Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

DuctiLoc: Energy-efficient Location Sampling with Configurable Accuracy

PANAGIOTA KATSIKOULI[1], DIEGO MADARIAGA[2], ALINE CARNEIRO VIANA[3], ALBERTO TARABLE[4] and MARCO FIORE[5]

[1]DIKU (Department of Computer Science), UCPH, Denmark
[2]NIC Chile Research Labs, Chile
[3]Inria, Palaiseau, France
[4]CNR-IEIIT, Italy
[5]IMDEA Networks Institute, Spain

Corresponding author: Panagiota Katsikouli (e-mail: pk@di.ku.dk).

ABSTRACT Mobile device tracking technologies based on various positioning systems have made location data collection ubiquitous. The frequency at which location samples are recorded varies across applications, yet it is usually pre-defined and fixed, resulting in redundant information, and draining the battery of mobile devices. In this paper, we first answer the question *"at what frequency should individual human movements be sampled so that they can be reconstructed with minimum loss of information?"*. Our analysis unveils a novel linear scaling law of the localization error with respect to the sampling interval. We then present DuctiLoc, a location sampling mechanism that utilises the law above to profile users and adapt the position tracking frequency to their mobility. DuctiLoc is energy efficient, as it does not rely on power-hungry sensors or expensive computations; moreover, it provides a handy knob to control energy usage, by configuring the target positioning accuracy. Controlling the trade-off between accuracy and sampling rate of human movement is useful in a number of contexts, including mobile computing and cellular networks. Real-world experiments with an Android implementation show that DuctiLoc can effectively adjust the sampling frequency to individual mobility habits and target accuracy level, reducing the energy consumption by 60% to 98% with respect to a baseline periodic sampling.

INDEX TERMS Human mobility, Trajectory data, Location sampling, Configurable accuracy, Energy efficiency

I. INTRODUCTION

OVER the past decade, the pervasive usage of smart devices and location-tracking systems has made it possible to study and understand human mobility at unprecedented scales. A number of studies based on measurements from millions of data subjects have repeatedly demonstrated that human mobility is highly regular [5] and predictable [30], as people tend to follow the same patterns over and over, and they do so in ways that are clearly periodic. Regularity is easily found in the both spatial and temporal dimensions of the movements of most individuals. As an example, let us consider Fig. 1, which shows heatmaps of the locations visited by three random users in our reference dataset (presented later in Section III-A). Although these plots summarize three weeks of data, a small set of frequently visited places emerges for all users, along with systematic paths connecting them. Likewise, Fig. 2 illustrates the temporal regularity of the mobility of the same users, as a clear periodicity emerges from the time series of their visited locations.

In this paper, we study whether the spatiotemporal regularity of human mobility entails the possibility of sampling individual movements *at reduced frequencies in an adaptive way while allowing for the reconstruction of trajectories* that retain a vast portion of their original level of detail. Intuitively, periodic visits to a limited set of important places through repeated routes may be captured with a smaller sampling effort, as opposed to completely random mobility that would need incessant and continuous sampling to be tracked effectively. The problem of identifying suitably reduced frequencies for human mobility sampling is in fact 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?"*.

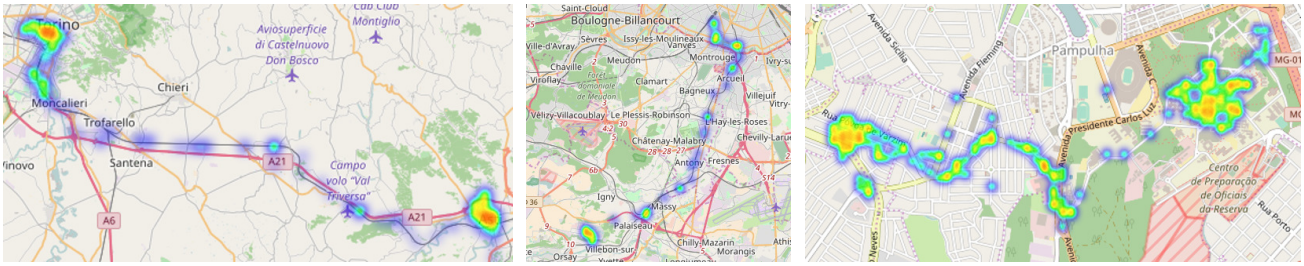


FIGURE 1: Heatmaps of locations visited by three distinct individuals located in different countries during three consecutive weeks: people tend to move among a limited set of specific locations, following repeated patterns. Figures best seen in color.

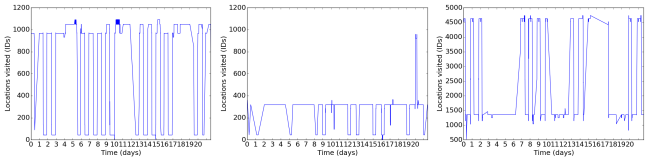


FIGURE 2: Location time series for three distinct users during three weeks: humans tend to revisit locations in a periodic fashion. The visited locations are mapped to sequential identifiers, upon discretization on a regular grid of 50 meters step.

In this context, we investigate the effect that increasingly reduced fixed sampling frequencies have on the quality of tracked movement, via a signal processing approach: we consider mobility patterns as signals over time, and carry out a spectral analysis of human mobility. We found that the spectra of the movements of 119 individuals have very similar, flat shapes; this suggests 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 carried out a quantitative analysis of the user localization error in movements reconstructed from regular sampling at different periodicities. Our results unveil a *linear scaling law of the error* with respect to the span of the constant sampling interval. This law corroborates the outcome of the spectral analysis, and has significant practical implications, as it allows controlling the trade-off between accuracy and cost of measurements of human mobility. Examples of practical applications in a number of fields include, but are not limited to: (i) mobile computing, where overly frequent GPS localization unnecessarily reduces the battery life of mobile devices; (ii) location-based service design, where unwarranted users' position data collection raises significant privacy concerns; (iii) cellular networks, where active probing of subscribers' positions is a costly task whose rate must be duly optimized; (iv) trajectory data compression, where information loss must be minimized. It is therefore crucial to precisely control the trade-off between sampling rate and accuracy of human mobility.

Inspired by these results, we focus on the first usage above (i.e., mobile computing and energy saving of tracking mobile devices) and develop DuctiLoc, a *ductile localization* technique that takes advantage of the newly unveiled linear scaling law to adjust the sampling frequency according to the user's mobility habits. We implement DuctiLoc as a mobile phone app and run experiments with real-world mobile users in different countries. Our experimental results highlight that:

- DuctiLoc reduces energy consumption by 60%-98% with respect to a high-frequency periodic sampling, without compromising the tracking quality;
- DuctiLoc successfully operates without relying on mobile device sensors such as accelerometer or gyroscope, which significantly reduces the usage of computational resources in the device;
- DuctiLoc enables a unique explicit configuration of the desired location accuracy, and –by correspondingly adapting the sampling frequency– an indirect control over the battery drain of the tracking process, which is not possible with previous approaches.

DuctiLoc is an effective, lightweight adaptive location sampling mechanism based on an original concept, which can operate in isolation, or complement more traditional techniques based on auxiliary sensors or/and alternative positioning systems. As such, DuctiLoc can support downstream applications that do not necessarily require high accuracy hence continuous sampling of the user position (such as navigation), and rather focus on different aspects like long stay periods at specific locations. Indeed, it is in these situations that smart sampling becomes an appropriate solution, as we better expound in the conclusions of the manuscript.

The paper is organized as follows. We first discuss related studies in Section II. In Section III, we present the setup and insights of our quantitative analysis of human mobility as a signal. Building on those results, we present the design of DuctiLoc in Section IV, while its implementation and experimental performance evaluation are in Section V. Finally, conclusive remarks are drawn in Section VI.

II. RELATED WORK

Previous studies have considered techniques for a simplified representation of human mobility, as well as practical solutions for the adaptive utilization of localization functions in mobile devices. Our work is related to both topics, and we discuss the relevant literature next. We also remark that this paper extends an earlier version of the study, which was

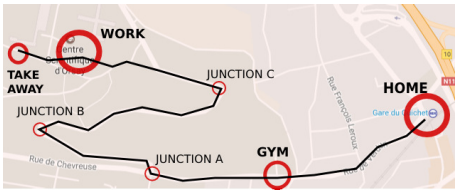


FIGURE 3: Toy example of a mobility trajectory. Labels denote important locations or turning points. The size of the circle around each location is proportional to the amount of time spent there by the user.

limited to the analysis of the effect of the sampling frequency on the quality of tracked movements [3]. The contributions of this previous study are presented and discussed in Section III, and serve as a basis for introducing our novel DuctiLoc mechanism.

A. STREAMLINING HUMAN MOBILITY

The problem of identifying the minimum sampling rate of individual movements is related to –but should not be confounded with– multiple well-researched topics in human mobility analysis.

A large literature has addressed the problem of spatial data trajectory compression. There, the objective is to maintain the *shape* of a spatial trajectory while simplifying its representation; representative examples include, *e.g.* [19], [20] and the references therein. Consider the toy example in Fig. 3, where a user leaves home, trains at the gym before work, and later goes to a take-away restaurant. A trajectory compression technique could approximate the shape of the spatial mobility as the sequence of home, junctions B and C, and take-away locations: then, map-matching based on these cardinal points would provide a fair description of the movement. However, trajectory compression neglects the temporal dimension of the mobility, portrayed in Fig. 3 as circles proportional to the time spent at each location. Our purpose is instead to recreate the complete mobility of the user, including these temporal features.

The problem we address is also different from sampling to detect important locations [21], [22], or from simplifying GPS trajectories to preserve location semantics [23], [24]. In the example of Fig. 3, important location detection is solved by sampling the trajectory so as to model the original distribution of time spent at home, work, gym, and take-away. However, that factually ignores the time ordering of visits, and does not capture transitions between frequent locations. We aim at identifying a sampling process that captures all these characteristics.

Finally, we are not interested in identifying and retaining locations that determine major changes in the heading of a trajectory [25]; nor we address the similar problem of calculating the current position of a target based on its traveled distance and direction of movement, known as dead reckoning [26], [27]. Indeed, we are not interested in simplifying a pre-recorded GPS trajectory, but in finding convenient sampling frequencies for human trajectory data.

B. ADAPTIVE LOCALIZATION

Our proposed solution dynamically changes the sampling frequency of GPS localization, which is an approach that has been widely investigated in the past. However, the vast majority of previous works utilize auxiliary sensors, or alternative positioning systems to trigger GPS recording. The accelerometer embedded in mobile devices has been the primary auxiliary system used in adaptive location-sensing mechanisms; for instance, accelerometer information allows minimizing the probability of exceeding a given positional error while maximizing the energy-saving when tracking users indoors [14].

Other sensors can complement or replace the accelerometer; as an example, the Bluetooth interface and the microphone can be used to implement an adaptive sampling of movements and locations associated to interesting events [16]. Such sensors can be further combined with a dedicated geographical zoning to limit GPS activation to important spatial movements [13]. Velocity information, *e.g.* from historical data, has also been used to trigger GPS localization only when required [17].

In other cases, alternative and less energy-hungry localization mechanisms are used as a partial replacement for GPS, or positioning data demanded by other apps running on the mobile device is reused to reduce GPS activation [18].

Our proposed solution, DuctiLoc, brings three major elements of novelty with respect to models in the literature. First, it develops a practical location sampling technique on top of novel insights on the representation of human mobility that were not capitalized upon before. Second, unlike previous approaches that rely on external sensors or systems, DuctiLoc does not need auxiliary elements beyond the main localization system (*e.g.* GPS); as proven by our experimental performance evaluation, this spares computational resources in the mobile device, and entails further savings on its battery consumption. Third, DuctiLoc is ductile, *i.e.* it allows configuring the desired level of localization accuracy, which can be also leveraged as a knob to adjust the energy usage of the localization process, and is an unprecedented feature in this type of tools.

III. SAMPLING INDIVIDUAL MOBILITY FOR TRAJECTORY RECONSTRUCTION

We aim at answering the question "*at what frequency should one periodically sample individual human movements so that they can be reconstructed from the collected samples with minimum loss of information?*". We approach the question in a systematic way, using a reference dataset collected by volunteers in several countries through diverse initiatives as presented in Section III-A. We first perform a spectral analysis of such data, in Section III-B: by considering human movements as a signal in time and studying its spectrum in frequency, we observe the absence of low sampling frequencies that can capture a large portion of the human mobility.

TABLE 1: Per-source users and weeks in the reference dataset.

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

Next, in Section III-C, we refine the result via an extensive quantitative investigation of the exact trade-off between the sampling frequency and the quality of the recorded movement.

A. REFERENCE DATASET

In this study, we employ a dataset of real-world individual mobility data collected from three different initiatives.

- The **MACACO** data was collected between July 2014 and December 2016 as part of the European collaborative project MACACO. The project collected GPS positioning information of volunteers located in Europe and South America with a regular sampling interval of one to five minutes, via a dedicated smartphone application[1].
- The **OpenStreetMap** (OSM) data was collected by volunteers who recorded and uploaded their trajectories as part of contributions to the OSM database[2]. The OSM project is a global crowdsourcing initiative aiming to map the whole world surface thanks to a vast community of supporters. The GPS traces uploaded by OSM participants typically feature 1-Hz frequency, and are freely available at the official OSM project website.
- The **Geolife** data was collected in Beijing, China, by researchers of Microsoft Research Asia, between April 2007 and August 2012 [4]. It consists of GPS trajectories recorded through different GPS loggers and smartphone applications. 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 at the official project website[3].

In order to build a consistent reference dataset, we first homogenize the GPS trajectory data from the three sources above. Specifically, we segment the mobility traces of all users into one-week trajectories and analyze them separately, under the rationale that human activities have been shown to have a weekly periodicity [28], [29] hence weekly logs let us capture most of the regularity of human mobility. These weekly trajectories have heterogeneous quality in terms of completeness of the mobility information, and many feature relatively long periods with missing or erroneous data: we thus filter the trajectories, retaining only those that contain complete GPS records in at least six out of seven distinct weekdays.

As a result, our reference dataset is composed of 1,052 weeks of the mobility of 119 different individuals, which cover sensibly different geographical spans and can encompass a single city or multiple continents. Tab. 1 provides a break down of the number of users and trajectories on a per-source basis. Fig. 4(a)-(c) give further detail, portraying the distribution of the number of weekly trajectories associated to a particular user, separately for each data collection initiative. The plots show that the vast majority of users contribute one to ten weeks of movement data, hence ensuring that the data is representative of a fairly diverse base of individuals. Also, our dataset contains 4 weeks of data on average for each individual and up to 48 weeks for a single user: these observation periods are long enough that the data of a single individual often captures irregular patterns due to non-periodic endeavours of the user.

It is to be noted that the sampling periodicity in the retained weeks is not uniform. Despite the filtering on completeness, 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 may have minor temporal gaps due to offline GPS receivers, or interruptions in the data collection service. Fig. 4d shows the Cumulative Distribution Function (CDF) of the sampling intervals observed in all one-week trajectories of our reference dataset. In 95% of cases, consecutive positioning samples are collected within 10 minutes from each other; in the case of the OSM and GeoLife sources, 90% to 95% of points are less than 10 seconds apart, as highlighted in Fig. 4e. In all cases, the sampling intervals above appear sufficient to capture well human movements on a weekly basis, and allow for a reliable approximation of the time-continuous mobility via a simple linear space-time interpolation on the available location samples. We then resample all interpolated trajectories with the same frequency, *i.e.*, 5 minutes, and use the resulting set of homogeneously granular trajectories as the ground truth for the remainder of the study.

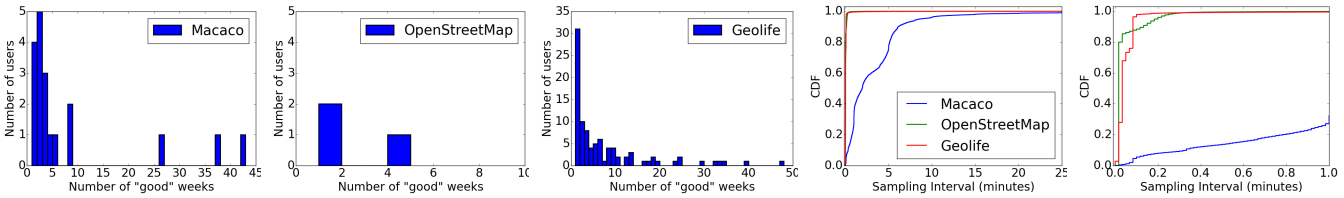
B. SPECTRAL ANALYSIS OF INDIVIDUAL MOBILITY

As we are interested in determining a proper sampling periodicity for human movements, a signal processing approach appears especially well suited. To this end, we explore mobility through the lenses of Fourier transformations: we translate the trajectory data in the frequency domain, and carry out a thorough spectral analysis of their frequency components.

1) Time series representation of mobility

As a preliminary step to the spectral analysis, we need to transform individual GPS trajectories into unidimensional time series; this poses a challenge since, even when ignoring altitude information, points in geographical trajectories are obviously bidimensional. We carry out an extensive evaluation of approaches to reduce bidimensional movements to

[1] https://macaco.inria.fr/macacoapp/.
[2] https://www.openstreetmap.org.
[3] https://www.microsoft.com/en-us/research/project/geolife-building-social-networks-using-human-location-history/.



(a) MACACO (b) OpenStreetMap (c) GeoLife (d) Full view (e) Zoomed view

FIGURE 4: (a)-(c) User distribution of one-week trajectories in our reference dataset. (d)-(e) Distribution of the sampling interval in one-week trajectories in our reference dataset: full interval span and sub-minute intervals. Figure best seen in color.

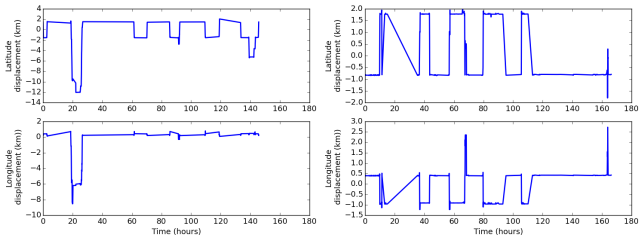


FIGURE 5: Unidimensional movement signals of two representative one-week trajectories (left and right). Plots refer to latitude (top) and longitude (bottom) displacements.

Latitude	Longitude
MACACO	
0.55	0.84
OpenStreetMap	
0.53	0.7
GeoLife	
0.57	0.62

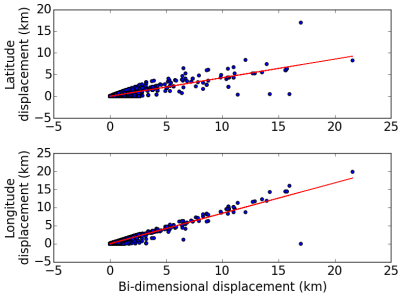


FIGURE 6: Correlation between unidimensional displacements and original traveled distance in the bidimensional space. (left) Coefficients separated by data source. (right) Scatterplot of latitude (top) and longitude (bottom) displacemnets with respect to the actual bidimensional movement, for trajectories in the MACACO data.

unidimensional signals, using approximated measures such as the movement velocity or the relative displacement from the center of mass, as well as applying transformations such as the enumeration of discretized locations in the Hilbert space. However, all these techniques introduce an exceeding amount of noise in the time series, which introduces artificial or unrealistic patterns in the user movements.

The difficulty of identifying a reliable univariate representation of mobility lets us opt 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 center of mass of the one-week trajectory. Formally, the displacements of the n-th positioning sample in a trajectory are denoted as $\bar{\phi}[n]$ and $\bar{\lambda}[n]$ for latitude and longitude respectively, computed as

$$\bar{\phi}[n] = \phi[n] - \frac{1}{N}\sum_{n'=0}^{N-1}\phi[n'], \tag{1}$$

$$\bar{\lambda}[n] = \lambda[n] - \frac{1}{N}\sum_{n'=0}^{N-1}\lambda[n'], \tag{2}$$

where $\phi[n], \lambda[n]$ are the latitude and longitude coordinates of the n-th GPS point, and N is the number of samples in the weekly trajectory. Other than making time series more easily compared across users and weeks, the transformations in (1)–(2) have the desirable property of generating zero-mean signals whose spectra have no DC component. Illustrations of our unidimensional description of individual movements are in Fig. 5, for two one-week trajectories.

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 analyze the correlation between the isolated latitude or longitude displacements and the actual traveled distance in the bidimensional space. Fig. 6 shows the per-source correlation coefficients, as well as the linear fitting on trajectories from the MACACO data. We observe consistently good correlations in all cases, and conclude that both dimensions, when taken separately, still provide fair approximations of the overall mobility. Interestingly, the correlation is always stronger for longitude than for latitude, indicating that participants to all data sources tend to move along an East-West axis rather than along a South-North one; we hypothesize that this may be due to the topology of the cities where data collections took place.

2) Frequency spectra of human mobility

We apply a Fast Fourier Transform (FFT) to the latitude and longitude displacement signals inferred from each one-week trajectory, so as to compute their spectral representation. 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 (*i.e.* the bandwidth of the signal), the Nyquist–Shannon sampling theorem

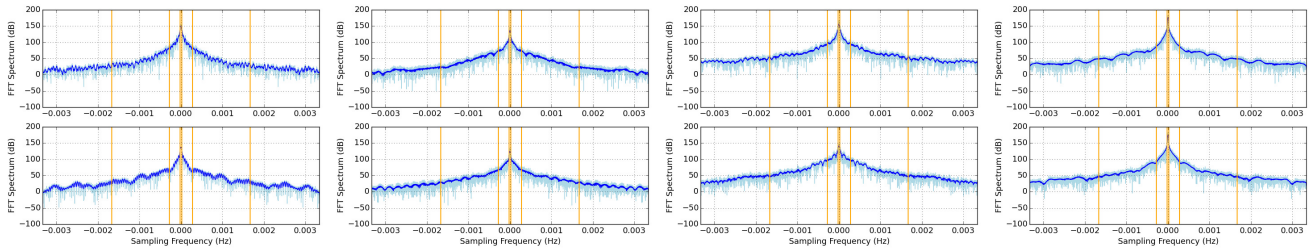


FIGURE 7: Frequency spectra of unidimensional movement signals, for a selection of one-week trajectories. Figure best seen in color.

guarantees that a sampling rate $2f_s$ is enough to achieve a lossless reconstruction of the original signal from its samples. For practical signals, the spectrum is 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 shows the spectra of the latitude (top) and longitude (bottom) displacement signals of a representative selection of one-week trajectories. The plots in first two columns refer to the signals in Fig. 5. The original spectra are in light blue, while a moving average curve that better displays the overall trends is in dark blue. The time granularity of the trajectory data, *i.e.* 5 minutes, sets the spectrum boundaries at $\pm 0.00\bar{3}$ Hz frequencies, while the 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). While we only present visualization for a subset of the data in Fig. 7, we found the overwhelming majority of spectra to be very much alike those in the reported plots.

Based on the spectra, we make two important remarks: (i) the hundreds of very diverse trajectories in our dataset all yield spectra with very similar shapes; (ii) the spectral shapes do not show evidence of a bandwidth threshold beyond which the signal power becomes clearly negligible, *i.e.* they do not identify a clear operational point for effective sampling. We can explain both these phenomena by considering that the unidimensional movement signals typically feature constant values linked by very steep transitions and deep spikes, as exemplified in Fig. 5: capturing such a behavior requires a near-infinite bandwidth, which results in spectra with a slow decay for high frequencies. In other words, although it exhibits a clear periodicity [30], human mobility is, in fact, a succession of long periods where individuals are almost static, with fast transitions between important locations [33]. While positions during stationary time intervals contribute to low-frequency spectral components and are hence easily captured by a sparse sampling, traveling causes discontinuities in the mobility signal and is much harder to sample. As a result, the spectra do not reveal interesting operation points in the trade-off between the sampling frequency (and its associated energy cost) and the quality of the reconstructed signal.

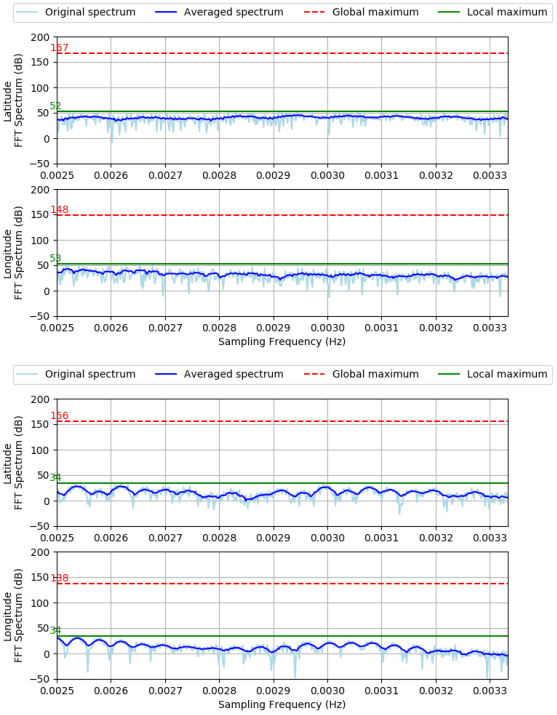


FIGURE 8: Zoomed-in spectra for two representative one-week trajectories, focusing on high frequency values.

It is also worth noting that a rightful doubt is whether the full spectra in Figure 7 provide a too high-level view of the transform, and hide important details. In particular, sporadic and uncommon movement patterns of the users may generate important high-frequency components that may be difficult to spot in the complete spectrum plots. We investigate this aspect by looking at zoomed-in versions of the full spectra, which focus on the highest frequencies only, as in the examples in Figure 8. While we only provide results for two representative trajectories for the sake of brevity, all spectra yield the same behavior: the amplitude of high-frequency components (highlighted by the solid green line in the plots) is in all cases orders of magnitudes (considering that the ordinate is in dB) lower than that of low-frequency components (outside the abscissa of the plots, but reported as the dashed red line). In other words, regular mobility patterns largely dominate over spurious ones, which confirms

previous findings in the literature [5], [30].

C. QUANTITATIVE ANALYSIS OF TRAJECTORY SAMPLING

Although disappointing in a sense, the outcome of our spectral analysis calls forth for further investigation to confirm, better understand and possibly model the apparent steady relationship between the quality and cost of individual mobility sampling. To this end, we perform an extensive quantitative analysis, and investigate the impact of different constant sampling frequencies on the quality of the mobility reconstructed from the collected samples.

1) Measuring error in reconstructed trajectories

We first create downsampled versions of the one-week trajectories in our reference dataset, using a wide range of sampling intervals, from 10 minutes to 12 hours. We then reconstruct complete trajectories from their retained sample by linearly interpolating such points. Finally, we assess how such time-continuous downsampled trajectories compare to the original ones: specifically, we measure the error in retrieving the individual trajectory from sampled data via the average Haversine distance. Given two points on Earth surface, $p_a = (\phi_a, \lambda_a)$ and $p_b = (\phi_b, \lambda_b)$, their Haversine distance is computed as

$$H(p_a, p_b) = R \cdot 2 \cdot \texttt{atan2}\left(\sqrt{\theta}, \sqrt{(1-\theta)}\right). \quad (3)$$

Here, `atan2` is a well-known function that returns a unambiguous phase value in Cartesian to polar coordinates conversion, and $\theta = \sin^2(\Delta_\phi/2) + \cos(\lambda_a) \cdot \cos(\phi_b) \cdot \sin^2(\Delta_\lambda/2)$, where $\Delta_\phi = \phi_b - \phi_a$, and $\Delta_\lambda = \lambda_b - \lambda_a$. Also, $R = 6,371$ km is the Earth radius. The average Haversine distance of two (original and downsampled) trajectories is the mean of all Haversine distances between each original sample and its counterpart, *i.e.* the sample associated to the same timestamp, in the downsampled trajectory.

2) Linear relationship of sampling frequency and error

In the following, we report results aggregated on a per-user basis, since we find all trajectories of a same individual to yield very similar properties. Also, while we show in detail the results of a few representative users, all individuals in our reference datasets exhibit similar behaviors.

Fig. 9 shows the evolution of the average Haversine error against the sampling interval for a choice of eight individuals. Each plot presents results for all of the one-week trajectories of one person: as multiple weekly 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, *i.e.* average Haversine distance, measured over all trajectories of the user.

Remarkably, a neat linear relationship characterizes all plots. Indeed, a simple linear model

$$\bar{H} = \alpha \cdot T = \alpha / f, \quad (4)$$

fits very well the mean error, as shown by the solid lines in Fig. 9. In (4), $\bar{H}$ is the average Haversine distance between the original and downsampled trajectories, T is the sampling interval, and f is its reciprocal, *i.e.* the sampling frequency. The result holds for all individuals in our dataset, as proven in Fig. 10a. The plot portrays the CDF, computed over all users, of the Root Mean Square Error (RMSE) between the linear fitting (*i.e.* the solid line in Fig. 9) and the mean Haversine distance at different sampling intervals (*i.e.* dots in Fig. 9); in other words, the RMSE quantifies the accuracy of the linear approximation in representing how the distance varies with the sampling frequency. In 95% of the cases, the RMSE is below 250 meters, and it drops to 100 meters in 50% of the cases; these are reasonable values considering that our data subjects travel tens of km per day.

The average RMSE value in Fig. 10a is computed over all sampling intervals, and the distribution of the error is not uniform across them: Fig. 9 shows how the error is in fact very small for short sampling intervals, and only increases when the location is sampled at every 2 hours or more. In other words, the linear scaling is very accurate for sampling frequencies below one hour; errors in the order of km, which may seem prominent in Figure 8, are incurred for infrequent sampling occurring at intervals of multiple hours. This proves that irregular patterns in the mobility of the same user do affect the variance of the Haversine distance for the same sampling frequency. Yet, the effect is directly proportional to the sampling interval, and becomes apparent in presence of infrequent (*i.e.* hourly or above) sampling. As a result, irregular patterns degrade the quality of the linear model only in settings where the expected accuracy of the trajectory reconstruction is already low (*i.e.* in the order of kilometers).

An important remark is that the only parameter of the fitting curve, *i.e.* the slope α 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 that intercurs between samples. Hence, it can be measured in meters per minute (m/min). From this perspective, our analysis indicates that *adding one minute to the sampling interval used to track one individual leads to an additional positioning error of α meters in her recorded trajectory, irrespective of the absolute span of the sampling interval.*

When looking at the value of α, 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 note some diversity in the values. The heterogeneity of α in our complete dataset is illustrated in Fig. 10b, which displays the CDF of the error-sampling ratio associated to all our user base. Over 98% of users have slopes uniformly distributed between 1 and 6 m/min. Hence, *for the vast majority of individuals, the inaccuracy of their recorded trajectory grows by 1 to 6 meters for each minute added to their movement sampling interval.*

All the results presented above are derived based on our reference dataset of over 1,000 weeks of mobility of 119

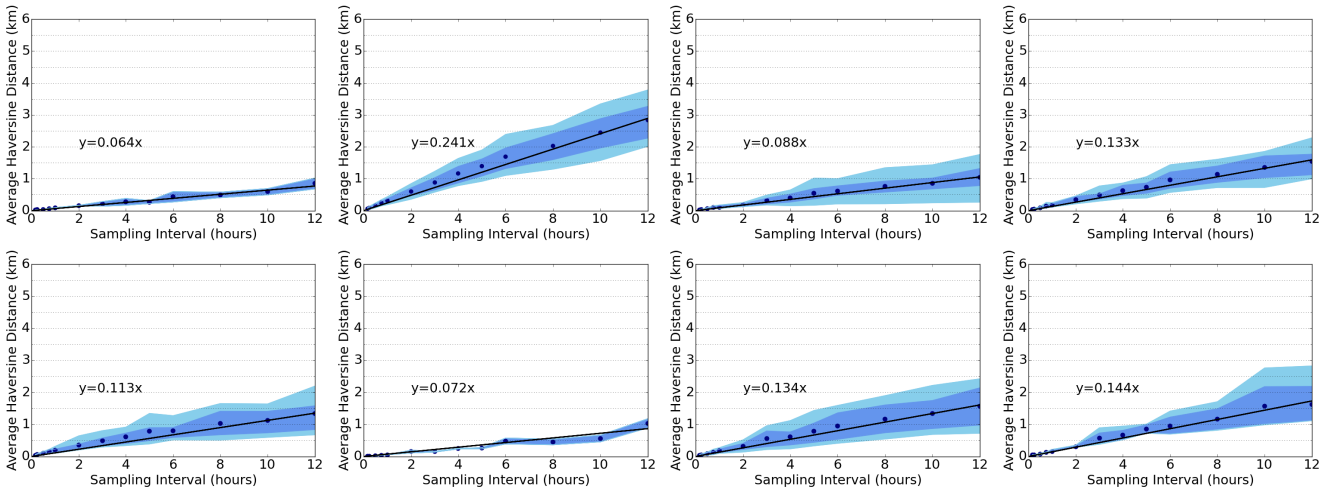
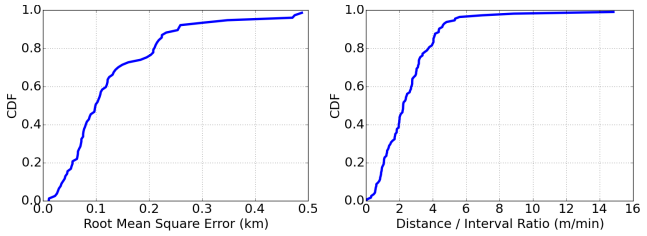


FIGURE 9: Average Haversine distance measured between the original and reconstructed trajectories of eight users, versus downsampling intervals from 10 minutes to 12 hours. Dots represent mean values. Dark and light shaded regions depict the 25-75% and 10-90% quantiles. Solid lines are the linear fittings on average points, whose equation is also reported in each plot. Figure best seen in color.



(a) Quality of fitting (b) Heterogeneity of α

FIGURE 10: (a) Distribution of the RMSE due to the linear approximation of the relation between the average Haversine distance and the sampling interval. (b) Distribution of the ratio α between the average Haversine distance and the sampling interval. Plots are for all users.

different users. While the linear relationship of the sampling interval and accuracy of the reconstructed trajectory yields in all of the considered cases, we cannot generalize our conclusions beyond the dataset we have access to. Yet, as mentioned in Section I, the wide consensus in the scientific community about the widespread and high regularity and predictability of human trajectories lets us argue that our analysis may hold for a vast majority of the users.

IV. DUCTILE LOCATION SAMPLING IN PRACTICE

We leverage the insight on the existence of a single parameter α regulating the relationship between localization accuracy and sampling frequency to design a practical technique for ductile localization, which we name DuctiLoc.

The rationale for our solution stems from the empirical observation that frequent sampling of GPS data tends to quickly drain the battery of a mobile device [1], [2]. A natural solution is then to sample the device position at a reduced frequency, which is dynamically adapted to the movements of the user. However, deciding which frequency should be employed at each time instant is not trivial, and the linear scaling law we identified in the previous section can be used to control the trade-off between energy consumption and localization accuracy.

More formally, the concept underpinning DuctiLoc is that the knowledge of the α value that characterizes the mobility of a given individual is sufficient to estimate the sampling frequency needed to achieve a given localization error. Indeed, a sampling frequency $f^*{=}\alpha/\bar{H}^*$ shall grant a target average error $\bar{H}^*$, according to (4). Several important remarks are in order, as follows.

- Given the high heterogeneity of α values observed in Fig. 10b, the relationship above allows adapting the sampling rate to the mobility habits of each individual in a significant way: for instance, the same localization accuracy can be achieved by sampling the position of volunteers in our dataset at a frequency that can vary sixfold, with obvious implications on the possible energy and resource savings for users whose mobility is characterized by lower α values.
- The target average error $\bar{H}^*$ is a configurable parameter, which allows de-facto controlling the level of localization accuracy, and adapting the sampling frequency f^* accordingly; indirectly, this also entails the possibility of controlling the energy consumption of the positioning tracking process, by varying $\bar{H}^*$.
- The linear model in (4) captures the mean positioning accuracy, hence f^* can be intended as the minimum sampling frequency that guarantees an error $\bar{H}^*$ averaged over all times; variance shall thus be expected in the instantaneous localization performance, and higher sampling rates can be used to combat that when appropriate, such as during periods of significant mobility.

We build on the considerations above to design DuctiLoc, as detailed next.

A. DUCTILOC DESIGN

The operation of our location sampling technique is outlined in Alg. 1. DuctiLoc receives as input the target localization error $\bar{H}^*$, which is the system parameter controlling the desired accuracy of the sampled trajectory; it also needs the value of the user-specific error-sampling ratio α, as defined in Section III-C2.

Algorithm 1: DUCTILOC pseudocode

```
   input: H̄*, α
1  T_max ← H̄*/α
2  do
3      (φ, λ, v) ← samplePosition
4      updateLocation (φ, λ)
5      v̄ ← (1 − γ) · v + γ v̄
6      T_next ← H̄*/v̄
7      T ← min {T_next, T_max}
8      sleep (T)
9  while (1)
```

The algorithm first computes T_{max} (line 1), *i.e.* the maximum time interval between any two sampling events. According to our previous discussion, we set $T_{max} = 1/f^*$, *i.e.* $T_{max} = \bar{H}^*/\alpha$, since this guarantees an average localization error $\bar{H}^*$.

The actual execution loop is then entered (lines 2–9). At each iteration, DuctiLoc samples the current position, collecting information on the latitude ϕ, longitude λ, and speed v (line 3); the latter information is usually provided by the positioning system itself, and can be computed from past location samples otherwise. The latitude and longitude information are used to update the location information, effectively sampling the user's trajectory (line 4). The velocity information is passed through an Exponentially Weighted Moving Average (EWMA) filter, so as to compute a robust estimate $\bar{v}$ of the user's speed over time (line 5).

The speed estimate $\bar{v}$ is employed to increase the sampling frequency above $f^* = 1/T_{max}$ when necessary. Specifically, if the user is found to be moving at a velocity that exceeds α, sampling at every T_{max} generates errors to the right of the mean $\bar{H}^*$, or, in other words, reduces the localization accuracy below the target. While this is expected as $\bar{H}^*$ is an average value obtained over the complete mobility of an individual, it makes sense to take advantage of the available information on the instantaneous velocity to improve DuctiLoc performance. Therefore, we use $\bar{v}$ to compute an alternative speed-based sampling interval T_{next}, and select the minimum of T_{max} and T_{next} as the time to the next sample collection (lines 6–8). We remark that this design factually turns $\bar{H}^*$ from an expectation into an upper-bound to the localization error, since high-mobility situations that would generate values above $\bar{H}^*$ are countered with an increased frequency of sampling; clearly, the scheme does not provide a guaranteed bound, as its performance depends on the precision and responsiveness of the velocity estimation process.

TABLE 2: Pearson's correlation of α with mobility features. For the $p-$values, we adopt the common notation: * $p < 0.1$, ** $p < 0.05$, and *** $p < 0.001$.

Feature	Correlation	$p-$value
Radius of gyration	0.73	***
Speed, average	0.31	**
Speed, standard deviation	0.57	***
Time fraction, mobile	-0.05	*
Time fraction, static	0.05	*
Visited locations	0.47	***
First popular location	0.08	**
Second popular location	0.11	***
Total traveled distance	0.27	***
Displacement	0.22	***
Entropy of visited locations	-0.10	**
Area	0.47	***
Regularity	0.18	***

B. ESTIMATING THE α PARAMETER

The DuctiLoc algorithm in Alg. 1 requires knowledge of the parameter α, which must be adjusted to the mobility of each individual as presented in Section III-C2 for users in our reference dataset. We propose two approaches to estimating α, as presented next.

1) Cold start

The baseline solution to determine α is running DuctiLoc in a *cold start* mode when first launched. In this mode, our scheme collects positioning information at a high frequency for a sufficient amount of time. This allows collecting training data about the mobility of the current user, and running the same procedure presented in Section III-C to compute the value of α: (i) downsampling the recorded data for intervals up to 12 hours, (ii) computing the average Haversine distance between each downsampled trajectory and the original one, and (iii) computing the slope of a linear fitting between the distance and the downsampling interval. We remark that this is equivalent to employing a traditional fixed sampling for the training period, while benefiting from DuctiLoc dynamic sampling afterwards; or, it can be achieved by leveraging historical movement data of the user, if available. In our experiments, we set the cold start sampling interval at one minute, and the collection period to two weeks. These settings allow for an accurate estimate of α that attains a good trade-off of localization accuracy and energy consumption reduction, as shown later in Section V.

2) Warm start

The cost of the cold start mode can be avoided if minimal information about the mobility of the user is available when DuctiLoc is first run. The *warm start* mode takes advantage of the fact that α can be inferred from high-level historical statistics on the movement patterns of an individual. Tab. 2,

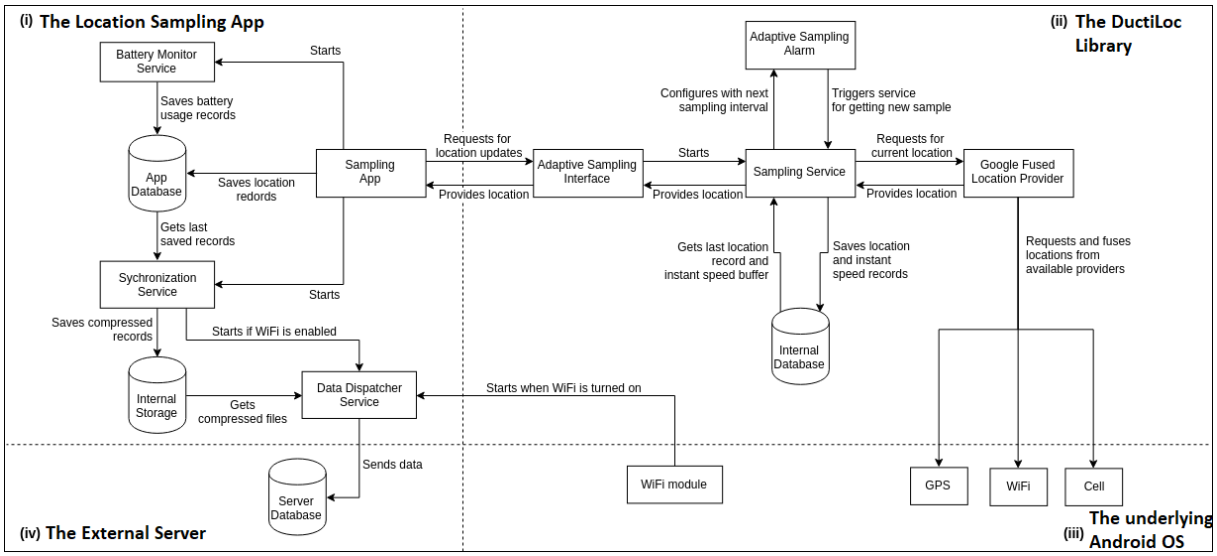


FIGURE 11: Overview of the implementation of DUCTILOC. The (i) location sampling app leverages DUCTILOC, coded as (ii) an Android library that relies on (iii) the basic functionalities of the OS. The app uploads the measurement data to (iv) an external project server.

computed on all users in our reference dataset, shows that the value of α correlates well with a range of features that have been largely used in the literature to characterize individual movement patterns. The Pearson's correlation coefficients are especially high for features that capture the breadth of movements of a person, *i.e.* the radius of gyration, number of visited locations, or area of the convex hull of visited locations: this implies that individuals who have more varied mobility incur an average localization error that grows faster under a sparser sampling. More generally, the observed substantial correlations point at the possibility of predicting the value of α from user-specific mobility statistics.

Indeed, a simple Multiple Linear Regression (MLR) model using a combination of five features with low collinearity (*i.e.* radius of gyration, total travelled distance, displacement, fraction of time the user is mobile, and regularity) achieves a coefficient of determination R^2 of 0.88, *i.e.* allows explaining 88% of the variance of α, with p-values lower than 0.05. Therefore, if similar statistics were available to the user based on previous mobility patterns, DuctiLoc could leverage them to derive a very good approximation of the value of α, without a need for the high-frequency sampling a cold start of the method relies upon. We remark that such information about the movement of the user can be represented with minimal data size, and is fully privacy preserving, which ease its permanent storage with respect to, *e.g.* complete trajectory data.

C. DUCTILOC IMPLEMENTATION

We implement DuctiLoc as a self-contained library for the Android operating system (OS), which can then be embedded in any Android app. In order to experiment with our technique, we also develop a dedicated Android location sampling app that relies on the DuctiLoc library, and transfers the collected data to a centralized server for further processing and analysis of the system performance.

Fig. 11 illustrates the organization of our DuctiLoc implementation. The diagram is separated into four components: (i) the location sampling app, (ii) the DuctiLoc library, (iii) the underlying Android OS, and (iv) the external server used for data collection. The location sampling app, parametrized with the values of $\bar{H}^*$ and α, requests location updates using the DuctiLoc library. The library internally relies on the Google Fused Location Provider to obtain location samples from the Android OS. Once DuctiLoc obtains a new location, it saves both the location and the current instant speed (calculated in our implementation using the previously collected sample) in an internal database. Next, DuctiLoc updates the EWMA-based speed $\bar{v}$ and calculates the time T to the next sample collection as per Alg. 1. The location sampling app receives the location samples from the DuctiLoc library, and saves them in an internal database. At the same time, the app is also responsible for collecting statistics about its own utilization of device battery and CPU, via the Android *dumpsys* tool. Dumpsys is a tool built into the Android OS, which allows obtaining precise information about the status of running services. Inside the location sampling app, a synchronization service compresses all collected records and stores the compact files in the internal storage of the mobile device; the data is finally sent to the external server when a Wi-Fi connection is available.

V. EXPERIMENTAL EVALUATION

We evaluate the performance of DuctiLoc via real-world experiments involving a small set of six volunteers located in countries in Europe and South America. The testers were

carefully selected so as to compose a representative group of individuals with markedly different mobility habits, from local dwelling to regular long-distance commuting, as proven by the results in Section V-B. Next, we first present the measurement campaign setup, and then report and discuss its results.

A. CAMPAIGN SETUP

All volunteers installed our experimental suite in their personal Android smartphones, upon signing a consent form that detailed the purpose of the study and of the collected information. Specifically, each data subject run for the duration of the evaluation campaign three different apps in parallel, denoted as follows:

- DuctiLoc: the location sampling app that relies on the DuctiLoc library to perform the sampling process, as presented in Section IV-C. As historical information about the mobility of the volunteers was not available at the start of the experiments, the cold start mode of DuctiLoc was employed, by collecting positioning data at 1-minute resolution over two consecutive weeks as previously mentioned.
- Fixed-sampling: an app performing location sampling at a fixed periodicity of one minute, which is used as (i) a way to collect ground-truth information about the trajectory of the user, and (ii) a worst-case performance baseline for energy consumption and CPU usage in the mobile device.
- Sensor-based: an app using an accelerometer-based location sampling process, by exploiting the fact that low-power accelerometer sensors embedded in modern smartphones can identify user movements so as to to avoid unnecessary location sampling in stationary conditions [36], [37]. More precisely, at every three minutes the app collects a short burst of measurements from the in-built accelerometer of the mobile phone, during two seconds. If the variance of values in the measured burst is greater than a threshold, a movement is detected, and a location sample is collected. For the choice of threshold, we selected the value 0.5 $(m/s^2)^2$. For the selection of this value, we got inspired by the paradigm proposed in [38]. We firstly applied a variety of values on our data collected by our volunteers as thresholds for detecting movement. We found 0.5 $(m/s^2)^2$ to give the best results. Then, we employed the data used in [39] to separately study the acceleration variance for humans under active, driving, walking and inactive state. We found driving to be the mobile status with the lower acceleration variance, with a median value of 0.72 $(m/s^2)^2$. Consequently, we chose 0.5 $(m/s^2)^2$ as a conservative threshold for our comparison experiments.

All three apps collect and send to the same external server the following data: (i) device-related information, including a unique random identifier for data pseudonymization, the Android version, and the manufacturer, model, and release of the device; (ii) the collected location samples, including the latitude, longitude, horizontal accuracy, and timestamp of the location; (iii) the energy consumption information, including the estimated power utilization of the app, in mAh; (iv) CPU-related information, including the total CPU time used by the app, in ms. The collected data allows comparing the three sampling methods in terms of energy consumption and CPU usage. Additionally, it lets us evaluate the localization error incurred by the trajectories reconstructed from DuctiLoc and the reference accelerometer-based solution with respect to the ground truth provided by the fixed-sampling baseline.

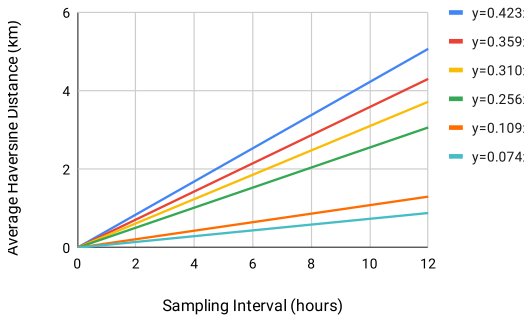


FIGURE 12: Linear fittings of the error versus sampling interval for each volunteer, and associated α values.

Experiments were carried out for a continued period of seven weeks. During this period, we had the volunteers vary the target error parameter of DuctiLoc, so as to investigate the capability of the mechanism to adapt to different user's settings; to this end, the data subjects updated the setting of $\bar{H}^* \in \{50, 100, 250, 500, 1000\}$ meters at regular time intervals of two weeks.

B. RESULTS

As a preliminary result to our performance evaluation, we show the linear fittings to the average Haversine distance as a function of the sampling interval, observed during the cold start phase of DuctiLoc. Fig. 12 summarizes the outcome for all volunteers. The different slopes and their associated coefficients (reported in the legend) illustrate that the volunteers involved in our experiments yield a variety of α values. Specifically, we observe values of α ranging from 1.23 m/min to 7.0 m/min, which are even more diverse values than those found in Section III-C2 for a user base of 119 individuals. Ultimately, this demonstrates that our choice of volunteers, although limited in size, encompasses a substantial variety of mobility patterns.

1) Localization accuracy

The main performance figures are in Fig. 13, which portray (i) the sampling rate, in the top plot, and (ii) the localization accuracy, in the bottom plot, for both DuctiLoc and the sensor-based solution that relies on accelerometer information. Results are shown as a function of the target error $\bar{H}^*$ used in DuctiLoc.

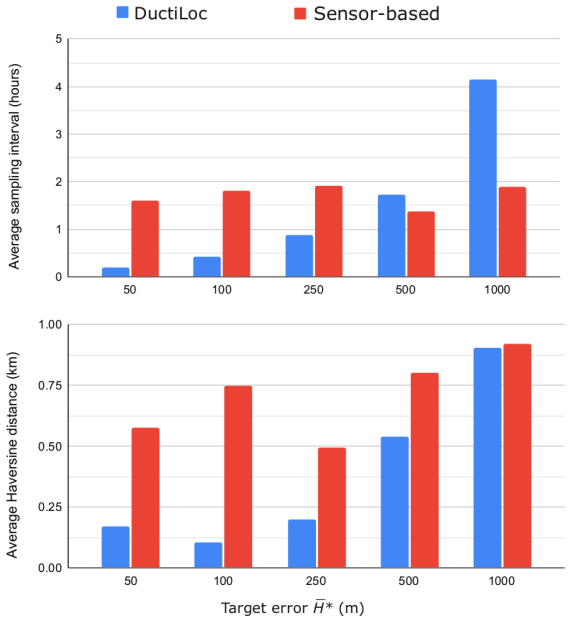


FIGURE 13: Localization accuracy performance of DUCTILOC and of the sensor-based benchmark, as a function of the target error $\bar{H}^*$. Top: average sampling interval across all volunteers; the fixed-sampling baseline yields a constant value at one minute. Bottom: average Haversine error with respect to the fixed-sampling baseline, across all volunteers.

We observe in the top plot that DuctiLoc can effectively adopt the sampling frequency to the expected accuracy, which results in a variable interval between subsequent samples. The average value of the interval varies from ten minutes, when $\bar{H}^*$ is set to 50 m, to more than four hours, when $\bar{H}^*$ is 1000 m. The sensor-based solution is obviously insensitive to the target error, and yields a fairly constant sampling interval with a mean value around 1.5 hours. The fixed-sampling approach clearly results in an identical rate of one sample per minute for any $\bar{H}^*$, which is not shown in the plot as it would be barely visible.

The difference above reflects on the mean localization error, computed using the average Haversine distance as explained in Section III-C1. Note that the positioning data returned by the fixed-sampling baseline is used in this case as the ground truth for that calculation: in other words, the result in the bottom plot of Fig. 13 can be interpreted as the error incurred by DuctiLoc and the sensor-based location sampling with respect to the trajectory information obtained with the fixed-sampling approach. The localization error of DuctiLoc grows with $\bar{H}^*$ as expected; in fact, it is remarkably close to the target, which demonstrates the capability of our method to adapt to the user requirements in terms of expected accuracy of position tracking.

Instead, the accuracy of the sensor-based solution fluctuates between 500 m and 800 m. This difference is to be ascribed to the fact that experiments with diverse values of $\bar{H}^*$ were carried out in separate weeks, during which the volunteers may have modified their mobility patterns due to, *e.g.* special events, vacations, or work and personal businesses. Here, the fact that the average sampling interval recorded by DuctiLoc is very consistent with the selected $\bar{H}^*$ further proves the robustness of our method to fluctuations in the movement habits of the user.

It is worth remarking that it is possible to fine-tune the sensor-based location sampling so as to reduce the error; for instance, this can be achieved by tailoring the sensitivity threshold for the variance of the accelerometer values. However, no inherent relationship exists between that threshold and the localization accuracy; also, the correct threshold will likely vary on an individual basis. Therefore, finding the correct parametrization of the sensor-based sampling that ensures a given localization error is reduced to a cumbersome trial-and-error task. Instead, the direct relation unveiled in Section III-C2 and used by DuctiLoc allows our solution to elegantly overcome this kind of parametrization problem.

Also, one may argue that the absolute accuracy performance of the sensor-based approach can be improved, *e.g.* by reducing the 3-minute interval between accelerometer data burst collections. While this is true, increasing the usage of the sensor would also grow the consumption of batteries and computational resources in the mobile device; and, those are already penalizing the sensor-based solution with respect to DuctiLoc under the current settings of accelerometer querying frequency. Such a behavior is discussed next, as part of our observations on the device resource utilization under different sampling schemes.

2) Energy efficiency and CPU usage

Reducing the location sampling frequency has a direct positive impact on the utilization of resources in the mobile device, as the OS positioning system (which possibly involved activating or probing the GPS receiver) is less solicited. Therefore, by adapting the sampling rate to the minimum value required to achieve the target localization accuracy, DuctiLoc is expected to yield savings on both battery and CPU.

This is demonstrated in Fig. 14, which shows the fractional gains in power consumption, in the top plot, and processing time, in the bottom plot, attained by DuctiLoc with respect to the fixed-sampling baseline. The plots illustrate the performance as a function of the target error $\bar{H}^*$: depending on that parameter, the reduction of battery usage ranges between 60% and 98%, and that of CPU is from 40% to 94%.

An important remark is that the parameter $\bar{H}^*$ allows controlling the level of energy efficiency of the localization process. Indeed, Fig. 14 depicts a clear negative correlation between the value of $\bar{H}^*$ and the resulting power and CPU consumption. Therefore, by acting on the desired accuracy input to DuctiLoc, the user has an effective knob to tune the demand for system resources entailed by the location sampling. This is not the case with the sensor-based approach, whose performance is also portrayed in Fig. 14: the power and CPU time consumption are fairly constant across

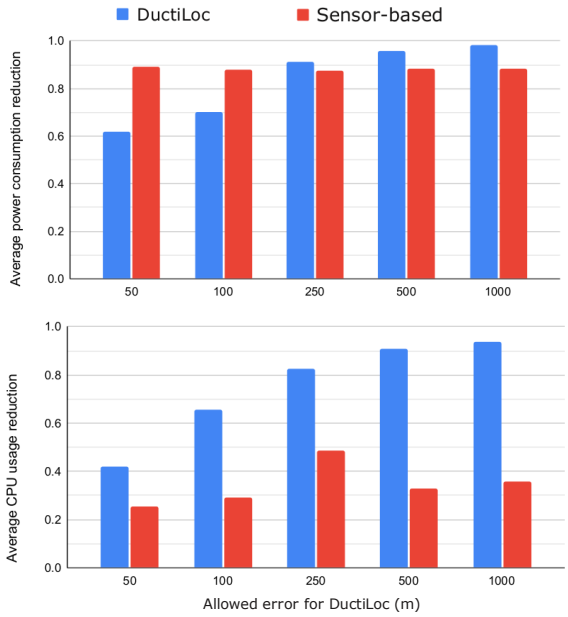


FIGURE 14: Utilization of mobile device resources by DUCTILOC and the sensor-based benchmark with respect to the fixed-sampling baseline, as a function of the target error $\bar{H}^*$. Top: average fractional reduction in power consumption across all volunteers. Bottom: average fractional reduction in CPU usage across all volunteers.

experiments with different values of $\bar{H}^*$, and the benchmark solution does not offer direct ways to control the system resource utilization.

Also, we note that the average reduction in energy usage of the sensor-based app is around 90%, whereas the CPU time required for localization is reduced by 25% to 49%. Thus, DuctiLoc largely outperforms the sensor-based approach in terms of CPU utilization cut, as it can save two to three times more computational resources. In terms of battery consumption, the trade-off between localization accuracy and power drain can be appreciated by comparing the bottom plot of Fig. 13 and the top plot of Fig. 14. Again, the outcome is clearly in favor of DuctiLoc: for instance, when $\bar{H}^*$ is set to 250 m, our solution yields a slightly lower energy usage while providing a 75% lower positioning error.

VI. DISCUSSION AND CONCLUSIONS

We found that the average error incurred by trajectories reconstructed from periodic samples scales linearly with the constant sampling interval, as shown in [3]. The result was first identified through measurement data collected by 119 users during more than 1,000 weeks, and then confirmed in a small-scale experimental evaluation with 6 volunteers; the consistence of the linear behavior across our heterogeneous user base lets us hypothesize that such a scaling law could be a universal property of human mobility.

The linearity of the relationship between error and sampling interval explains the absence of an operational point for the effective sampling of human movements, which was also corroborated by the outcome of a spectral analysis of individual movement patterns. Therefore, the trade-off between localization accuracy and sampling frequency can be fully modeled via a single parameter, *i.e.* the slope α of the linear scaling law. The slope quantifies the added error induced by a unit increase in the sampling interval, and we proved that it can vary by almost one order of magnitude depending on the target person.

Building on these insights, we designed DuctiLoc, a ductile location sampler that is lightweight and can adjust the sampling frequency to the preferred accuracy level of each individual. Experiments with real-world users demonstrated that DuctiLoc offers control on the positioning error, which can be also used as a means to effectively modulate the usage of energy and CPU resources by the localization process in the mobile device.

In addition to the practical application above, the seemingly general linear scaling law of the positioning error with respect to the sampling interval may also be a very useful tool in a number of other contexts. Thus, the importance of controlling the trade-off between sampling rate and accuracy of human mobility is further evidenced. Examples include applications in the following domains.

- In Location-Based Services (LBS) the excessively frequent collection of user locations is expensive from both energy and communication perspectives, and raises privacy concerns. Here, our results may help more informed decisions on the minimum frequency of position querying that can support each service. For instance, in location-based social network services where users share their attraction routes, favourite hotspots or traffic jam events, the important information lies in the visited locations and the amount of time spent there [40]–[42]; in these scenarios, it is important to be able to fetch locations only during specific (and long) stays, and our solution can help achieving that efficiently, avoiding unnecessary tracking in between those stays.
- Precise knowledge of subscribers' locations is valuable information for mobile operators, for both network management and value-added service development [32]. Yet, operators make today very limited use of active probing to update the location of inactive user equipment, as it is an expensive procedure, and favor less controllable passive measurements [31]. In typical 4G/5G deployments, user positions are probed in a deterministic way, with a periodicity of a few hours that is identical for all subscribers. Operators could instead use our findings to develop active probing solutions that are not uniform across the whole user population, but are instead tailored to the mobility of each subscriber, ensuring lower cost and higher accuracy.
- Mobility data compression is an open field of research, which seeks solutions aiming at simplifying individual trajectories so as to allow storing them in very large amounts [19], [20]. As discussed in Section II-A, our

problem is a superset of the trajectory compression one. Therefore, our findings can also be leveraged to help the task of trajectory compression and make it more viable in practice at the data collection stage, by sampling the user movements with a loss of information that is controllable.

To conclude our discussion, we would like to clarify the limitations of our work. First, our analysis in Section III is based on a dataset of trajectories of 119 individuals, whereas our experimental evaluation of DuctiLoc in Section V relies six testers; while the acknowledged regularity of human mobility yields promises for the wider applicability of our results, at this time we cannot generalize them beyond such data subjects and additional tests would be needed to that end.

As a second point, the question we posed in Section I and the subsequent analyses aim at being as general as possible. However, specific applications that leverage some form of human mobility sampling may have unique requirements that do not fit our approach. For instance, different services may rely on forms of accuracy that are not well represented by our average Haversine error; they may have requirements in terms of maximum variance of the accuracy that is not captured by our mean analysis; or they may accommodate sampling approaches that are not constant but adaptive to, *e.g.* the social or environmental context of the user or device. It is thus important to understand that our study does not aim at providing direct support to the design of any specific practical service relying on mobile device localization: instead, our investigations are a sensible starting point for the tailored design of applications built on top of trajectory sampling.

ACKNOWLEDGMENT

This work was supported by the Regional Government of Madrid through grant no. 2019-T1/TIC-16037 NetSense. We would also like to appreciate the support from INRIA, STIC AmSud LINT project (code 22-STIC-07), and ANR MITIK project (PRC AAPG2019).

REFERENCES

[1] Y. Qi, C. Yu, Y.-J. Suh, and S. Y. Jang, "Gps tethering for energy conservation," in *Wireless Communications and Networking Conference (WCNC), 2015 IEEE*. IEEE, 2015, pp. 1320–1325.

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

[3] P. Katsikouli, A. C. Viana, M. Fiore, and A. Tarable, "On the sampling frequency of human mobility," in *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, Dec 2017, pp. 1–6.

[4] 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*. ACM, 2008, pp. 312–321.

[5] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, p. 779, 2008.

[6] T. S. Azevedo, R. L. Bezerra, C. A. V. Campos, and L. F. M. de Moraes, "An analysis of human mobility using real traces," in *2009 IEEE Wireless Communications and Networking Conference*, April 2009, pp. 1–6.

[7] R. Becker, R. Cáceres, K. Hanson, S. Isaacman, J. M. Loh, M. Martonosi, J. Rowland, S. Urbanek, A. Varshavsky, and C. Volinsky, "Human mobility characterization from cellular network data," *Commun. ACM*, vol. 56, no. 1, pp. 74–82, Jan. 2013. [Online]. Available: http://doi.acm.org/10.1145/2398356.2398375

[8] G. Chen, A. Carneiro, and M. Fiore, "Takeaways in large-scale human mobility data mining," in *IEEE International Symposium on Local and Metropolitan Area Networks*, 2018.

[9] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, no. 1, pp. 3–42, Apr 2006. [Online]. Available: https://doi.org/10.1007/s10994-006-6226-1

[10] H. Senaratne, M. Mueller, M. Behrisch, F. Lalanne, J. Bustos-Jiménez, J. Schneidewind, D. Keim, and T. Schreck, "Urban mobility analysis with mobile network data: A visual analytics approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 5, pp. 1537–1546, 2018.

[11] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the crowd: The privacy bounds of human mobility," *Scientific Reports*, vol. 3, 2013.

[12] K. Sila-Nowicka, J. Vandrol, T. Oshan, J. A. Long, U. Demšar, and A. S. Fotheringham, "Analysis of human mobility patterns from gps trajectories and contextual information," *International Journal of Geographical Information Science*, vol. 30, no. 5, pp. 881–906, 2016. [Online]. Available: https://doi.org/10.1080/13658816.2015.1100731

[13] N. Deblauwe and G. Treu, "Hybrid gps and gsm localization — energy-efficient detection of spatial triggers," in *2008 5th Workshop on Positioning, Navigation and Communication*, March 2008, pp. 181–189.

[14] C.-W. You, P. Huang, H. hua Chu, Y.-C. Chen, J.-R. Chiang, and S.-Y. Lau, "Impact of sensor-enhanced mobility prediction on the design of energy-efficient localization," *Ad Hoc Networks*, vol. 6, no. 8, pp. 1221 – 1237, 2008, energy Efficient Design in Wireless Ad Hoc and Sensor Networks. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1570870507001667

[15] H. Ma, D. Zhao, and P. Yuan, "Opportunities in mobile crowd sensing," *IEEE Communications Magazine*, vol. 52, no. 8, pp. 29–35, Aug 2014.

[16] K. K. Rachuri, C. Mascolo, M. Musolesi, and P. J. Rentfrow, "Sociablesense: Exploring the trade-offs of adaptive sampling and computation offloading for social sensing," in *Proceedings of the 17th Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '11. New York, NY, USA: ACM, 2011, pp. 73–84. [Online]. Available: http://doi.acm.org/10.1145/2030613.2030623

[17] J. Paek, J. Kim, and R. Govindan, "Energy-efficient rate-adaptive gps-based positioning for smartphones," in *Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '10. New York, NY, USA: ACM, 2010, pp. 299–314. [Online]. Available: http://doi.acm.org/10.1145/1814433.1814463

[18] Z. Zhuang, K.-H. Kim, and J. P. Singh, "Improving energy efficiency of location sensing on smartphones," in *Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '10. New York, NY, USA: ACM, 2010, pp. 315–330. [Online]. Available: http://doi.acm.org/10.1145/1814433.1814464

[19] 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*, ACM, 2014, pp. 371–380.

[20] 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*, ACM, 2010, pp. 402–405.

[21] B. Thierry, B. Chaix, and Y. Kestens, "Detecting activity locations from raw gps data: a novel kernel-based algorithm," *International Journal of Health Geographics*, vol. 12, no. 1, p. 14, 2013.

[22] 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*, vol. 392, no. 6, pp. 1459 – 1473, 2013.

[23] 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)*, ACM, 2009, pp. 33–40.

[24] 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)*, ACM, 2011, pp. 8:1–8:6.

[25] C. Long, R. C.-W. Wong, and H. V. Jagadish, "Direction-preserving trajectory simplification," *Proc. VLDB Endow.*, vol. 6, no. 10, pp. 949–960, Aug. 2013.

[26] 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), ACM, 2006, pp. 19–26.

[27] 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)*, ICST, 2008, pp. 52:1–52:10.

[28] 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)*, ACM, 2010, pp. 341–352.

[29] 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*, ACM, 2011, pp. 305–316.

[30] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.

[31] M. Ficek and L. Kencl, "Inter-Call Mobility model: A spatio-temporal refinement of Call Data Records using a Gaussian mixture model," in *Proceedings of IEEE INFOCOM*, 2012, pp. 469-477.

[32] 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)*, ACM, 2013, pp. 321–326.

[33] 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)*, ACM, 2010, pp. 239–254.

[34] L. Ang, K. P. Seng, A. M. Zungeru and G. K. Ijemaru, "Big Sensor Data Systems for Smart Cities", *IEEE Internet of Things Journal*, vol. 4, no. 5, pp. 1259-1271, Oct. 2017.

[35] Yu, X., Zhao, H., Zhang, L., Wu, S., Krishnamachari, B., and Li, V. O., "Cooperative sensing and compression in vehicular sensor networks for urban monitoring," *IEEE International Conference on Communications*, pp. 1-5. May, 2010.

[36] J. Kwapisz, G. Weiss, and S. Moore, "Activity recognition using cell phone accelerometers," *ACM SigKDD Explorations Newsletter*, 2011, vol. 12, no 2, p. 74–82.

[37] M. Puyau, A. Adolph, F. Vohra, I. Zakeri, and N. Butte, "Prediction of activity energy expenditure using accelerometers in children," *Medicine & Science in Sports & Exercise*, 2004, vol. 36, no 9, p. 1625-1631.

[38] D. Brown, S. LaPoint, R. Kays, W. Heidrich, F. Kümmeth, and M. Wikelski, "Accelerometer-informed GPS telemetry: Reducing the trade-off between resolution and longevitys," *Wildlife Society Bulletin*, 2012, vol. 36, no 1, p. 139-146.

[39] D. Garcia-Gonzalez, D. Rivero, E. Fernandez-Blanco, M.R. Luaces. "A Public Domain Dataset for Real-Life Human Activity Recognition Using Smartphone Sensors," *Sensors* 2020, vol. 20, no 8: 2200.

[40] Z. Wang, M. Lu, X. Yuan, J. Zhang, H. van de Wetering, "Visual Traffic Jam Analysis Based on Trajectory Data," *IEEE Transactions on Visualization and Computer Graphics* 2013, vol. 19, no. 12, p. 2159-2168.

[41] R. Ong, F. Pinelli, R. Trasarti, M. Nanni, C. Renso, S. Rinzivillo, F. Giannotti, "Traffic Jams Detection Using Flock Mining," *ECML PKDD* 2011.

[42] L. Ferrari, A. Rosi, M. Mamei, F. Zambonelli, "Extracting Urban Patterns from Location-based Social Networks," *ACM SIGSPATIAL International Workshop on Location-Based Social Networks* 2011.



PANAGIOTA KATSIKOULI is a researcher at the University of Copenhagen. She received the Diploma and M.S. degrees in computer engineering and informatics from the Polytechnic University of Patras, Greece, in 2011 and 2013, respectively, and the Ph.D. degree in informatics from the University of Edinburgh, U.K., in 2017. From 2017 until now she has been a researcher in various institutes, such as Inria, University College of Dublin and Technical University of Denmark. Her research interests include distributed algorithms, blockchain technology, human mobility, data analytics, applications of machine learning and distributed algorithms for mobility data.



DIEGO MADARIAGA is a Ph.D. student in Computer Science at the University of Chile. His research interests include machine learning for anticipatory networking, user mobility, network privacy/security and QoS monitoring.



ALINE CARNEIRO VIANA is a permanent Reseach Director (DR) at Inria, where she leads the team TRiBE team. After a 1-year sabbatical leave at the TKN Group of the TU-Berlin, Germany, she got her habilitation degree from UPMC - Sorbonne Universités, France in 2011. Dr. Viana got her PhD in Computer Science from the UPMC - Sorbonne Universités in 2005. Her research addresses the design of solutions for tactful networking, smart cities, mobile and self-organizing networks with the focus on human behavior analysis. She is a recipient of the French Scientific Excellence award since 2015 and for 6 years now, was nominated in 2016 as one of the "10 women in networking/communications that you should Watch", and got got the Medal of Honor 2021 Edition from Federal University of Goiás, Brazil, for her distinguished society contributions.



ALBERTO TARABLE Alberto Tarable (S'00-M'02) received the Laurea degree (summa cum laude) in 1998 and the Ph.D. degree in Electronic Engineering in February 2002, both from Politecnico di Torino. From 2002 to 2012, he worked as a researcher in the Department of Electronics and Telecommunications of Politecnico di Torino. From 2012, he holds a research position in the Institute of Electronics, Computer and Telecommunication Engineering of the Italian National Research Council. His research interests include signal processing for wireless networks, physical-layer design, smart radio environments and quantum communications.



MARCO FIORE is Research Associate Professor at IMDEA Networks Institute and CTO at Net AI Tech Ltd. He received MSc degrees from University of Illinois at Chicago in IL, USA and Politecnico di Torino in Italy, a PhD degree from Politecnico di Torino in Italy, and a Habilitation a Diriger des Recherches (HDR) from Université de Lyon in France. He held tenured positions as Associate Professor at Institut National des Sciences Appliquées de Lyon in France, and Researcher at Consiglio Nazionale delle Ricerche in Italy. He has been a visiting researcher at Rice University in TX, USA, Universitat Politecnica de Catalunya in Spain, and University College London (UCL) in UK. Dr. Fiore is a senior member of IEEE, and a member of ACM. He was a recipient of a European Union Marie Curie fellowship and a Royal Society International Exchange fellowship. He leads the Networks Data Science group at IMDEA Networks Institute, which focuses on research at the interface of computer networks, data analysis and machine learning.