\approx 1\%$ of the PIs are highly repetitively visited, presenting more than 98% of repetitiveness. As the system model for Mexico city does not contain PIs, we have calculate repetitiveness using the cellular network's antennas. Figure 5(b) shows the CDF of the average repetitiveness per week day in Mexico considering cellular antennas instead of PIs. The results are similar to counterpart results for PIs, majority of antennas are barely used, while a very small amount of them is used several times by the same subscriber. 6% of the antennas present up to 50% of repetitiveness, 90% have up to 85% of repetitiveness, and less than 1% of antennas present more than 98% of repetitiveness. For all cities, the average repetitiveness difference between weekdays and weekends is 4.5%.

5.2. Displacement behavior

In order to evaluate how much physical space users cover, and how they move about this space, we look at their maximum displacement, their desire lines, and their radius of gyration,

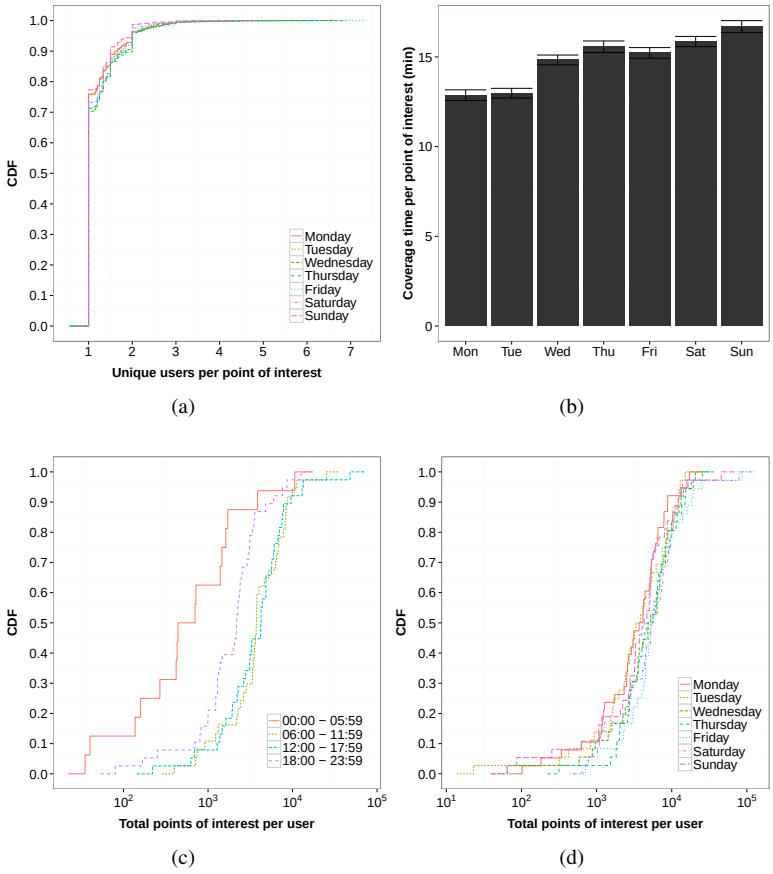


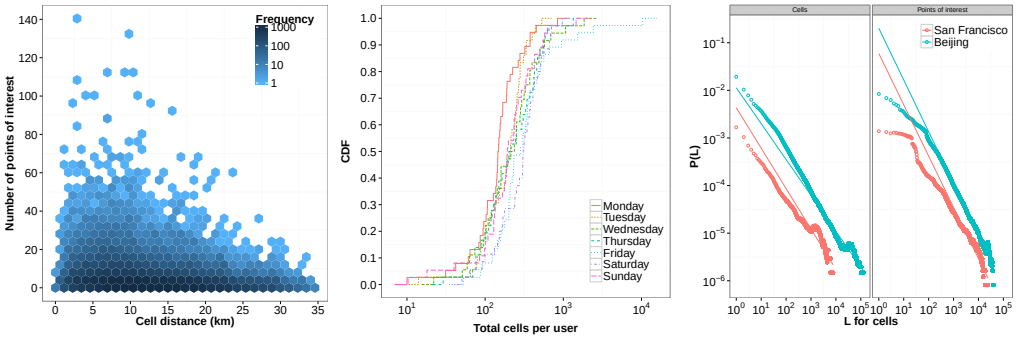
Figure 3: (Better seen in colors) (a) Number of users and (b) mean coverage time per PI and day of the week in Beijing. Number of visited PIs per user per (c) period of the day and (d) per day of the week in Beijing.

which is the linear size occupied by each user's trajectory after a time t, thus the disc given by this radius is an estimate of the area size covered by a user at time t. We define this quantity as:

$$r_g^u(t) = \sqrt{\frac{1}{n}\sum_{p=1}^{n}(\overrightarrow{r}_p^u - \overrightarrow{r}_{cm}^u)^2} \tag{2}$$

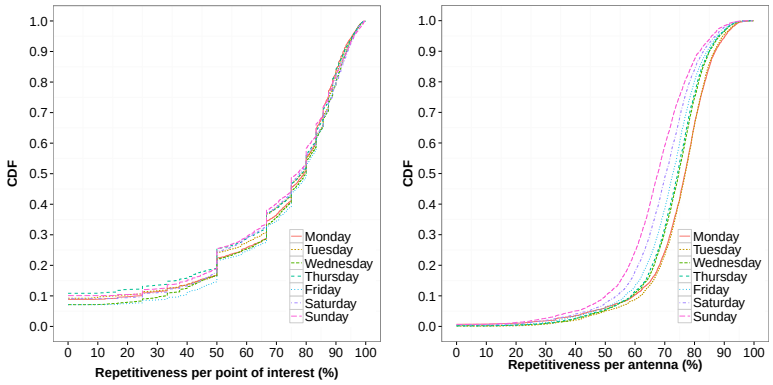
where $\overrightarrow{r}_p^u$ represents all the trajectory points $p = 1, \ldots, n$ of the user u and $\overrightarrow{r}_{cm}^u = \frac{1}{n}\sum_{p=1}^{n}\overrightarrow{r}_p^u$ is the center of mass of the trajectory. $r_g^u(t)$ captures thus how broadly the users travel as opposed to the actual distance traveled. In our results, the unit of the radius of gyration is meters. During t all the trajectories of each user u are considered in the calculation of r_g, which is expected to grow with the growth of t. The following results analyse r_g for different durations of time, periods of the days, whole days and the 2 months of the dataset.

In general, the radius of gyration is higher on weekends than on weekdays (Table 4). Figure 6(a) shows the CDFs of the radius of gyration per period of the day on weekdays and weekends. *On weekdays, the earliest period of the day presents the smallest radius of gyration.* That is



(a) (b) (c)

Figure 4: (Better seen in colors) (a) Bin plot of the number of PIs per cell distance in Beijing. (b) Total number of cells visited per user per day of the week in Beijing. (c) L rank for cells (left) and PIs (right) in San Francisco and Beijing.



(a) (b)

Figure 5: (Better seen in colors) Repetitiveness of PIs per days of the week in (a) Beijing and (b) Mexico.

coherent with human routines, from 06:00 to 23:59 we perform more activities and are more susceptible to displacement that covers a wider area. Contrarily, *from 00:00 to 05:59 people are more stationary performing at most short trajectories and likely at home, sleeping, etc.* Median radius of gyration per user in the period from 00:00 to 05:59, is 92% shorter than the radius of gyration from 06:00 to 11:59 in Beijing. Considering all cities, except Mexico, the radius of gyration from 00:00 to 05:59, is 57% shorter than the radius of gyration from 06:00 to 11:59. For instance, the average radius of gyration is 759 meters from 00:00 to 05:59 and about 10 km from 06:00 to 11:59 in Beijing. Due to the sparsity of the Telco dataset, the radius of gyration tends to be larger on Mexico than on the other cities. Therefore, we have made a separate analysis for Mexico dataset on displacement aspects. Figure 7(a) shows the radius of gyration for subscribers per period of the day on weekdays and weekends in Mexico. Radius of gyration from 00:00 to 05:59, is 69% shorter than the radius of gyration from 06:00 to 11:59. For instance, the radius of gyration is 13.9 km from 00:00 to 05:59 and 46.4 km from 06:00 to 11:59.

Differently, on weekends the radius of gyration from 00:00 to 05:59 grows 49% in Mexico,

46% in New York, 6% in Tokyo, and 35% in Beijing when compared to weekdays. That is due to the nightlife activities which increase the late night mobility of the users. As *a probable consequence of the higher mobile behavior on weekends' late nights, there is a reduction on the average radius of gyration from 06:00 to 11:59 on weekends when compared to same period on weekdays.* For instance, it is 41, 2.5, 1.1, and 4.6 km in Mexico, New York, Tokyo, and Beijing, which it is 12%, 20%, 49% and 53% less than on the same period on weekdays, respectively. That is likely due to the people waking up later on weekends than on weekdays.

Figure 6(b) depicts the CDFs of the radius of gyration per day of the week in Beijing. On average, *users tend to journey over a larger area as the week passes by from Monday to Saturday, with the exception of Sunday which has r_g comparable to Monday.* For instance in Beijing, the radius of gyration on Monday, Wednesday and Friday is 1.6, 1.8 and 2.3 km, respectively. Considering all cities, except Mexico, it is 1.7, 1.8, and 2.3 km, for the same days, respectively. The radius of gyration has the highest values on Friday and Saturday, the latter 2.9 km in Beijing and 2.8 km average considering all cities except Mexico. Due to the peak on Saturday, average radius of gyration is higher on weekends than on weekdays, 1.9 and 2.5 km, respectively in Beijing and 1.9 and 2.2 km for all cities. Figure 7(b) shows the same results for Mexico. The behavior present on the GeoLife and OpenStreetMap is also present on the Telco dataset, i. e., peak on Friday and Saturday, 69 and 71 km, respectively. Moreover, the average radius of gyration grows as the week passes by.

Figure 6(c) shows the CDF of the final radius of gyration per user in Beijing. The steady and constant increase on the CDF curve shows that *users are almost equally distributed by their radius of gyration.* To further analyze the radius of gyration, we have grouped users by their final radius of gyration into four groups: $r_g \leq 10^4$, $10^4 < r_g \leq 10^5$, $10^5 < r_g \leq 10^6$, and $r_g > 10^6$ meters. Figure 6(d) shows the average radius of gyration for each of the groups up to the hour on the x-axis. The confidence intervals are shown as shadows around the curves. The saturation on the curves shows an upper bound for the movement area on each of the groups. An interesting aspect is how fast each of the groups reach (or approaches) their saturation values. For instance, at the end of the first day, 69%, 17%, $\approx$ 1,% and $\approx$ 1% of the final r_g has been reached in the groups $r_g \leq 10^4$, $10^4 < r_g \leq 10^5$, $10^5 < r_g \leq 10^6$, and $r_g > 10^6$, respectively. On one week, users on the same groups have reached 88%, 40%, 35%, and 12% of their final r_g. It means that *users whose mobility is more confined tend to reach the upper boundary of their movement proportionally faster than the ones who journey over larger areas.*

The concept of *desire lines* states that people tend to choose the shortest-paths to arrive to their destinations. To verify that, we have compared the length of each traveled leg (Section 4.2) against the length of the corresponding shortest path considering the same initial and final points of the original leg. Dividing the length of the original leg by the length of the shortest path allows us to measure how longer the path made by a person is from the shortest path. We have used Google Directions API[8] to compute the shortest path. The API receives the coordinates of both initial and final points and a *travel mode*, i. e., transportation mode. Then, it returns the shortest path considering the restrictions imposed by the existing routes and obstacles in the city for a specific transportation mode. Note that, we only have considered transportation modes where people have decision control of their paths. This excludes for example, buses, boats or trains.

Legs traveled by *walk*, *run*, and *bike* had their lengths divided by their respective shortest paths computed while using the API in *walking* mode. Google Directions API indeed has a *bicycling* travel mode, but at the moment, it does not contain routes in Beijing. Therefore, we use *bicycling* mode when available in the evaluated cities. Legs traveled by *taxi*, *car*, and *motorcycle* had their lengths divided by results of the API in the *driving* mode.

[8] https://developers.google.com/maps/documentation/directions

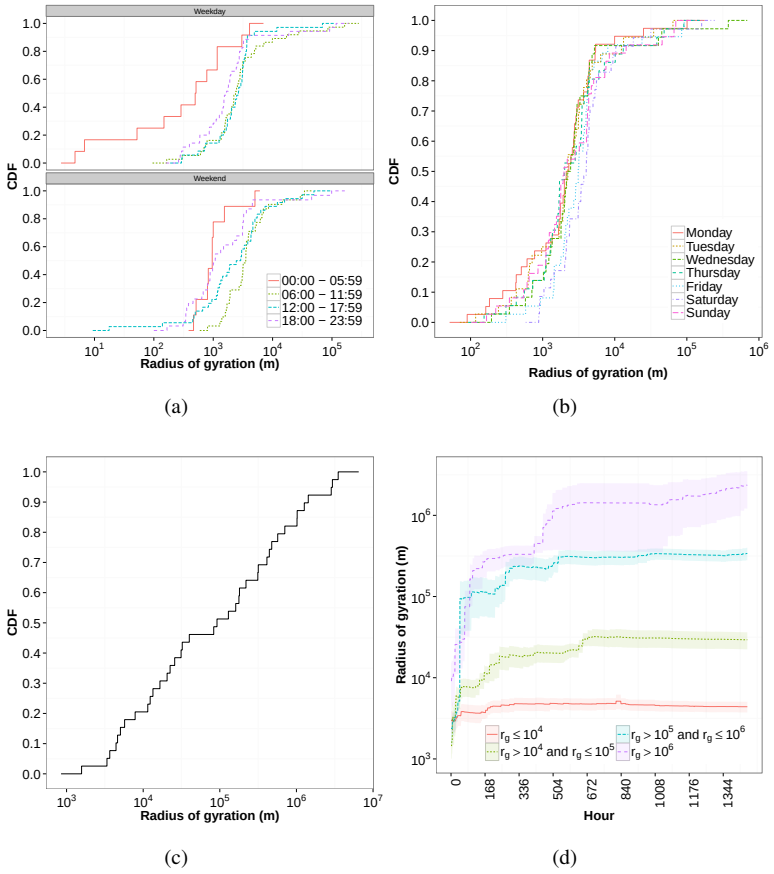
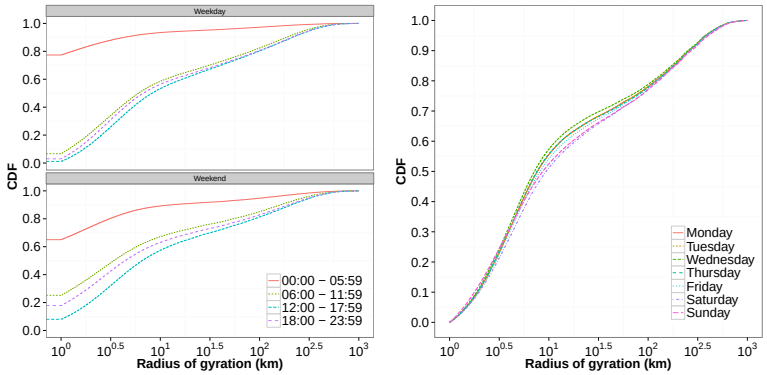


Figure 6: (Better seen in colors) Radius of gyration for users per (a) period of the day, weekdays (top), and weekends (bottom), (b) day of the week, (c) for all periods and days, and (d) for all periods and days grouped by final radius of gyration in Beijing.

Figure 8(a) shows the CDF of the ratio between the original legs length and the shortest path, by transportation mode and period in Beijing. It shows that the periods from 00:00 to 05:59, from 06:00 to 11:59, from 12:00 to 17:59, and from 18:00 to 23:59 present, respectively, 36%, 62%, 52% and 74% of the legs measuring, at most, half longer than the shortest path. For all other cities, the average percentages for the same periods are 44%, 53%, 57%, and 73%, respectively. These results show that on late night people tend to walk around not directly going to their destination. Indeed, on late night people tend to go for bars, night clubs and are more susceptible to create routes that are way longer than the shortest ones. On the other hand, periods representing early morning and early night show high percentage of legs closest to the shortest one and describing how people go directly to their destinations, e. g., work, home, etc. The period containing the early afternoon hours present an intermediate percentage of legs close to the shortest path. Indeed, this period mixes people walking around careless about shortest paths (e.g, someone shopping, or looking for restaurants), and people more concerned about being on time (e.g, people coming back to work from lunchtime). Moreover, it is possible to see that the length ratio changes in



(a) (b)

Figure 7: (Better seen in colors) Radius of gyration for subscribers per (a) period of the day, weekdays (top) and weekends (bottom) and (b) day of the week in Mexico.

function of the transportation mode. For instance, *walk* and *taxi* modes present trajectories that are closer to the shortest path. That is probably due to the human capacity of being able to identify the trajectories, mainly when one has the knowledge of the neighborhood, which is the case of the participants of GeoLife experiment. Taxis tend to be equipped with GPS-enable devices and route planning software in order to find the addresses and the better (shorter in time and/or cost) routes. Besides, Table 4 shows the percentage of trajectories half longer than the shortest-path for all cities. We conclude that *regardless of the transportation mode, people tend to be oriented by the shortest paths.*

We have also analyzed the length ratio grouped per weekdays and weekends and per transportation mode. For Beijing, the median length ratio is 1.1 and 1.3 on weekdays and weekends, respectively. Those values are consistent for all cities, on average 1.1 and 1.2, respectively. Additionally, on all cities, *bike*, *car*, *taxi*, and *walk* presented median length ratio of 1.05, 1.07, 1.12, and 1.39 for weekdays and 1.06, 1.2, 1.33, and 1.72 for weekends, respectively. It shows that *people on our datasets were presenting more routes closer to the shortest ones on weekdays than on weekends.* That is interesting because it measures a difference in people's behavior on weekdays and weekends.

People's mobility is generally *confined*. Even if people are not using the shortest routes, they are at least not going far from their home location. To check how that premise occurs in our scenario, we have measured how confined the trajectories are by their maximum displacement. Maximum displacement is the distance between a trajectory's initial and farthest point (not necessarily the last point). Figure 8(b) presents the CDF of the maximum displacement for all trajectories grouped by period of the day in Beijing. It shows that *90% of the trajectories per period of the day have, at most, 10 km maximum displacement in Beijing. Considering all cities, except Mexico, this values is 75%.* Similar findings are present on the analysis of maximum displacement per day. For Mexico, the displacement is generally larger than 10 km due to the coarse-grained nature of the CDR dataset. Figure 8(c) presents the maximum displacement per period of the day for Mexico City. 57% of the trajectories per period of the day have, at most, 10 km. This value is lower than the other cities because of the sparsity of the dataset. For instance, 75% of the trajectories have, at most, 35 km. Additionally, there is a significant difference between the maximum displacement from 00:00 to 05:59 to the other periods, which is not observed nor on

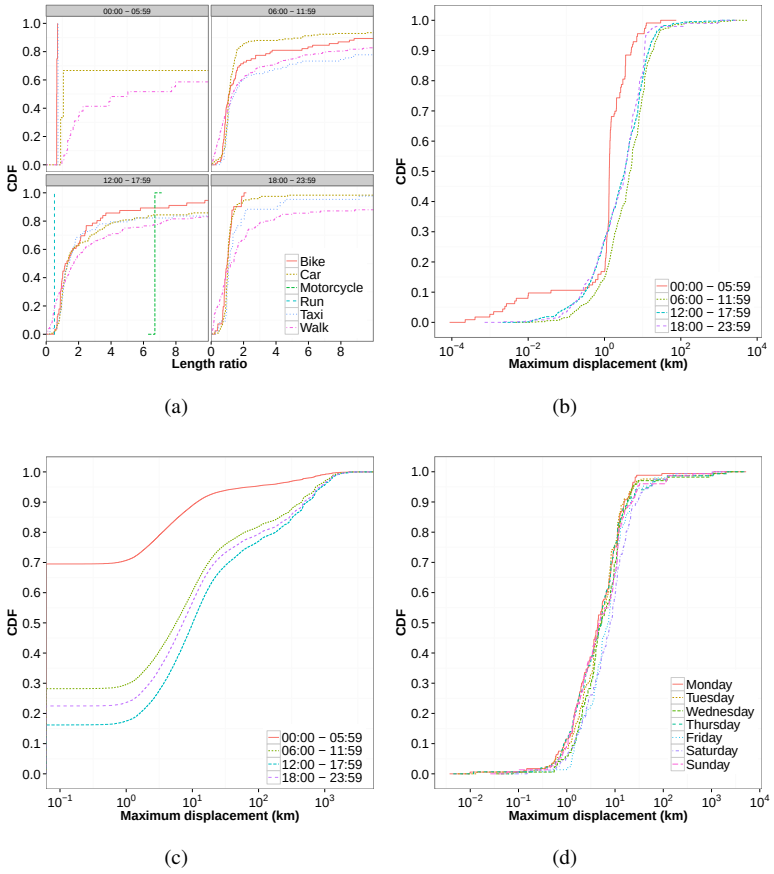


Figure 8: (Better seen in colors) (a) CDF of the length ratio per transportation mode grouped per period in Beijing. (b) Maximum displacement per period in Beijing. (c) Maximum displacement per period in Mexico. (d) Maximum displacement per day of the week in Beijing.

GeoLife neither on OpenStreetMap. This is due to the difference between a fine-grained mobility and sparse mobility. Generally, mobility is more frequent than calls from 00:00 to 05:59.

Figure 8(d) shows the CDF of the maximum displacement per user per day of the week in Beijing. For example, the median maximum displacement from Sunday to Thursday ranges from 4.4 to 5.2 km, and it is higher on Friday and Saturday, 7.2 and 7.9 km, respectively. This is a reoccurring behavior in all the cities. Median maximum displacement from Sunday to Thursday ranges from 6.3 to 6.9 km, and on Friday and Saturday, 7.1 and 7 km, respectively. This result shows that generally people do not move far away from their starting point, and presents high confinement. This result reinforces the findings for the radius of gyration. In fact, *there is a 96% correlation between users' maximum displacement and radius of gyration.*

5.3. Spatiotemporal behavior

People's mobility and visiting behavior may reflect some of their preferences and lifestyle. To better understand the visiting behavior of people, we have classified the categories of our PIs

(refer to Sec. 4.3) in 9 classes. This classification aims to group together, on the same class, PIs whose categories carry similar meaning. For example, class *Education* (which we abbreviate as *Edu*) groups together all PIs with category "school", "university", and "library". Similarly, the remaining 95 categories were classified into more 8 classes. Table 5 describes the classes and some of the categories they contain.

Figure 9(a) shows the amount of coverage time each of classes provided per period of the day in Beijing. Among all classes, *Food*, *Shop*, *Trvl*, and *Rel* present higher values on at least one period of the day. From 00:00 to 05:59, PIs in the *Food* class presents the highest coverage, 30 minutes on average. On London, San Francisco, and New York this period presents highest coverage for class *NL*, which is understandable due to the night-life related activities in those cities. From 06:00 to 11:59, *Shop*, *Food* and *Trvl* have the highest amount of coverage time, 115, 61, and 51 minutes, respectively. That is probably due to the shopping and breakfast-related PIs before using the transportation to work or study-related places. Similarly, from 12:00 to 17:59, the order of the classes with the highest coverage is the same: *Shop*, *Food* and *Trvl*, 211, 100, and 64 minutes, respectively. All other cities present mostly similar results from 12:00 to 17:59, with the inclusion of *Srvc* class being significant together with *Shop* and *Food*, they are among the top 4 classes that more provide coverage. This is probably related to lunch and transportation back home. Differently, from 18:00 to 23:59, *Rel* class in Beijing shows the highest coverage time, 93 minutes, with large confidence interval. It is still unclear why the average coverage time for *Rel* is the highest in this period, but the large confidence interval is due to the few occurrences of this class. It might be the case that few users share a particular religious ceremony during night time. This period is different from the other cities we have evaluated and it is likely related to local circumstances. Aside from that, from 18:00 to 23:59 *Shop* and *NL* have 50 and 30 minutes of coverage time, respectively. For all other cities, those two classes are among the top 3 that most provided coverage in this period.

Figure 9(b) presents further investigation for the coverage time per points' of interest class. It shows the same data presented on Figure 9(a) grouped by weekdays and weekends instead of periods in Beijing. It is possible to see a significant difference between the coverage time on weekdays and weekends for *Shop* class. For instance, on average, it is 87 minutes on weekdays and 323 on weekends. On weekends, people normally have more time to spend on shopping areas than on weekdays and that is the probable cause of this difference. However, this aspect depends on the opening hours during weekends, e. g., *Shop* class in Paris class has, on average, 404 minutes of coverage on weekdays and 137 during weekends. Paris during weekends has a very limited number of venues opened compared to weekdays. Similarly, *Food* class has higher average coverage time on weekends than on weekdays in Beijing. *On metropolitan areas, shopping malls tend to concentrate food- and shopping-related venues, which is likely the reason behind those two classes having high coverage time both on weekdays and weekends.* Furthermore, *NL* has 22%, 40%, and 8% higher coverage time on weekends than on weekdays for New York, Tokyo and Beijing, respectively, which is expected due to people having more time to spend on night life-related venues than on weekdays.

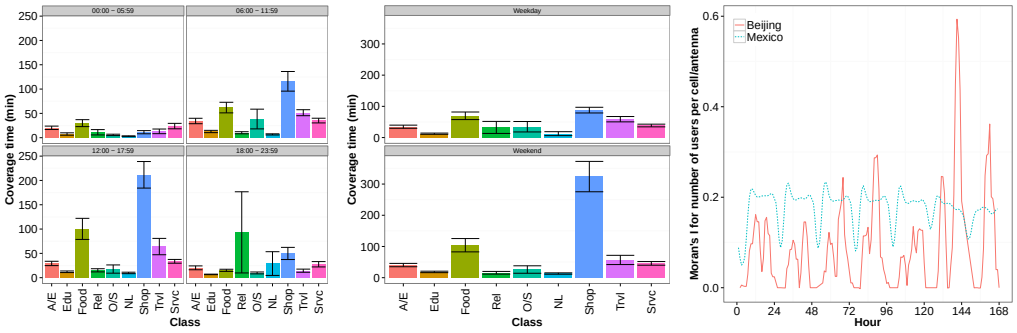
To further comprehend how people together explore the city on a spatiotemporal fashion, we use Moran's I spatial autocorrelation index [39] on snapshots of time. Spatial autocorrelation evaluates the correlation of variables among nearby locations in space and Moran's I can be defined as:

$$I = \frac{N}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(X_i - \bar{X})(X_j - \bar{X})}{\sum_i (X_i - \bar{X})^2} \tag{3}$$

where N is the number of locations, X is the studied random variable, $\bar{X}$ is the mean of X, and w_{ij} is the weight between X_i and X_j. When $I > 0$ there is positive autocorrelation and when $I < 0$

Table 5: Classes and some of their categories

Class	*Abbreviation*	*Categories*
Arts & Entertainment	A/E	aquarium, casino, etc.
Education	Edu	school, university, etc.
Food	Food	cafe, restaurant, etc.
Religion	Rel	church, mosque, etc.
Outdoor & Sports	O/S	gym, stadium, etc.
Night Life	NL	bar and night club
Shopping	Shop	book store, shopping mall, etc.
Travel	Trvl	bus station, subway station, etc.
Services	Srvc	atm, dentist, doctor, etc.



(a) (b) (c)

Figure 9: (Better seen in colors) Coverage time provided by each of the PIs class per (a) period of the day and (b) per weekday (top) and weekends (bottom) in Beijing. (c) Hourly Moran's I for the number of users on the cells (Beijing) and connected to the antennas (Mexico).

negative autocorrelation. We aim to calculate the correlation between the number of people that visits a cell and its surrounding cells with the time. Therefore, in our context, N is the number of cells the city was divided into, X is a cell, w_{ij} is the inverse of the distance between cell centers, X_i is the number of people in cell i, and $\bar{X}$ is the mean of people that visited all cells. To aggregate the temporal aspect on the spatial correlation, we calculate I on snapshots of 1 hour, i. e., we sum up the number of people that visited all the cells during one hour and calculate I. Figure 9(c) shows the hourly Moran's I for the number of users per cell during one week, from 10th to 16th November, 2008 in Beijing and for the number of subscribers per antenna in Mexico. In order to remove noise in the plot curves, it has been smoothed with a sliding window of size 4 hours applying the average. It is interesting to see the periodical behavior that matches with the diurnal activities on both cities. It means that people tend to crowd some popular areas and their neighborhoods on certain hours of the day. It is particularly true around lunch time on metropolitan areas when people go to common areas of restaurants. Moreover, the right end of the curve represents the weekend on both curves, in which the autocorrelation is higher for Beijing and slightly lower on Mexico. The difference lays in the nature of both datasets. First, Mexico dataset has slightly less users making calls during weekends, i.e., there are less people sharing the antennas, thus spatiotemporal correlation is lower than on weekdays. On the other hand, on a mobility dataset as GeoLife, people increase their mobility during weekends, and, as consequence they gather on common leisure areas more than weekdays.

6. Discussion and open issues

Looking ahead, we see a wide range of possible research directions in both short- and long-term bond to the human mobility routine analysis.

Mobility from different nature. Even if we have presented mobility analysis from both CDR- and GPS-based trajectories (refer to Section 5), we envision other possibilities. Services such as Foursquare and Instagram provide large-scale human data collection, which often contains geolocalized information. Similar to a CDR record, user positioning is only available when he performs an activity, which in this context indicates his presence (e. g., *check-in* on Foursquare) at a certain location. This comprises a whole area of research called Participatory Sensing Networks [40, 41]. As with a CDR, a sequence of check-ins represent a user trajectory, e. g., temporally concatenating them, or by inference [42]. Human mobility and routine analysis from this source is still unexplored, to the best of our knowledge. Therefore, we claim the same analysis herein presented could be performed with different natures of datasets, such as, check-ins. Concluding results could be of extreme value to this still not-well-exploited mobility source.

Dataset collection. Analysis presented in our paper rely on the precision in the mobility information given by the dataset. For instance, more precise the geolocalized information is, more we can understand and better analysis/models can be inferred from user's mobility. As a vast spectrum of work can take advantage of dataset analysis in the area of human mobility and network, experiments aiming to collect rich information from users are of enormous value for the research community. Information regarding the user, his device, and surrounding scenario, e. g., fine-grained mobility, battery level, running applications, access points, bluetooth devices, network traffic usage can contribute to a deeper understanding on how we interact with the network and with the environment around. Examples of research projects in this direction are the European MACACO [43] and the French Priva'Mov [44] projects. In general, there is a trade-off between the number of participants in the collected data and the amount of private information collected from them. Due to privacy concerns users are less inclined to participate in projects that collect private information regarding their behavior. Therefore, the limitated availability of valuable datasets is still an impacting factor in the mobility research field.

Forwarding protocols. An important problem on intermittently connected networks is how to couple human mobility patterns with message forwarding algorithms [29]. In this area, mobility has been widely studied when it comes down to encounters among nodes. Contrarily, trajectories behavior could be studied to improve protocols based on store-and-forward late delivery. More related to how far a person routinely moves (see Section 5.2) or to how frequently he/she visits the same places in a city (see Section 5.1) are important characteristics to define the potential of a peer to be chosen to keep a message to be routed. For instance, important routers on a pocket switched network could be taken from the set of few individuals routinely going further than 10 km a day, thus carrying the messages farther.

Routine-based data prefetching. Prefetching has been shown to effectively reduce user perceived latency. An interesting approach might be to merge trajectory information provided by mobility datasets, as the ones used in this work, with datasets or models describing demanded content in urban scenarios. Analysis results could be then used in the design of smart data prefetching approaches. Based on the routine of a user, the prefetching service could store static content of often visited web pages in more adapted locations visited by the user. This content could be downloaded using access points present in the locations he routinely visits. For instance, an

application for mobile phones could download content in advance at home using WiFi and this would be later presented along with content downloaded using the mobile cellular network in real time. This approach could alleviate the usage of the cellular network by shifting to inexpensive networks (such as WiFi) the load of static content, such as images.

Urban planning. Monitoring, distributing, and processing traffic information may enable better strategic planning and encourage better use of public transportation. Applications may take advantage of the driver's routinary behavior in order to improve aspects of the vehicular networks. For instance, traffic information, e. g., accidents, construction sites, traffic jams may be exchanged between vehicles. On a routinary scenario, the human mobility characteristics considered in this work can be used to forecast the situation for the next days and inform the driver, for example, about possible alternative roads. Besides, a service may identify points of interest based on driver's mobility patterns. Identify parking lots and its availability in number of free spots, or suggest the best charging station for electrical vehicles based on the driver's routine and battery conditions. On a collaborative scenario, a carpooling service could suggest people to get or to offer a ride based on regular driver's destinations and passengers interested on ride-sharing.

Customized ad-campaign. A service to advertise products on roadside signs may merge different sources of information in order to display targeted marketing. By crossing information from the people's routinary trajectories (cf. Section 5.1), interests, and traffic conditions, a service could present ads that match people's interest on a certain area of the city likely having him on its vicinity.

7. Conclusions

In this paper, we have made an extensive analysis of human mobility on several categories of datasets describing mobility in several large metropolitan areas, in order to unveil common aspects present in human mobility. We firstly presented our system model, which unifies different datasets into a common representation of urban scenario. Then, we presented analyses on the visiting patterns to PIs. Results unveiled a clear repetitivity on people's visiting behavior. Additionally, we have proposed a metric to measure the repetitiveness of people's visits to PIs. Next, we have evaluated displacement in their trajectories. The main conclusions are two: people have a tendency to use shortest-path when moving around, and their mobility is confined, i. e., displacement is generally limited to 10 km. Finally, we have zoomed out from the per-user analysis to a spatial autocorrelation. It shows that the regular patterns found in human mobility are not restricted by the scale of the dataset, since GeoLife and Telco datasets are orders of magnitude distant on the number of users. Moreover, they represent mobility in different ways and in different granularities. Still, their spatial autocorrelation shows the same routinary regular behavior.

References

[1] K. C. Seto, M. Fragkias, B. Gneralp, M. K. Reilly, A meta-analysis of global urban land expansion, PLoS ONE 6 (8) (2011) e23777. doi:10.1371/journal.pone.0023777.
URL http://dx.doi.org/10.1371%2Fjournal.pone.0023777

[2] D. S. Hamermesh, Routine, Working Paper 9440, National Bureau of Economic Research (January 2003).
URL http://www.nber.org/papers/w9440

[3] J. A. Throgmorton, B. Eckstein, Desire lines: The Chicago area transportation study and the paradox of self in post-war America (Nov. 2000).
URL http://www.nottingham.ac.uk/3cities/throgeck.htm

[4] D. Brockmann, L. Hufnagel, T. Geisel, The scaling laws of human travel, Nature 439 (7075) (2006) 462–465. doi:10.1038/nature04292.

[5] A. Socievole, F. De Rango, A. Caputo, Wireless contacts, facebook friendships and interests: Analysis of a multi-layer social network in an academic environment, in: Wireless Days (WD), 2014 IFIP, 2014, pp. 1–7. doi:10.1109/WD.2014.7020819.

[6] A. Galati, K. Djemame, C. Greenhalgh, A mobility model for shopping mall environments founded on real traces, Networking Science 2 (1-2) (2013) 1–11. doi:10.1007/s13119-012-0011-1.
URL http://dx.doi.org/10.1007/s13119-012-0011-1

[7] N. Kiukkonen, B. J., O. Dousse, D. Gatica-Perez, L. J., Towards rich mobile phone datasets: Lausanne data collection campaign, in: Proc. ACM Int. Conf. on Pervasive Services, 2010.

[8] Y. Zheng, X. Xie, W.-Y. Ma, Geolife: A collaborative social networking service among user, location and trajectory, IEEE Data Eng. Bull. 33 (2) (2010) 32–39.

[9] M. Piorkowski, N. Sarafijanovic-Djukic, M. Grossglauser, A parsimonious model of mobile partitioned networks with clustering, in: Communication Systems and Networks and Workshops, 2009. COMSNETS 2009. First International, IEEE, 2009, pp. 1–10. doi:10.1109/comsnets.2009.4808865.
URL http://dx.doi.org/10.1109/comsnets.2009.4808865

[10] R. Amici, M. Bonola, L. Bracciale, A. Rabuffi, P. Loreti, G. Bianchi, Performance assessment of an epidemic protocol in {VANET} using real traces, Procedia Computer Science 40 (0) (2014) 92 – 99, fourth International Conference on Selected Topics in Mobile & Wireless Networking (MoWNet 2014). doi:http://dx.doi.org/10.1016/j.procs.2014.10.035.
URL http://www.sciencedirect.com/science/article/pii/S1877050914014021

[11] H. Wang, F. Calabrese, G. Di Lorenzo, C. Ratti, Transportation mode inference from anonymized and aggregated mobile phone call detail records, in: Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on, 2010, pp. 318–323. doi:10.1109/ITSC.2010.5625188.

[12] C. Song, Z. Qu, N. Blumm, A.-L. Barabási, Limits of Predictability in Human Mobility, Science 327 (5968) (2010) 1018–1021. doi:10.1126/science.1177170.
URL http://dx.doi.org/10.1126/science.1177170

[13] A.-L. Barabasi, The origin of bursts and heavy tails in human dynamics, Nature 435 (2005) 207.
URL http://www.citebase.org/abstract?id=oai:arXiv.org:cond-mat/0505371

[14] G. Ramos-Fernández, J. Mateos, O. Miramontes, G. Cocho, H. Larralde, B. Ayala-Orozco, Lévy walk patterns in the foraging movements of spider monkeys (Ateles geoffroyi), Behavioral Ecology and Sociobiology 55 (3) (2004) 223–230. doi:10.1007/s00265-003-0700-6.
URL http://dx.doi.org/10.1007/s00265-003-0700-6

[15] R. P. D. Atkinson, C. J. Rhodes, D. W. MacDonald, R. M. Anderson, Scale-free dynamics in the movement patterns of jackals, Oikos 98 (1) (2002) pp. 134–140.
URL http://www.jstor.org/stable/3547620

[16] G. M. Viswanathan, V. Afanasyev, S. V. Buldyrev, E. J. Murphy, P. A. Prince, H. E. Stanley, Lévy flight search patterns of wandering albatrosses, Nature 381 (6581) (1996) 413–415. doi:10.1038/381413a0.
URL http://dx.doi.org/10.1038/381413a0

[17] M. F. Shlesinger, J. Klafter, G. Zumofen, Above, below and beyond Brownian motion, Am. J. Phys. 67 (12).

[18] I. Rhee, M. Shin, S. Hong, K. Lee, S. Chong, On the levy-walk nature of human mobility: Do humans walk like monkeys? (2007).

[19] M. C. Gonzalez, C. A. Hidalgo, A.-L. Barabasi, Understanding individual human mobility patterns, Nature 453 (7196) (2008) 779–782. doi:10.1038/nature06958.

[20] C. Song, T. Koren, P. Wang, A.-L. Barabási, Modelling the scaling properties of human mobility, Nature Physics 6 (10) (2010) 818–823. doi:10.1038/nphys1760.
URL http://dx.doi.org/10.1038/nphys1760

[21] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, J. Rowland, A. Varshavsky, Identifying important places in people's lives from cellular network data, in: Proceedings of the 9th International Conference on Pervasive Computing, Pervasive'11, Springer-Verlag, Berlin, Heidelberg, 2011, pp. 133–151.
URL http://dl.acm.org/citation.cfm?id=2021975.2021988

[22] R. Becker, R. Caceres, K. Hanson, J. Loh, S. Urbanek, A. Varshavsky, C. Volinsky, A tale of one city: Using cellular network data for urban planning, IEEE Pervasive Computing 10 (4) (2011) 18–26.

[23] Y. Liang, X. Zhou, B. Guo, Z. Yu, Understanding the regularity and variability of human mobility from geo-trajectory, in: Web Intelligence and Intelligent Agent Technology (WI-IAT), 2012 IEEE/WIC/ACM International Conferences on, Vol. 1, 2012, pp. 409–414. doi:10.1109/WI-IAT.2012.163.

[24] S. Motahari, H. Zang, P. Reuther, The impact of temporal factors on mobility patterns, in: System Science (HICSS), 2012 45th Hawaii International Conference on, 2012, pp. 5659–5668. doi:10.1109/HICSS.2012.572.

[25] E. Cho, S. A. Myers, J. Leskovec, Friendship and mobility: User movement in location-based social networks, in: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '11, ACM, New York, NY, USA, 2011, pp. 1082–1090. doi:10.1145/2020408.2020579.
URL http://doi.acm.org/10.1145/2020408.2020579

[26] A. Noulas, S. Scellato, R. Lambiotte, M. Pontil, C. Mascolo, A tale of many cities: universal patterns in human

urban mobility, PloS one 7 (5) (2012) e37027.
URL https://scholar.google.de/scholar.bib?q=info:gf9wW7OT2uQJ:scholar.google.com/&output=citation&hl=de&ct=citation&cd=0

[27] V. Conan, J. Leguay, T. Friedman, Characterizing pairwise inter-contact patterns in delay tolerant networks, in: Proceedings of the 1st International Conference on Autonomic Computing and Communication Systems, Autonomics '07, ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), ICST, Brussels, Belgium, Belgium, 2007, pp. 19:1–19:9.
URL http://dl.acm.org/citation.cfm?id=1365562.1365588

[28] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, J. Scott, Impact of human mobility on opportunistic forwarding algorithms, Mobile Computing, IEEE Transactions on 6 (6) (2007) 606–620. doi:10.1109/TMC.2007.1060.

[29] P. Hui, A. Chaintreau, J. Scott, R. Gass, J. Crowcroft, C. Diot, Pocket switched networks and human mobility in conference environments, in: Proceedings of the 2005 ACM SIGCOMM Workshop on Delay-tolerant Networking, WDTN '05, ACM, New York, NY, USA, 2005, pp. 244–251. doi:10.1145/1080139.1080142.
URL http://doi.acm.org/10.1145/1080139.1080142

[30] A. Natarajan, M. Motani, V. Srinivasan, Understanding urban interactions from bluetooth phone contact traces, in: Proceedings of the 8th International Conference on Passive and Active Network Measurement, PAM'07, Springer-Verlag, Berlin, Heidelberg, 2007, pp. 115–124.
URL http://dl.acm.org/citation.cfm?id=1762888.1762904

[31] Y. Wang, B. Krishnamachari, T. Valente, Findings from an empirical study of fine-grained human social contacts, in: Wireless On-Demand Network Systems and Services, 2009. WONS 2009. Sixth International Conference on, 2009, pp. 153–160. doi:10.1109/WONS.2009.4801861.

[32] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, J. Scott, Pocket Switched Networks: Real-world mobility and its consequences for opportunistic forwarding, Tech. Rep. UCAM-CL-TR-617, University of Cambridge, Computer Laboratory (Feb. 2005).
URL http://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-617.pdf

[33] Z.-R. Peng, Urban transportation strategies in chinese cities and their impacts on the urban poor, Transportation Research Board 85th Annual Meeting (2005) 14.

[34] J. Lin, Bicycles in beijing, project: Transportation (Green Design and the City).

[35] S. Wang, J. Sun, C. Shao, F. Wang, Evaluation on Vehicle Restriction Measure in Beijing, Ch. 39, pp. 433–443. doi:10.1061/41123(383)40.

[36] Z. Xin, Subway line 6 to start running in december (November 2012).
URL http://www.chinadaily.com.cn/beijing/2012-11/26/content_15998073.htm

[37] ISO, ISO 8601:1988. Data elements and interchange formats — Information interchange — Representation of dates and times, 1988, see also 1-page correction, ISO 8601:1988/Cor 1:1991.
URL http://www.iso.ch/cate/d26780.html

[38] D. B. Carr, A. R. Olsen, D. White, Hexagon mosaic maps for displaying univariate and bivariate geographical data, Cartography & Geographical Information Systems 19 (1992) 228–236.

[39] P. A. P. Moran, Notes on Continuous Stochastic Phenomena, Biometrika 37 (1/2) (1950) 17–23. doi:10.2307/2332142.
URL http://dx.doi.org/10.2307/2332142

[40] T. H. Silva, P. O. S. Vaz de Melo, J. M. Almeida, J. Salles, A. A. F. Loureiro, A comparison of foursquare and instagram to the study of city dynamics and urban social behavior, in: Proceedings of the 2Nd ACM SIGKDD International Workshop on Urban Computing, UrbComp '13, ACM, New York, NY, USA, 2013, pp. 4:1–4:8. doi:10.1145/2505821.2505836.
URL http://doi.acm.org/10.1145/2505821.2505836

[41] T. Silva, P. Vaz De Melo, J. Almeida, A. Loureiro, Challenges and opportunities on the large scale study of city dynamics using participatory sensing, in: Computers and Communications (ISCC), 2013 IEEE Symposium on, 2013, pp. 000528–000534. doi:10.1109/ISCC.2013.6755000.

[42] L.-Y. Wei, Y. Zheng, W.-C. Peng, Constructing popular routes from uncertain trajectories, in: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12, ACM, New York, NY, USA, 2012, pp. 195–203. doi:10.1145/2339530.2339562.
URL http://doi.acm.org/10.1145/2339530.2339562

[43] Mobile context-Adaptive CAching for COntent-centric networking.
URL https://macaco.inria.fr

[44] Priva'Mov - mobilité et vie privée.
URL http://liris.cnrs.fr/privamov/