On The Sampling Frequency of Human Mobility

Panagiota Katsikouli, Aline Carneiro Carneiro Viana, Marco Fiore, Alberto Tarable

On The Sampling Frequency of Human Mobility

Panagiota Katsikouli, Aline Carneiro Viana, Marco Fiore, Alberto Tarable

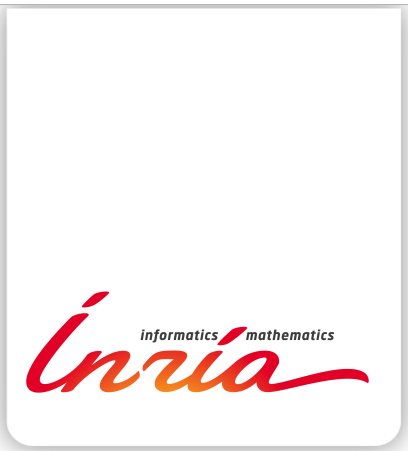
TECHNICAL REPORT

N° 487

February 2017

Project-Teams INFINE

ISSN 0249-0803 ISRN INRIA/RT--487--FR+ENG



On The Sampling Frequency of Human Mobility

Panagiota Katsikouli*, Aline Carneiro Viana†, Marco Fiore‡, Alberto Tarable‡

Project-Teams INFINE

Technical Report n° 487 — February 2017 — 19 pages

Abstract: Recent studies have leveraged tracking techniques based on positioning technologies to discover new knowledge about human mobility. These investigations have revealed, among others, a high spatiotemporal regularity of individual movement patterns. Building on these findings, we aim at answering the question *"at what frequency should one sample individual human movements so that they can be reconstructed from the collected samples with minimum loss of information?"*. Our quest for a response leads to the discovery of *(i)* seemingly universal spectral properties of human mobility, and *(ii)* a linear scaling law of the localization error with respect to the sampling interval. Our findings are based on the analysis of fine-grained GPS trajectories of 119 users worldwide. The applications of our findings are related to a number of fields relevant to ubiquitous computing, such as energy-efficient mobile computing, location-based service operations, active probing of subscribers' positions in mobile networks and trajectory data compression.

Key-words: Human mobility, spatiotemporal trajectories, sampling frequency, sampling interval

This work was supported by the EU FP7 ERANET program under grant CHIST-ERA-2012 MACACO.

* University of Edinburgh
† INRIA
‡ CNR - IEIIT

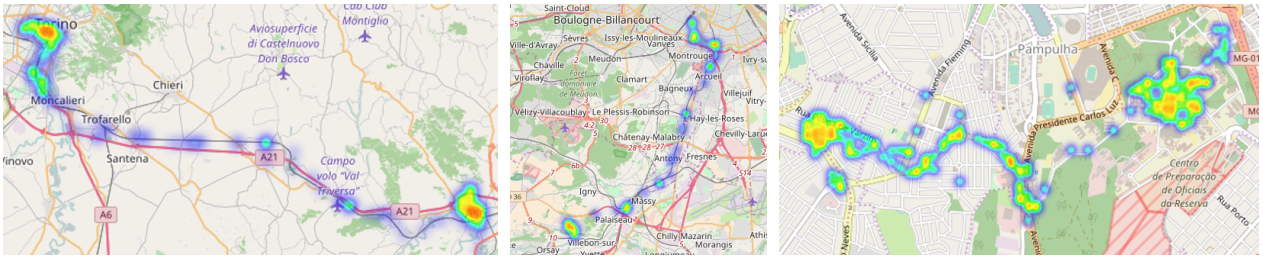
RESEARCH CENTRE
SACLAY – ÎLE-DE-FRANCE

1 rue Honoré d'Estienne d'Orves
Bâtiment Alan Turing
Campus de l'École Polytechnique
91120 Palaiseau

Investigations sur la fréquence d'échantillonnage de la mobilité

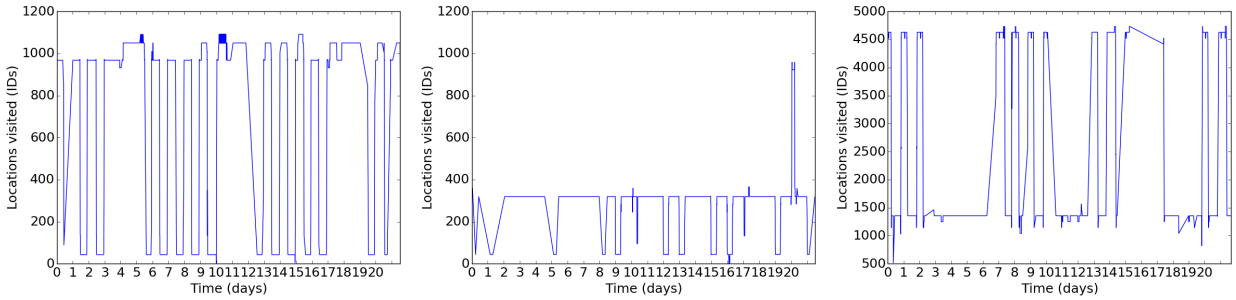
Résumé : Des études récentes ont mis à profit des techniques de suivi basées sur des technologies de positionnement pour étuder la mobilité humaine. Ces recherches ont révélé, entre autres, une grande régularité spatio-temporelle des mouvements individuels. Sur la base de ces résultats, nous visons à répondre à la question «à quelle fréquence doit-on échantillonner les mouvements humains individuels afin qu'ils puissent être reconstruits à partir des échantillons recueillis avec un minimum de perte d'information? Notre recherche d'une réponse à cette question nous a conduit à la découverte de (i) propriétés spectrales apparemment universelles de la mobilité humaine, et (ii) une loi de mise à l'échelle linéaire de l'erreur de localisation par rapport à l'intervalle d'échantillonnage. Nos résultats sont basés sur l'analyse des trajectoires GPS de 119 utilisateurs dans le monde entier. Les applications de nos résultats sont liées à un certain nombre de domaines pertinents pour l'informatique omniprésente, tels que l'informatique mobile économe en énergie, les opérations de service basées sur l'emplacement, le sondage actif des positions des abonnés dans les réseaux mobiles et la compression des données de trajectoire.

Mots-clés : Mobilité humaine, trajectoires spatio-temporelles, fréquence d'échantillonnage, intervalle d'échantillonnage



(a) User A (b) User B (c) User C

Figure 1: Heatmaps of locations visited by three distinct users during three weeks: humans tend to commute between a limited set of specific locations. Figure best seen in color.



(a) User A (b) User B (c) User C

Figure 2: Location time series for three distinct users during three weeks: humans tend to revisit locations in a periodic fashion. The visited locations here are mapped to a unidimensional space through discretization on a regular grid with step 50 meters and assignment of a unique sequential identifier.

1 Introduction

Over the past few years, the pervasive usage of smart devices and location-tracking systems has made it possible to study and understand human mobility at unprecedented scales. An important feature that was found to characterize human mobility is regularity; we tend to follow the same patterns over and over, and we do so in ways that are clearly periodic [7]. Movement regularity is easily found in the movements of most individuals. As an example, consider Fig. 1(a)-(c), which show heatmaps of the locations visited by three random users in the datasets employed in our study and presented in Sec. 2. Although each plot conveys three weeks of data, a clear set of frequently visited places emerges for all users, along with systematic paths connecting them. Likewise, Fig. 2(a)-(c) illustrate the temporal dimension of regularity for the same users; movement periodicity becomes apparent in the time series of the visited locations.

In this paper, we investigate whether the regularity of human mobility results in the possibility of sampling individual movements at reduced frequencies, while retaining a vast portion – if not all – of the original full-detail trajectories. Intuitively, periodic visits to a limited set of important places through repeated routes may be captured with a reduced sampling effort than required for, e.g., a completely random mobility.

Note that our problem is different from the widely addressed subjects of simplifying, com-

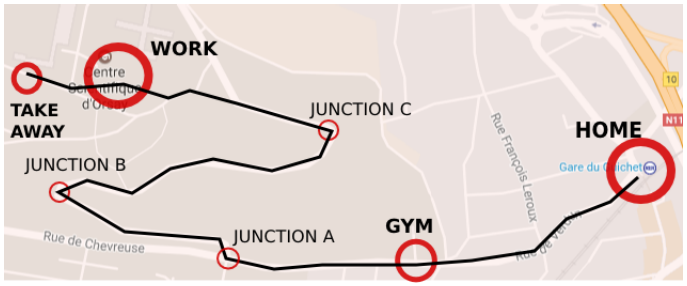


Figure 3: Toy example of individual mobility trajectory. Labels denote important locations or turning points. The size of the circle around each location is proportional to the amount of time spent there by the target user.

pressing or approximating spatial trajectories [9, 16]. In those cases, the objective is to maintain the trajectory shape, while we aim at preserving the temporal dimension of movements as well. Consider the toy example in Fig. 3, where a user leaves home, trains at the gym before reaching work and, later on, goes to a take-away restaurant. The shape of the spatial trajectory could be well approximated as the sequence of home, turning junctions B and C, and take-away locations: map-matching based on these cardinal points would allow describing the whole movement. However, individuals visit places for a purpose, carrying out activities that have different duration. For instance, in our example, the size of the circle around each location is proportional to the amount of time spent there. Our purpose is then to recreate the complete original mobility of the individual, including these temporal features.

Our problem is also different from sampling to detect important locations [23, 3]. In the example of Fig. 3, important location detection is solved by sampling the trajectory so as to be able to model the original distribution of time spent at home, work, gym, and take-away. However, approaches for the detection of such frequently visited places ignore the time ordering of visits, and do not capture transitions between frequent locations. Instead, our holistic perspective accounts for all these characteristics.

Overall, our problem is equivalent to posing the question *"at what frequency should one sample individual human movements so that they can be reconstructed from the collected samples with minimum loss of information?"*. As we deal with information encoded in sequences (of movements), information theory is a sensible instrument to answer the question. Interestingly, information theory has already been employed in a seminal work on human mobility analysis to demonstrate the low entropy of individual movements, and hence their high predictability [22]. These previous results may suggest that trajectories can be fully described with limited sampled data. However, in Sec. 3 we show that this is not the case, by adopting a different approach based on signal processing. To that end, we consider mobility patterns as signals over time, and carry out a spectral analysis of human mobility. The spectra of the movements of 119 different individuals tend to have very similar, flat shapes that suggest the absence of convenient sampling frequency thresholds –even specific to single users– beyond which the error in the reconstructed trajectories drops significantly.

Stimulated by this finding, we carry out a quantitative analysis of the user localization error in movements reconstructed from regular sampling at different periodicities. Results in Sec. 4, based on our reference datasets of GPS trajectories, unveil a linear scaling law of the error with respect to the sampling interval. This law corroborates the outcome of the spectral analysis, and has significant practical implications, as it controls the trade-off between accuracy and cost of measurements of human mobility. Therefore, our findings have applications in a number of fields,

including energy-efficient mobile computing, location-based service operation, active probing of subscribers' positions in mobile networks, and trajectory data compression. These usages are unfolded in Sec. 6, which also concludes the paper.

2 Reference dataset

Our study employs real-world individual mobility data from three different sources.

- MACACO data was collected between July 2014 and December 2016 as part of the activities of the European collaborative project of the same name, funded by the EU CHIST-ERA program[1]. A dedicated monitoring application, running on smartphones of volunteers in several countries in Europe and South America, recorded GPS positioning information at regular time intervals, typically ranging from one to five minutes. GPS logs were then uploaded to the project servers in France. Due to data privacy regulations in that country, the data is not publicly available.

- OpenStreetMap (OSM) data was collected by volunteers who recorded and uploaded their trajectories as a contribution to the OSM database[2]. The OSM project is a global crowd-sourcing initiative, aiming at mapping the whole world surface thanks to the activity of a vast user community. GPS traces uploaded by participants typically feature 1-Hz frequency, and are a useful tool to validate urban or rural road layouts in the maps. They are freely available on the official OSM project website.

- Geolife data was collected in Beijing by Microsoft Research Asia from April 2007 to August 2012 [26]. It consists of GPS trajectories recorded through different GPS loggers and smartphone apps. Although sample rates vary significantly across users and time periods, the vast majority of GeoLife positioning data is recorded at intervals from one to five seconds. GeoLife traces are publicly available on the official website.

GPS trajectories cover notably different geographical and temporal spans, even for data coming from the same source. Depending on the user, movements can range from single cities to multiple continents, and can last from days to years. Moreover, the quality of the data for a single user is typically very heterogeneous over time, with periods of days or weeks where GPS logs are erroneous or completely absent. In order to build a consistent reference dataset, we segmented the mobility traces of all users into one-week trajectories, and analysed them separately[3]. During each week, we also bounded the mobility of each individual to the regions where the activity is mainly concentrated. Bounding regions avoid biases introduced by singularities, such as, e.g., international journeys performed just once in months; depending on the user and week, these regions span from 400 to $3,000$ km^2, and users spend from 85% to 100% of their time within them. Clearly, bounding regions can also create temporal gaps in the weekly traces, whenever users visit areas outside them.

We then filtered the one-week geographically-bounded trajectories based on their quality. Specifically, we only retained trajectories that contain complete GPS records in at least six out of seven distinct week days. Ultimately, our reference dataset is composed of 1052 weeks of mobility of 119 different individuals. Tab. 1 provides a break down of these numbers on a per-source basis. A legitimate question is whether the data is dominated by a few users, i.e.,

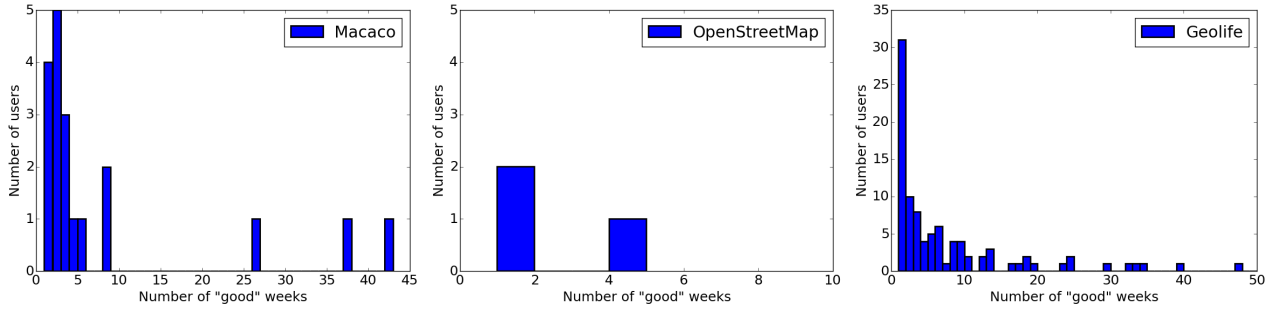
[1] `https://macaco.inria.fr/macacoapp/`

[2] `https://www.openstreetmap.org`

[3] The rationale behind our choice is that many human activities have been shown to have a weekly periodicity [10, 21]. Hence, using one-week GPS logs let us capture both repetitiveness and regularity of human mobility.

Dataset	Users	Weeks
MACACO	19	164
OpenStreetMap	4	7
GeoLife	96	881

Table 1: Number of users and weeks in our reference dataset from each data source.



(a) (b) (c)

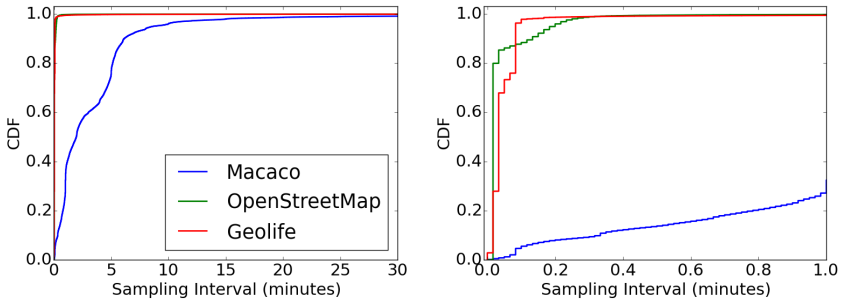
Figure 4: Distribution of one-week trajectories in our reference dataset, separated by data source. (a) MACACO. (b) OpenStreetMap. (c) GeoLife.

if the majority of weeks refers in fact to a limited set of users, which could bias the analysis. Fig. 4(a)-(c), portraying the distribution of the number of weekly trajectories, show that this is not the case. Indeed, the vast majority of users contribute one to ten weeks of movement data, and the few users who exceed that range provide around 50 weeks of mobility at most. Overall, our reference dataset encompasses a quite diverse base of individuals.

As mentioned above, the different techniques employed to collect the GPS positioning information lead to uneven recording intervals across, and even within, the original data sources. In addition to this, weekly trajectories have temporal gaps due to offline GPS receivers, interruptions in the data collection service, or users travelling outside the bounding regions we introduce. Fig. 5(a) shows the cumulative distribution function (CDF) of the sampling intervals observed in all one-week trajectories of our reference dataset. We remark that in almost all cases intervals are shorter than 15 minutes. More precisely, in trajectories from the lower-granularity MACACO data, 40% and 65% of GPS points are separated by less than 1 and 5 minutes, respectively. In trajectories from the OSM and GeoLife data, 90% to 95% of points are less than 10 seconds apart, as highlighted in Fig. 5(b). We argue that the sampling intervals in the one-week trajectories of our reference dataset sufficiently capture human movements and hence, we will consider them as our ground-truth in the remainder of the study.

3 Spectral analysis of human mobility

From a spectral analysis viewpoint, answering the question we posed in the Introduction, *"at what frequency should one sample individual human movements so that they can be reconstructed from the collected samples with minimum loss of information?"*, is equivalent to considering human movements as a signal in time, and studying its spectrum in frequency.



(a) (b)

Figure 5: Distribution of the sampling intervals in one-week trajectories in our reference dataset, separated by data source. (a) Full sampling interval span. (b) Zoom on sub-minute sampling intervals. Figure best seen in color.

3.1 Individual movement as a unidimensional signal

First, we need to transform individual GPS trajectories into unidimensional time series. Even when neglecting altitude information, points in geographical trajectories are obviously bidimensional. We carried out an extensive evaluation of approaches to reduce bidimensional movements to unidimensional signals, using approximated measures such as the velocity or the relative displacement from the centre of mass, and transformations such as enumeration of discretized locations in the Hilbert space. However, all of the techniques we tested introduced an exceeding amount of noise in the process, disrupting individual movements or introducing unrealistic jumps in the mobility of users.

We then opted for a parallel study of the two dimensions of the geographical space, by considering them in isolation. Instead of using the absolute values of latitude and longitude as unidimensional time series, we replace them with the signed latitude and longitude displacements from the corresponding centre of mass of the one-week trajectory. Formally, the displacements of the i-th point in a trajectory, denoted as $\widetilde{lat}_i$ and $\widetilde{lon}_i$ respectively, are computed as

$$\widetilde{lat}_i = lat_i - \frac{1}{n}\sum_{j=0}^{n} lat_j \qquad \text{and} \qquad \widetilde{lon}_i = lon_i - \frac{1}{n}\sum_{j=0}^{n} lon_j \tag{1}$$

where lat_j, lon_j are the latitude and longitude coordinates of the j-th GPS point, and n is the number of points in the trajectory. Other than making time series more easily comparable and readable across users and weeks, the transformations in (1) have the property of generating zero-mean signals whose frequency spectra have null constant components. Graphical examples of the unidimensional representations of individual movements are in Fig. 7(a)-(c), for three random one-week trajectories: the plots portray latitude (top) and longitude (bottom) displacement time series.

By considering the transformation above on the two geographical dimensions in isolation, we do not introduce errors; yet, we may lose properties that only emerge when the two dimensions are considered jointly. To verify whether such a problem exists, we analysed the correlation between the isolated latitude or longitude displacements and the actual travelled distance in the bidimensional space. Tab. 2 shows the correlation coefficients, separated depending on the data source; an illustrative representation is provided in Fig. 6, for all one-week trajectories extracted

Dataset	Latitude	Longitude
MACACO	0.55	0.84
OpenStreetMap	0.53	0.7
GeoLife	0.57	0.62

Table 2: Correlation coefficients of latitude and longitude with respect to the actual travelled distance in the bidimensional space, separated by data source.

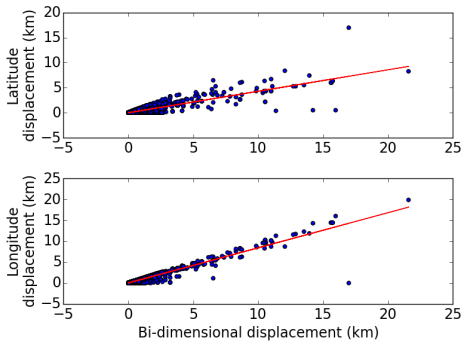


Figure 6: Scatterplot of latitude (top) and longitude (bottom) with respect to the actual travelled distance in the bidimensional space, for trajectories from the MACACO data.

from the MACACO data. We observe consistently good correlations in all cases, which lets us conclude that both dimensions, when taken singularly, still provide decent approximations of the overall mobility. Interestingly, the correlation is always stronger for longitude than for latitude, indicating that participants to all of the initiatives that gathered the GPS data have a tendency to move along an East-West axis rather than along a South-North one.

3.2 Frequency spectra of human mobility

We apply the Fast Fourier Transform (FFT) in order to compute the spectral representation of the finite-length sequences that represent the one-week latitude and longitude displacement signals. Let $x[n]$, $n = 0, \ldots, N-1$ be a sequence (or signal) of length N. Then, its spectral counterpart,

$$X[k] = \sum_{n=0}^{N-1} x[n] \mathrm{e}^{-j2\pi kn/N}, \quad k = 0, \ldots, N-1$$

can be computed with $\mathcal{O}(N \log N)$ operations. It is well known that $X[k]$ is useful to highlight periodic components in the sequence $x[n]$. In particular, if the sequence $x[n]$ is obtained by sampling a continuous-time periodic signal $x(t)$ within a single period, then $X[k]$ will be equal to the sequence of coefficients of the Fourier series of $x(t)$.

The frequency spectrum of a signal yields information about the sampling frequency needed to reconstruct the original time series with a small error. For an ideal signal, whose spectrum drops to zero after some frequency threshold f_s (the bandwidth of the signal), the Nyquist—Shannon sampling theorem guarantees that a sampling rate $2f_s$ is enough to allow a lossless reconstruction of the original signal from its samples. For practical signals, the spectrum is

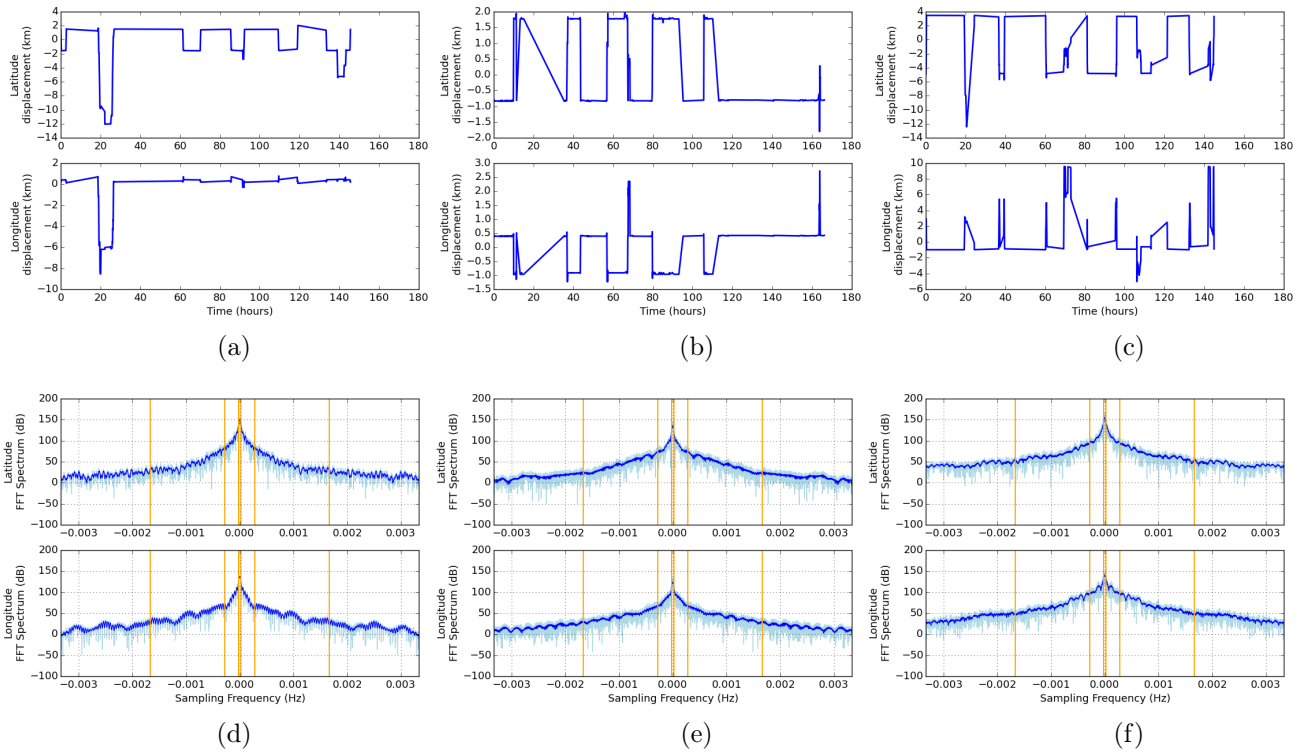


Figure 7: Unidimensional movement signals and associated frequency spectra for three random one-week trajectories. Figure best seen in color.

not strictly limited, but it features limited amounts of noise. In those cases, the spectrum is mostly concentrated within a finite support and shows a negligible amount of power beyond the frequency threshold; again, sampling at a rate twice the threshold, allows reconstructing the original signal with minimum error.

Fig. 7(d)-(f) show the spectra of the signals in Fig. 7(a)-(c). The original spectra are in light blue, while a moving-average that better displays the overall trends is in dark blue. Vertical orange lines outline the frequencies that correspond to sampling intervals of 10 minutes (farthest from the central frequency), 1 hour and 12 hours (closest to the central frequency), respectively. We make the two following important remarks: *(i)* despite the diversity of the latitude and longitude displacement time series across the different one-week trajectories, all spectra have very similar shapes; *(ii)* the shapes do not show evidence of any bandwidth threshold beyond which the spectra become clearly negligible, making it impossible to identify an operational point for effective sampling. We can explain both facts remarked above by noticing that the time series show very steep transitions and deep spikes, so that the resulting spectrum shows a slow decay for high frequencies.

We found the observations above to hold for the vast majority of one-week trajectories. Illustrative examples are provided in Fig. 8, which shows a gallery of representative spectra of signals from our reference dataset. Considering the heterogeneity of our user base, we hypothesize that the features identified above could be a universal property of human mobility spectra.

Our conclusion is that, although it shows some clear periodicity [22], human mobility is in

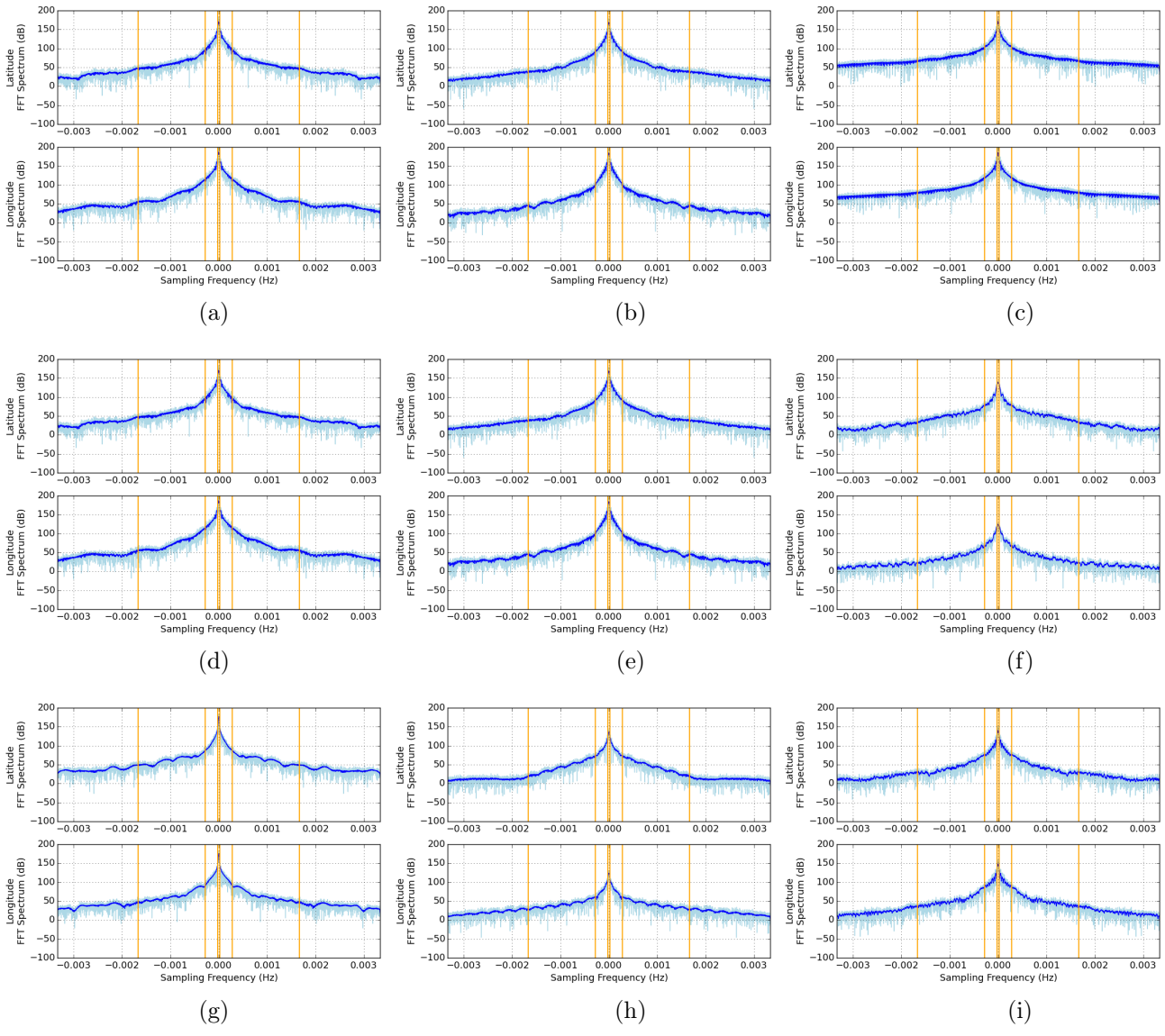


Figure 8: Frequency spectra for 9 representative one-week trajectories. Figure best seen in color.

fact a sequence of long periods where individuals are almost static and fast transitions between such important locations. While positions during stationary time intervals contribute to low-frequency spectral components and are hence easily captured by a sparse sampling, travelling causes discontinuities in the mobility signal and is much harder to sample. As a result, considering that sampling at higher frequencies has a cost but leads to a better quality of the reconstructed signal, the spectra do not reveal whether, e.g., collecting samples at every 10 minutes is obviously more efficient than doing the same at every hour. Although disappointing in a sense, this outcome calls forth for an extensive quantitative analysis of the exact trade-off between the quality and cost of sampling in the context of human mobility. We address this aspect in the next section.

4 Quantitative analysis of human mobility sampling

We perform an experimental analysis on our reference dataset of one-week trajectories, investigating the impact of different sampling frequencies on the quality of the mobility reconstructed from the collected samples. To this end, we create downsampled versions of the trajectories, using a wide range of representative sampling intervals, from 10 minutes to 12 hours. Longer intervals are deemed unreasonable for one-week-long time series. We then create a reconstructed version of the complete trajectories by linearly interpolating the samples, and assess how such reconstructed trajectories compare to the original ones.

We measure the error in retrieving a complete individual trajectory from sampled data by using the average Haversine distance. Given two points on Earth's surface, $p_a = (lat_a, lon_a)$ and $p_b = (lat_b, lon_b)$, we define $\Delta[lat] = lat_b - lat_a$, and $\Delta[lon] = lon_b - lon_a$. The Haversine distance between the two points is then

$$D(p_a, p_b) = R \cdot 2 \cdot atan2\left(\sqrt{\phi}, \sqrt{(1-\phi)}\right), \tag{2}$$

where $\phi = \sin^2(\Delta[lat]/2) + \cos(lon_a) \cdot \cos(lat_b) \cdot \sin^2(\Delta[lon]/2)$, and $R = 6,371km$ is the Earth's radius. The average Haversine distance of a one-week trajectory is the mean of all Haversine distances between the points of the reconstructed and original mobility recorded at the same time instant.

Fig. 9 shows the evolution of the average Haversine error against the sampling interval, for a representative set of nine individuals in our reference dataset. Each plot presents results for all of the one-week trajectories of a specific user: as multiple one-week trajectories are aggregated in every plot, we outline the mean (dots), 25-75% quantiles (dark blue region) and 10-90% quantiles (light blue region) of the error measured over all involved trajectories.

A surprisingly clear linear relationship characterizes all curves. Fittings on a simple linear model, depicted as solid lines in Fig. 9, show an excellent match for all the considered users. Fig. 10 provides a zoom on sub-hour sampling intervals. The linear fitting remains a decent approximation even for very sparse sampling intervals, i.e., low frequencies. In fact, the linearity of the relation between the Haversine distance and the sampling interval is a common trait of all individuals in our user base. Fig. 11 portrays the Root Mean Square Error (RMSE) of the linear fitting on the mean values of the average Haversine distance versus sampling intervals that range between 10 minutes and 12 hours. The probability mass of the distribution is below 250 meters, which is a very reasonable RMSE for people travelling tens of km per day.

We also highlight that the only parameter of the fitting curve $y = ax$, i.e., the slope a, has an important physical meaning: it characterizes the ratio between the average Haversine distance and the sampling interval, or, equivalently, it explains the mean additional error of the reconstructed trajectory when increasing the time between samples. Hence, it can be measured

in meters per minute (m/min). When looking at its value, we remark that it is not identical across users: the plots in Fig. 9 also report the equation of the linear fit, and we can remark some diversity in the distance/interval ratio. Therefore, we study this heterogeneity in Fig. 12, which portrays the CDF of the distance/interval ratio associated to all individuals in our reference dataset. There, we note that 90% of users have slopes that are uniformly distributed between 1 and 4 m/min: although some diversity exists in the slope, it is well within the same order of magnitude for all our user base.

Summarizing our findings, we assert that the average error incurred by trajectories reconstructed from periodic samples scales linearly with the time interval that separates the samples. Note that this result is well aligned with the outcome of our spectral analysis in Sec. 3: the linearity of the relationship between error and sampling interval explains the absence of an operational point for the effective sampling of human movements. In addition, we find the linear scaling law to be characterized by a comparable parameter, i.e., the error-to-interval ratio, across all our user base; depending on the individual, the error typically grows 1 to 4 meters when adding one minute to the inter-sample time. Such a reduced range on a simple and seemingly general scaling law provides a very useful reference for practitioners, as expounded in Sec. 6.

5 Related Work

Reducing the number of points on a curve or line is a research topic going as far back as the $70s$. It has met with recent research revival after the immense amount of GPS data being generated every day by mobile devices. The majority of the works on GPS trajectory simplification or compression aim at reducing the size of pre-recorded GPS trajectories, while maintaining the shape of the trace.

The best-known method for line simplification, initially introduced to approximate lines in cartography and recently used for trajectory simplification, is the Douglas-Peucker [4] algorithm. The algorithm assumes knowledge of the complete trace and starts producing a line connecting the beginning and ending of the initial curve. Recursively, in each round, it selects the point that is the furthest away from the current simplification. When the error of approximation is below a user-defined threshold, the algorithm terminates. Douglas-Peucker variants that give approximations at different resolutions have been proposed in [1, 15]. Basic offline algorithms for compressing trajectories are evaluated in [16]. The goal of these algorithms is again to reduce the amount of points preserved in the simplification and maintain the shape of the trace.

Direction preserving methods [14] aim at simplifying location data but preserving the overall direction of the input trace, by maintaining locations that would impose a great change in the original direction of the trajectory, if absent. Towards a similar goal, authors in [24, 13] suggest the use of dead reckoning in order to calculate the current position of a target based on her previous positions. A sample of the current geographical location is stored if it deviates from the estimation significantly. This line of work attempts to achieve adaptive sampling of location data, however, it requires the recording of a new sample at all times in order to compare it with the estimated position. What is more, the accumulation of the approximation error results in samples that might deviate significantly from the original trajectory. Historical data of cell sequences has been used in [19] to estimate the current location of a moving target. Similar to dead reckoning, these methods do not guarantee that all important information of a user's mobility will be maintained in the sample and only attempt to approximate the shape of the original trajectory.

An interesting characteristic of this line of work, however, is the online-processing of location data, also known as streaming spatial data. There are cases when it is useful to have methods that

allow for the real-time compression of GPS traces; for example, when the method operates on a lightweight-device (like a smartphone, or tablet) where we do not want to maintain a complete copy of the trajectory. An efficient online, multiresolution algorithm is suggested in [9], giving small sized simplifications of small approximation errors. An experimental comparison of online simplification algorithms can be found in [8]. In those cases, however, the main goal is again to simplify a mobility trace so that it approximates the original one. What is more, in all these works, the GPS samples are taken at a pre-defined sampling interval and each algorithm, at each time instant, makes a decision whether to store the sample or not.

The problem we are tackling in this work is substantially different. Firstly, we are not interested in simplifying a pre-recorded GPS trajectory but we aim at finding the sampling frequency at which to sample geographical locations. Our ultimate goal is to reconstruct a user's movement completely, including both the spatial and temporal characteristics of her mobility (see Figure 3).

In [12], the authors ask when should various sensors (including GPS) be sampled in order to get a new location position. The goal here is to decrease the energy consumption and provide accurate approximations depending on the application at hand.

In [2], the authors are suggesting heuristics to simplify GPS trajectories with the goal to maintain both the shape of the trajectory but also its semantics. They define the importance of each point based on its direction and its distance from its neighbours. In our work, we do not focus on mining or sampling the important locations, but the overall movement of a user in sparse samples.

Works on defining and mining important locations from mobility traces can be found in [18, 11, 27] and references therein. In [18], the authors aim at understanding human mobility and analyze a set of mobility characteristics. In their work, points of interest are defined with respect to frequency of visit, per user. When considering a multi-user scenario, [11] define as important the locations visited by many users. In [27], the important locations are popular and culturally significant places (e.g, museums, shopping centers).

To the best of our knowledge, this is the first work to thoroughly study the problem of finding a good sampling frequency at which to sample human mobility so that a user's movement can be accurately reconstructed.

6 Discussion and conclusions

In the light of the results presented in this work, a laconic answer to the original question posed in the Introduction is *"it depends on the error one can afford"*. Fortunately, our findings are more informative than that, and provide a simple scaling law for the user positioning error, with general validity and limited parametrization diversity across individuals. The scaling law we identified has practical applications in a number of contexts. Below, we list and discuss a few usages that look especially promising.

- *Energy-efficient mobile computing.* It has been shown that frequent sampling of GPS data tends to quickly drain the batteries of a mobile device [20, 25]. A natural solution to this is to sample the device position with reduced frequency. However, deciding which periodicity should be employed is not trivial. The linear scaling law we identified can be easily extended to control the tradeoff between energy consumption and localization accuracy. It can also become an important building block in techniques for the dynamic adaptation of GPS data collection frequency in mobile devices –based not only on, e.g., context, but also on the desired accuracy.

- *Location-based service operation.* Location-based services (LBS) rely on positioning information about their users. Yet, an excessively frequent collection of user locations is expensive from both energy and communication perspectives, it raises privacy concerns, and it can ultimately bother customers. Our results may help towards taking informed decisions, reducing the periodicity of localization depending on the level of approximation in the user's position that can be tolerated by the service.

- *Active probing of subscribers' positions in mobile networks.* Precise knowledge of subscribers' locations is a valuable information for mobile operators, for both network management and value-added service development [17, 6]. Actively probing mobile devices for their position in a country-scale mobile network is a computationally expensive task, which has traditionally pushed operators to favor less controllable passive measurements [5]. In fact, mobile subscribers are localized based on the antenna sector they are associated with, and those sectors cover hundreds of m^2 in the best case. In this context, our results suggest that running active probing at, e.g., every hour, would not decrease significantly the measurement accuracy –as the incurred error is comparable to the antenna sector coverage.

- *Trajectory data compression.* A straightforward application of our results is data compression. If large amounts of trajectories must be stored, and memory becomes an issue, one could sample the original movement data at some fixed frequency, and store only the samples. This is a lossy operation, yet the scaling law we identified provides reasonable indications about the incurred loss of information.

To summarize, in our work, we analyzed a collection of fine-grained GPS trajectories in a quest to understand whether human mobility can be sampled at reduced frequencies so that a user's movement can be reconstructed with minimum loss of information. We found that there is a linear scaling law of the localization error with respect to the sampling frequency across all users; in fact we found that in the vast majority of our user base, an added minute of sampling interval increases the localization error from 1 to 4 meters.

Our current research directions involve the exploration of the impact that more complex interpolation techniques have on the shape and parametrization of the scaling law. We are also investigating the physical reasons behind the diversity of the linear scaling law parameter among different individuals. Finally, we are working on an online smart algorithm that, building on our results, tailors the sampling frequency to the user mobility over time.

References

[1] M. Chen, M. Xu, and P. Franti. A fast $o(n)$ multiresolution polygonal approximation algorithm for gps trajectory simplification. *Trans. Img. Proc.*, 21(5):2770–2785, May 2012.

[2] Y. Chen, K. Jiang, Y. Zheng, C. Li, and N. Yu. Trajectory simplification method for location-based social networking services. In *Proceedings of the 2009 International Workshop on Location Based Social Networks*, LBSN '09, pages 33–40, New York, NY, USA, 2009. ACM.

[3] B. C. Csáji, A. Browet, V. Traag, J.-C. Delvenne, E. Huens, P. V. Dooren, Z. Smoreda, and V. D. Blondel. Exploring the mobility of mobile phone users. *Physica A: Statistical Mechanics and its Applications*, 392(6):1459 – 1473, 2013.

[4] D. H. Douglas and T. K. Peucker. *Algorithms for the Reduction of the Number of Points Required to Represent a Digitized Line or its Caricature*, pages 15–28. John Wiley & Sons, Ltd, 2011.

[5] M. Ficek, T. Pop, P. Vláčil, K. Dufková, L. Kencl, and M. Tomek. Performance study of active tracking in a cellular network using a modular signaling platform. In *Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services*, MobiSys '10, pages 239–254, New York, NY, USA, 2010. ACM.

[6] A. Finamore, M. Mellia, Z. Gilani, K. Papagiannaki, V. Erramilli, and Y. Grunenberger. Is there a case for mobile phone content pre-staging? In *Proceedings of the Ninth ACM Conference on Emerging Networking Experiments and Technologies*, CoNEXT '13, pages 321–326, New York, NY, USA, 2013. ACM.

[7] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, June 2008.

[8] N. Hönle, M. Grossmann, S. Reimann, and B. Mitschang. Usability analysis of compression algorithms for position data streams. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '10, pages 240–249, New York, NY, USA, 2010. ACM.

[9] P. Katsikouli, R. Sarkar, and J. Gao. Persistence based online signal and trajectory simplification for mobile devices. In *Proceedings of the 22Nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '14, pages 371–380, New York, NY, USA, 2014. ACM.

[10] R. Keralapura, A. Nucci, Z.-L. Zhang, and L. Gao. Profiling users in a 3g network using hourglass co-clustering. In *Proceedings of the Sixteenth Annual International Conference on Mobile Computing and Networking*, MobiCom '10, pages 341–352, New York, NY, USA, 2010. ACM.

[11] S. Khetarpaul, R. Chauhan, S. K. Gupta, L. V. Subramaniam, and U. Nambiar. Mining gps data to determine interesting locations. In *Proceedings of the 8th International Workshop on Information Integration on the Web: In Conjunction with WWW 2011*, IIWeb '11, pages 8:1–8:6, New York, NY, USA, 2011. ACM.

[12] M. B. Kjærgaard, S. Bhattacharya, H. Blunck, and P. Nurmi. Energy-efficient trajectory tracking for mobile devices. In *Proceedings of the 9th International Conference on Mobile Systems, Applications, and Services*, MobiSys '11, pages 307–320, New York, NY, USA, 2011. ACM.

[13] R. Lange, F. Dürr, and K. Rothermel. Online trajectory data reduction using connection-preserving dead reckoning. In *Proceedings of the 5th Annual International Conference on Mobile and Ubiquitous Systems: Computing, Networking, and Services*, Mobiquitous '08, pages 52:1–52:10, ICST, Brussels, Belgium, Belgium, 2008. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

[14] C. Long, R. C.-W. Wong, and H. V. Jagadish. Direction-preserving trajectory simplification. *Proc. VLDB Endow.*, 6(10):949–960, Aug. 2013.

[15] P.-F. Marteau and G. Ménier. Speeding up simplification of polygonal curves using nested approximations. *Pattern Analysis and Applications*, 12(4):367–375, 2009.

[16] J. Muckell, J.-H. Hwang, C. T. Lawson, and S. S. Ravi. Algorithms for compressing gps trajectory data: An empirical evaluation. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '10, pages 402–405, New York, NY, USA, 2010. ACM.

[17] E. M. R. Oliveira and A. C. Viana. From routine to network deployment for data offloading in metropolitan areas. In *SECON*, 2014.

[18] E. M. R. Oliveira, A. C. Viana, C. Sarraute, J. Brea, and I. Alvarez-Hamelin. On the regularity of human mobility. *Pervasive and Mobile Computing*, pages –, 2016.

[19] J. Paek, K.-H. Kim, J. P. Singh, and R. Govindan. Energy-efficient positioning for smartphones using cell-id sequence matching. In *Proceedings of the 9th International Conference on Mobile Systems, Applications, and Services*, MobiSys '11, pages 293–306, New York, NY, USA, 2011. ACM.

[20] Y. Qi, C. Yu, Y. J. Suh, and S. Y. Jang. Gps tethering for energy conservation1. In *2015 IEEE Wireless Communications and Networking Conference (WCNC)*, pages 1320–1325, March 2015.

[21] M. Z. Shafiq, L. Ji, A. X. Liu, and J. Wang. Characterizing and modeling internet traffic dynamics of cellular devices. In *Proceedings of the ACM SIGMETRICS Joint International Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS '11, pages 305–316, New York, NY, USA, 2011. ACM.

[22] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[23] B. Thierry, B. Chaix, and Y. Kestens. Detecting activity locations from raw gps data: a novel kernel-based algorithm. *International Journal of Health Geographics*, 12(1):14, 2013.

[24] G. Trajcevski, H. Cao, P. Scheuermanny, O. Wolfsonz, and D. Vaccaro. On-line data reduction and the quality of history in moving objects databases. In *Proceedings of the 5th ACM International Workshop on Data Engineering for Wireless and Mobile Access*, MobiDE '06, pages 19–26, New York, NY, USA, 2006. ACM.

[25] S. Vanini, F. Faraci, A. Ferrari, and S. Giordano. Using barometric pressure data to recognize vertical displacement activities on smartphones. *Computer Communications*, 87:37 – 48, 2016.

[26] Y. Zheng, Q. Li, Y. Chen, X. Xie, and W.-Y. Ma. Understanding mobility based on gps data. In *Proceedings of the 10th International Conference on Ubiquitous Computing*, UbiComp '08, pages 312–321, New York, NY, USA, 2008. ACM.

[27] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma. Mining interesting locations and travel sequences from gps trajectories. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 791–800, New York, NY, USA, 2009. ACM.

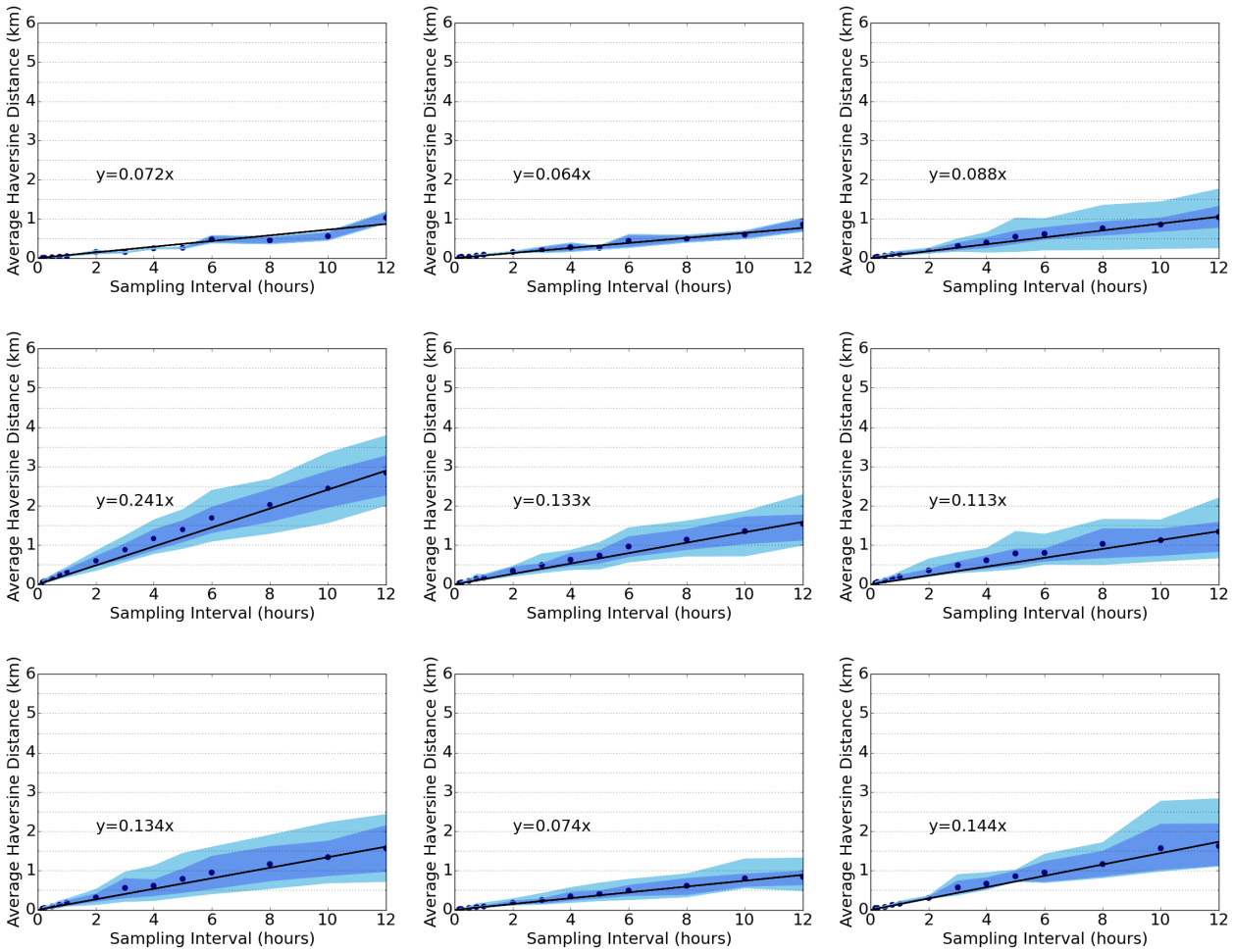


Figure 9: Average Haversine distance between the original and the reconstructed trajectories of nine users, versus sampling intervals between 10 minutes and 12 hours. Dots represent mean values. Dark and light shaded regions depict the 25-75% and 10-90% quantiles. Solid lines are the linear fittings on average points. Figure best seen in color.

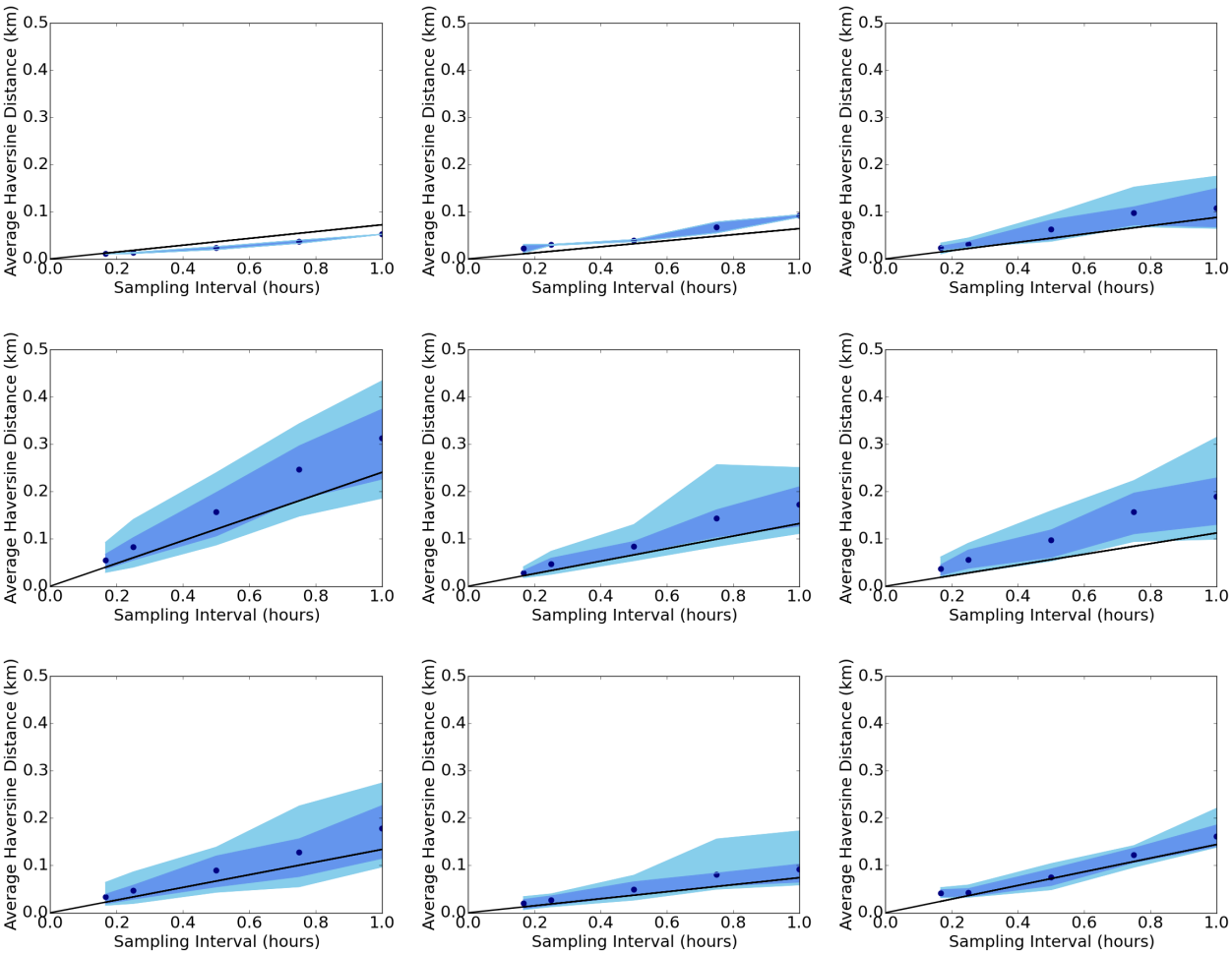
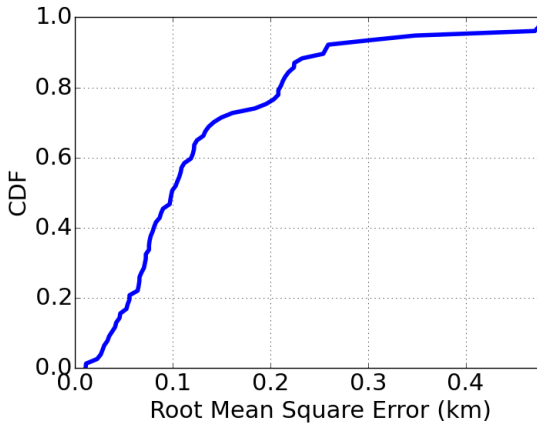


Figure 10: Average Haversine distance between the original and the reconstructed trajectories of nine users, versus sampling intervals between 10 minutes and 1 hour. Dots represent mean values. Dark and light shaded regions depict the 25-75% and 10-90% quantiles. Solid lines are the linear fittings on average points. Figure best seen in color.



11: Distribution of the RMSE incurred by approximating the trajectory between the average

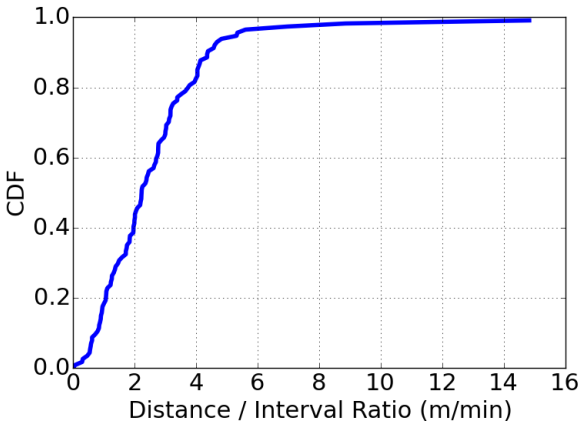
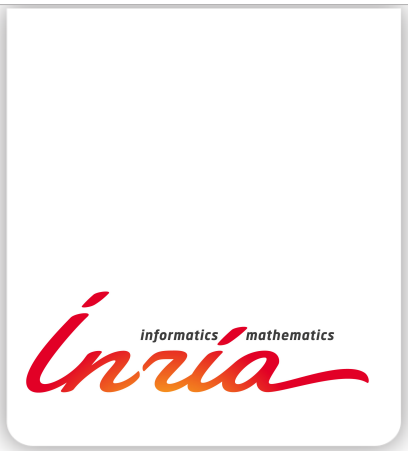


Figure 12: Distribution of the ratio between the average Haversine distance and the sampling interval, for all users.

ne distance and the sampling interval as a line, for all users.



RESEARCH CENTRE
SACLAY – ÎLE-DE-FRANCE

1 rue Honoré d'Estienne d'Orves
Bâtiment Alan Turing
Campus de l'École Polytechnique
91120 Palaiseau

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-0803