Moving and Calling: Mobile Phone Data Quality Measurements and Spatiotemporal Uncertainty in Human Mobility Studies

Chapter · May 2013

DOI: 10.1007/978-3-319-00615-4_14,

CITATIONS
11

READS
223

4 authors:

Ana-Maria Olteanu Raimond
Institut national de l'information géographiq…
48 PUBLICATIONS **237** CITATIONS
SEE PROFILE

Corina Iovan
Institut national de l'information géographiq…
8 PUBLICATIONS **76** CITATIONS
SEE PROFILE

Thomas Couronné
Telenor
30 PUBLICATIONS **313** CITATIONS
SEE PROFILE

Zbigniew Smoreda
Orange Labs, Paris, France
105 PUBLICATIONS **1,599** CITATIONS
SEE PROFILE

Moving and Calling: Mobile Phone Data Quality Measurements and Spatiotemporal Uncertainty in Human Mobility Studies

Corina Iovan, Ana-Maria Olteanu-Raimond, Thomas Couronné and Zbigniew Smoreda

Abstract In the past few years, mobile network data are considered as a useful complementary source of information for human mobility research. Mobile phone datasets contain massive amount of spatiotemporal localization of millions of users. The analyze of such huge amount of data for mobility studies reveals many issues such as time computation, users sampling, spatiotemporal heterogeneities, semantic incompleteness. In this chapter, two issues are addressed: (1) location sampling aiming at decreasing computation time without losing useful information on the one hand and to eliminate data considered as noise in the other hand and (2) users sampling whose goal is to select users having relevant information. For the first issue two measures allowing eliminating redundant information and ping-pong positions are proposed. The second issue requires the definition of a set of measures allowing estimating mobile phone data quality. New methods to qualify mobile phone data at local and global level are proposed. The methods are tested on one-day mobile phone data coming from technical mobile network probes.

1 Introduction

Human mobility analysis is an important issue in social sciences, and mobility data are among the most sought-after sources of information in economic forecasting, geography, transportation engineering and urban planning. Mobility studies have received increasing attention in the past few years, particularly the ones conducted on mobile phone location data (González et al. 2008; Song et al. 2010;

C. Iovan (✉) · A.-M. Olteanu-Raimond · T. Couronné · Z. Smoreda
Sociology and Economics of Networks and Services department,
Orange Labs R&D, Paris, France
e-mail: corina.iovan@orange.com

A.-M. Olteanu-Raimond
Laboratoire Cogit, Institut Géographique National, Saint-Mandé, France

D. Vandenbroucke et al. (eds.), *Geographic Information Science at the Heart of Europe*, Lecture Notes in Geoinformation and Cartography,
DOI: 10.1007/978-3-319-00615-4_14,

Andrienko et al. 2010; Onnela et al. 2011; Calabrese et al. 2011a, b; Olteanu Raimond et al. 2012; Becker et al. 2013). Based on mobile network data, different aspects can be tackled: collective spatial and temporal mobility patterns, patterns of travel behavior at the individual level, collective behavior at large scales, trajectory pattern mining or trajectory clustering (Song et al. 2010; Vieira et al. 2010). Human mobility models were successfully applied in various topics such as activities detection (Olteanu Raimond et al. 2012; Phithakkitnukoon et al. 2010), human mobility prediction (Song et al. 2010), tourism applications (Olteanu Raimond et al. 2011; Steenbruggen et al. 2011), traffic estimation (Phithakkitnukoon et al. 2010; Caceres et al. 2007) or commuting patterns (Zhang et al. 2010; Sevtsuk and Ratti 2010; Yuan et al. 2011).

Such datasets, daily collected by cellular service providers for billing and troubleshooting purposes and appropriately anonymized for privacy, contain massive amount of spatiotemporal localization of millions of users. The analysis of such a large amount of data is very costly from the time computation point of view. Moreover, they are often incomplete and/or have heterogeneous spatiotemporal resolutions. Thus, the use of mobile phone data for socioeconomic forecasting, in general, and human mobility studies, in particular, requires the sampling of data according to some selection criteria to create a subsample of statistically sound records, the definition of different assumptions to improve data by adding semantic information or the definition of adapted models to infer human behavior.

In this chapter, we focus on sampling issues. Two aspects are considered. The first concerns location sampling which consists of eliminating locations considered as redundant, on the one hand, and erroneous (i.e., locations produced by ping-pong phenomenon), on the other hand. Location sampling improves both computation time by eliminating redundant and erroneous locations but also mobile phone data quality by detecting and eliminating locations which could introduce a bias. The second aspect concerns user sampling, i.e., how to select users described by data having a minimal quality required to mobility analysis? We propose a user sampling approach based on the definition of a set of measures allowing estimating mobile phone data quality.

Mobile phone data quality estimation has already been studied in the literature. For example, mobile phone individual trajectories were compared with actual individual trajectories provided from GPS (Kang et al. 2012; Schulz et al. 2012) or with data access records (Ranjan et al. 2012; Zhao et al. 2011) in order to determine the bias and the characteristics of human mobility when mobile phone data are used. Measures allowing characterization of individual trajectory (e.g. length, travel distance, direction) (Andrienko et al. 2011; Hasan et al. 2012) or the territory of trajectories such as entropy (Song et al. 2010; Ranjan et al. 2012), eccentricity (Kang et al. 2012), radius of gyration (González et al. 2008; Song et al. 2010; Ranjan et al. 2012) and convex hull were proposed end tested (Csáji et al. 2012). Although the high number of measures proposed in the literature, the list is not exhaustive and these measures do not cover the temporal aspect of data, which is of great importance when studying location and user sampling.

The chapter is structured as follows. In the Sect. 2, mobile phone data are briefly described. Section 3 introduces the proposed measures to filter locations. In Sect. 4 user sampling is discussed. We first present the impact that user sampling methods could have on human mobility studies by analyzing the correlation between the communication and itinerancy events. Then, the new measures to qualify mobile phone data (local and global measures) and a decision process allowing to chose "representative" users (e.g. user sampling) are described. Section 5 concludes and suggests some directions for future work.

2 Mobile Phone Data

Each telecom operator collects and stores for a given period customers' mobile phone activities for billing or for technical measurement purposes. This type of collection is called "passive collection", since recordings are made automatically. There are three main types of mobile phone data collected through "passive collection": Call Detail Records (CDR) data representing cell phone billing records, probe data and Wi-Fi data. In this chapter, only probe data are briefly described, since these data are used to test the proposed measures. For more information about mobile phone data for human mobility studies, interested readers can refer to (Smoreda et al. 2013).

Probe data are issued from mobile network probes (MSC data), they are anonymous (each mobile phone SIM identifier is replaced with an identifier consisting in a unique integer) and contain both cell localized communication events (i.e. calls and SMS) and itinerancy events: handover (HO) and location area update (LAU). The location of mobile phone users is limited to the base station location. The base station is composed by at least three antennas, each antenna having a spatial coverage. Figure 1 shows an example of mobile phone localization according to different events that occur.

HO data are generated while an active communication is transferred from one cell of the mobile network to another, when a mobile device is on the move; LAU records are generated when a device changes location area, even if the user is not in communication (for Paris, a location area groups on average 150 cells).

In this chapter, anonymous data from a French cellular network operator are used. Spatially, it covers the Parisian region (12,012 km^2–4,638 m^2), and contains recordings of one weekday (Thursday, the 2nd of April 2009) of over 4 million mobile phone users (accounting for a total of 122,208,870 records).

3 Location Sampling

The goal of location sampling is to remove location points considered as "noise". Two such cases are considered here, duplicate location points (Sect. 3.1) and location points generated by the ping-pong phenomenon (Sect. 3.2).

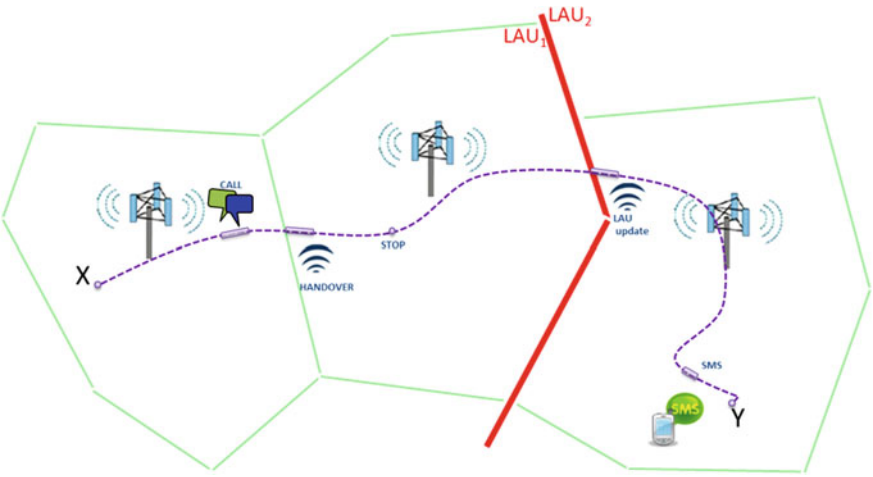

Fig. 1 An example of mobile phone network localization data types for one user travelling from X to Y

Location sampling is especially important when dealing with a huge number of records and consists in an efficient and lossless data reduction strategy. We propose to apply location sampling user by user, thus we use the concept of trajectory for a better understanding of the proposed measures and notations.

A user trajectory is defined as a set of locations in time and space. For each record, a geometry point in a geographic coordinate system is added. In the remainder of this article, by language abuse, the term point will be used to identify a user's location in cartographic coordinates.

Let t_j be the trajectory of user u_i defined as a sequence of points: $p_k \in P$, $P = \{p_1, p_2, \ldots, p_n\}$. A simplified illustration of a user's trajectory is given in Fig. 2 where a user's points p_k, defined by the location of the base station (latitude, longitude) and timestamp (T) serving an event (call, SMS, itinerancy), are sequentially connected to form a trajectory illustrated on a 2D plan.

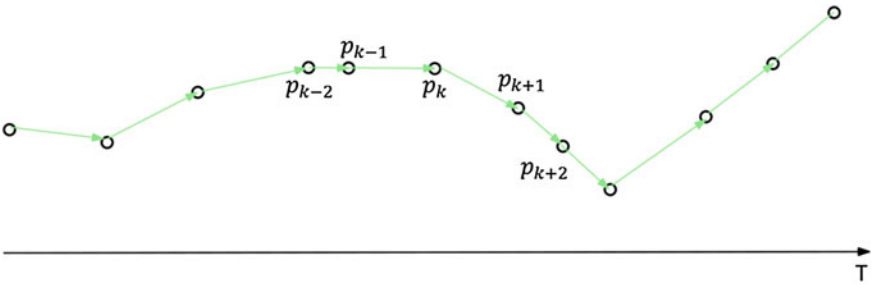

Fig. 2 Creating a user's trajectory t_j from successive location points p_k

3.1 Duplicated Location Points

Let's consider n consecutive locations, having the same geographic coordinates and recorded in a time interval below a threshold. We consider that these duplicate locations do not provide additional information. We agree that locations do not represent the absence of movement, but the lack of more detailed information about the movement. Our assumption is that by keeping one location point and eliminating the other (n − 1 location points) the temporal information about the presence at that geographic position is not lost and they are not essential to be used as an indicator of the a user's stationary state. This assumption is true, only and only if, the time interval is small. We propose a threshold equal to one minute. In this way, the computation process can be faster when individual approaches are carried out. The location is represented by a point defined by its geographic coordinates at a given time.

The set of points removed during this step is given by Eq. 1:

$$\{p_k | (T(p_k) - T(p_{k-1})) \leq 1 \text{ min } \& \; D(p_k - p_{k-1}) = 0 \text{ m}\} \tag{1}$$

where $T(p_k)$ is the timestamp recorded for point p_k and $D(p_k - p_{k-1})$ is the distance between two consecutive points. The percentage of duplicate points (dp) removed is thus given by Eq. 2:

$$dp = \frac{card(p_k)}{n} * 100 \tag{2}$$

where n stands for the total number of points and $card(p_k)$ stands for the cardinality of the set of points removed after the duplicate point filtering step. For one day of data consisting of a total of 122,208,870 points, 24 % of them are eliminated after this filtering step.

3.2 Ping-Pong Points

Since a mobile is connected to the cell providing the best coverage, a change in the cell to which the mobile connects occurs in time. When the mobile is located at cell edge or at the border between two location areas, it might give rise to the "ping-pong handover": within a short period of time (less than 10 s), the mobile switches back to the old cell, fluctuating between the two neighboring antennas. This phenomenon can also occur when an equal intensity strength signal is received from two or more base stations, when the mobile is located in a coverage hole or in areas shadowed by high buildings.

The ping-pong phenomenon has been studied by the cellular network research community, which proposed several approaches to tackle such events (Gudmundson 1991; Pollini 1996). Handover algorithms can be decomposed in two steps, initiation and decision. The first one consists in deciding when to request a handover while the

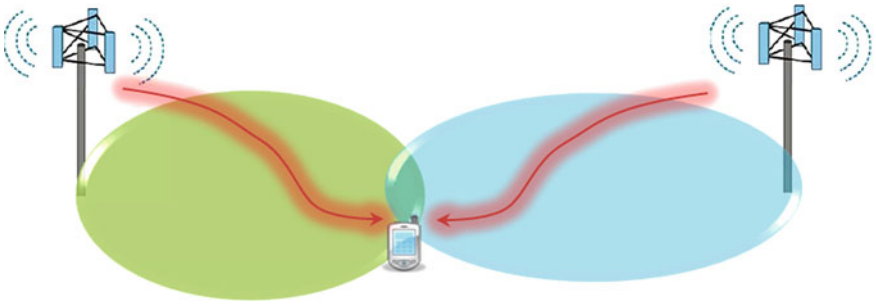

Fig. 3 The "ping-pong" handover phenomenon

latter one is based on signal strength comparison between the current and neighboring base station. Several handover types based on channel usage, microcellular and multilayered systems and network characteristics have been analyzed by state of the art researches [interested readers are invited to refer to (Ekiz et al. 2005)]. Most existing solutions focus on improving the decision in handover algorithms, which consists in comparing the differential signal power level between the serving and target base stations to a constant threshold value. This is mainly achieved by increasing this hysteresis threshold, designed to reduce the ping-pong effect during handover. Some methods propose handover algorithms performing at a sub-cell level, providing a more precise location of the mobile handset inside the cell (Feher et al. 2012). Other solutions take into account different types of location information to assist in the reduction of unnecessary handovers (He et al. 2010).

In human mobility studies based on cellular network traces, user location points issued from the ping-pong phenomenon are considered as noise. Since user trajectories are considered from a set of location points, noise should be filtered by a data pre-processing step, which operates on mobile handset communication logs.

Recently, in (Haoyi et al. 2012) such noise points were filtered out by mapping network cells to non-overlapping regions and identifying each region by the full set of cell towers covering the region. Then, user trajectories are considered by taking into account location points from the region having the longest hourly stay.

This produces anomalous points in a user trajectory as the users' handset is registered either with one LAU or the neighbor one, as illustrated in the simplified example in Fig. 3 which depicts a part of a network made of two base transceiver stations (BTS) each providing a coverage area (illustrated by oval shapes) for mobile stations. A mobile handset close to coverage borders of both BTS, is served by the left BTS (green cell) first, but it is then attached to the right BTS (blue cell), then switched back to the left BTS, and so on in a short time interval. The mobile device is stationary but in the data we observe erratic movements between the two positions.

In Fig. 4, the approach that allows detecting the ping-pong handover is illustrated. Points p_k (illustrated by circles) belonging to a user's trajectory (depicted by

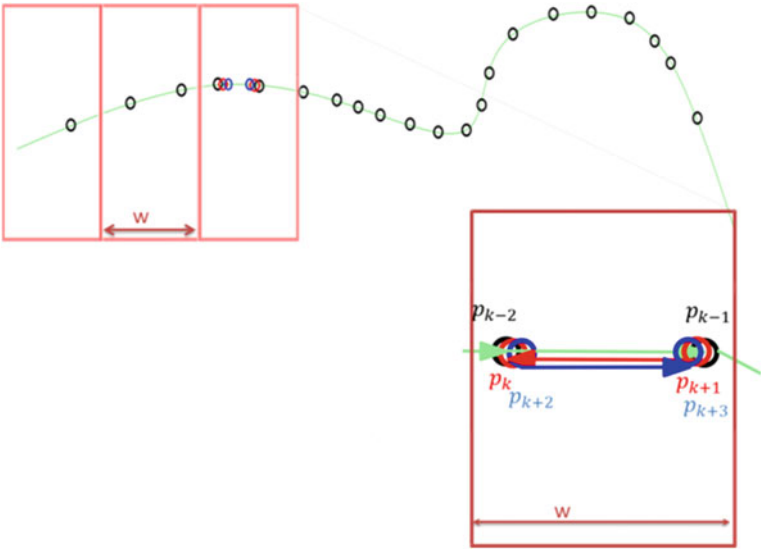

Fig. 4 The ping-pong handover detection

a continuous green line) are analyzed during a sliding spatial and temporal interval of width w (depicted by the rectangle of size w) in the upper left image and excerpt of successive points considered during the analysis window (in the lower right image). The ping-pong handover appears between points p_{k-2} and p_{k+3}, which are the points which should be considered in the user's trajectory as the first and the final points of the trajectory in w.

To identify the ping-pong handover (pp_{HO}) phenomenon, we define a sliding window of size w and analyze successive points of a trajectory belonging to the analysis window, denoted W in the following.

Between two consecutive points, p_{k-1}, and p_k, we compute the spatial distance $D(p_k - p_{k-1})$ and the temporal interval $T(p_k) - T(p_{k-1})$. Subsequently, the velocity of moving from a location point p_{k-1} to the following location point p_k is computed as follows:

$$v_{p_k} = \frac{D(p_k - p_{k-1})}{T(p_k) - T(p_{k-1})} \tag{3}$$

The heading direction (Zheng et al. 2008), $h_{p_{k-1}}$ is computed between the successive points, by considering North as the basis of the heading direction (cf. simplified illustration in Fig. 5). After computing pairwise heading direction between all points from a user's trajectory, we compute the heading change (Zheng et al. 2008) as:

$$hc_{p_k} = \left| h_{p_{k-1}} - h_{p_k} \right| \tag{4}$$

to identify location points having a heading change of 180°.

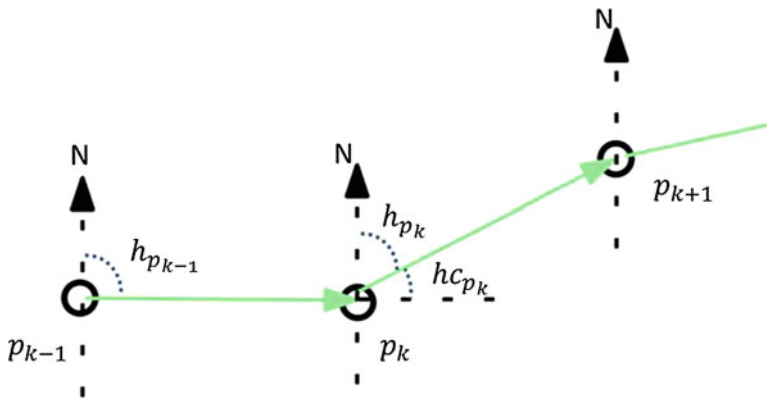

Fig. 5 Computing heading direction and heading change between three successive points from a user's trajectory

For all points inside the analysis window W, we pairwise compute velocity and heading change between successive points. The decision if a point is issued from a ping-pong phenomenon is taken based on the velocity between successive points and the heading change value. If the velocity is higher than a threshold, and the heading change is equal to 180°, the points are discarded as considered issued from a ping-pong handover phenomenon. The value of this threshold is empirically set at 200 km/h in our experiments.

The set of points issued from a ping-pong phenomenon is given by Eq. 5:

$$\{p_k | v_{p_k} > 200 \text{ km/h} \quad \& hc_{p_k} = 180^\circ\} \tag{5}$$

To assess the validity of the proposed approach for ping-pong point detection, we computed pairwise velocity, heading and heading change for a subset of 10 million points of the entire dataset of 122,208,870 points. Figure 6 below, shows the cumulative distribution function of points with a velocity between 0 and 500 km/h before (blue line) and after (red line) ping-pong filtering. It highlights the fact that over 90 % of the points belong to users moving with a speed lower than 100 km/h (which would correspond to vehicles or motorbikes usually used to move in urban areas). The effect of the ping-pong filtering is illustrated by the red line and shows that point velocity values belong after the filtering to lower bounds. Figure 7 illustrates point velocity plotted against the heading change value. The ping-pong phenomenon is shown in Fig. 7 through the straight peak at 180° which would correspond to the start of a handover phenomenon.

After the point filtering procedures a total of 40 % of points have been discarded: 24 % were due to duplicate points filtering and the remaining 16 % were points issued from ping-pong phenomenon.

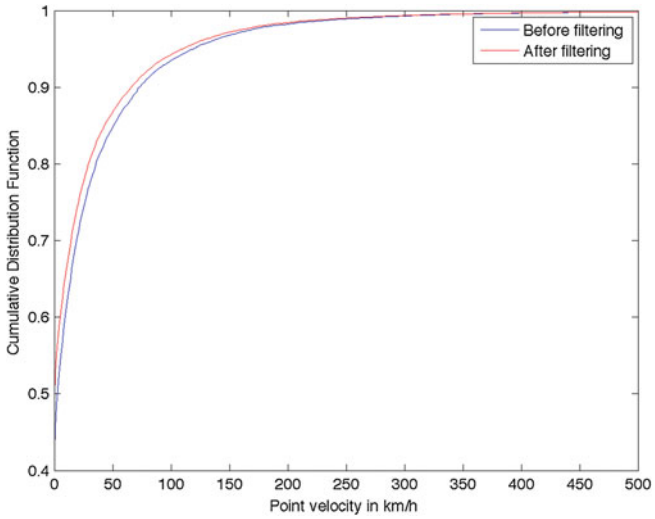

Fig. 6 Cumulative distribution function points velocity

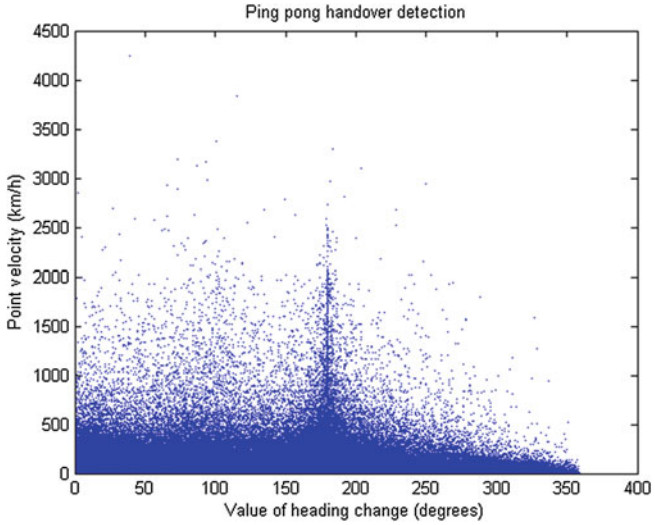

Fig. 7 Detecting ping-pong handover. A peak at 180° in the value of heading change marks the ping point phenomenon

4 User Sampling

The aim of this section is to introduce several quality measures for mobile phone data in order to sample users considered as reliable for a given application.

4.1 User Sampling Issue

Since, the set of observed locations for a user is dependent of his communication activity and/or the telecom network, the data describing the daily user activity are incomplete and heterogeneous. To overcome this drawback, some researchers (González et al. 2008) select randomly a sample of users from their dataset. Others try to optimize this method, by taking into account only users with a high number of recorded events (Song et al. 2010; Onnela et al. 2011). According to the desired application, the criteria could be to use only users having at least 0.5 calls per hour (Song et al. 2010), or users having records during each day for the study period. While this approach seems statistically sound (more user location points makes the analysis more precise), location points are generated whenever a communication or a LAU event is recorded. Thereby, user locations are biased, as they are depending of the users' calling frequency, mobility and the operator network. These evident shortcomings in estimating user mobility (Andrienko et al. 2012) have recently been raised in (Zhao et al. 2011; Couronné et al. 2011; Tiru and Ahas 2012).

In this section, we propose to estimate the correlation between the user's communication and itinerancy to better assess the bias of users sampling based on frequency activity. Using probe data, the relationship between communication and mobility patterns is studied. Analysis was conducted on two types of data: communication (voice calls, SMS) and itinerancy records. Figure 8 depicts communication frequency plotted against the median number of mobility records (LAU) for each user.

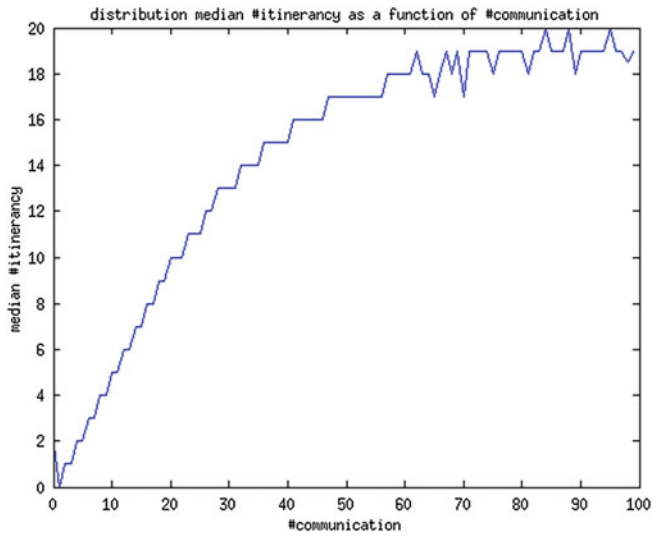

Fig. 8 Median number of local area change as a function of communication frequency

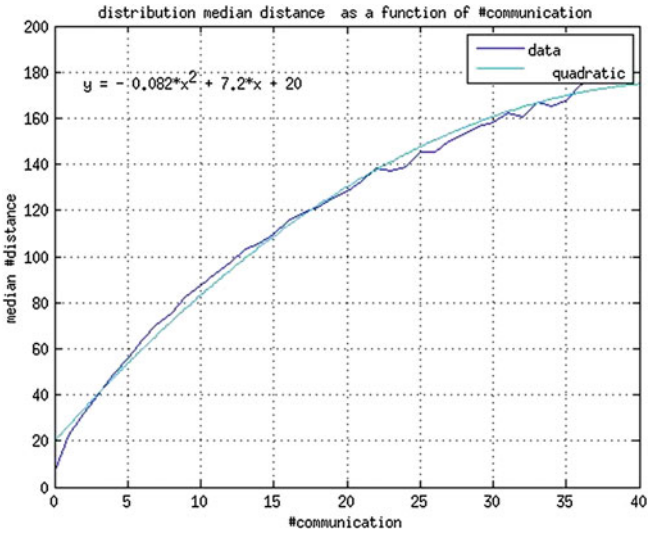

Fig. 9 Median daily distance traveled (km) as a function of communication events frequency

Ninety percent of users have less than 30 communication events (calls or SMS) during the observed day. For this group, we notice a clear, almost linear correlation between the frequency of communications and the median number of location area changes (daily mobility indicator).The curve reaches a plateau at about 50 communications per day and then the communication-itinerancy link disappears. People who communicate extremely frequently can no longer be distinguished by their median itinerancy.

To complete this study, the daily traveled distance (i.e. the Euclidian distance between locations defining the daily trajectory) per user was computed using all localized records (calls, SMS, HO, and LAU) and compared to communication events distribution (see Fig. 9).

We looked for a regression model to fit our data, where y is the median daily distance in km and x is the number of communication events (call, SMS). It appears that the best model is a quadratic function:

$$y = -0.082\,x^2 + 7.2\,x + 20 \tag{6}$$

This analysis confirms our observation showed in Fig. 9: the higher the number of communication events, the less the cumulative travelled distance increases.

A significant correlation between user mobility events and communication frequencies confirms our intuition that in mobile phone usages both phenomena are interrelated. A highly mobile person has in fact greater probability to use a mobile phone that someone who only commutes between a few places where s/he can also communicate *via* a landline telephone, VoIP, etc. In the same way, the higher mobility in the city context is frequently associated with a distant coordination via a mobile phone (Diminescu et al. 2009), and the mobile communication

is also linked to a management of the mobility itself: delays, traffic problems, last minute adjustments. Finally, correspondents of a highly mobile person learn with time which is the most adapted communication channel to reach this person, they will also contribute to reinforce the observed correlation.

From the point of view of human mobility analysis based on mobile phone data, the obtained results show that a careful examination of the sampling methods is of high importance. Selecting users with frequent communication traces, i.e., with many cell localizations, seems to introduce a clear bias because people having more mobile communications are also in a more mobile class of the general population.

In this context, the following questions arise: Which measures can be defined to obtain a statistically representative sampling of general human mobility patterns? How to choose the relevant measures for a specific application? Our approach is to use not only the frequency of records to sample representative users but also criteria that take into account data precision, accuracy or resolution. Moreover, data quality is a relative concept having different meanings to different consumers: data which is good enough for one user/application might not be of acceptable quality for another one. This is why, our approach consists in defining a set of different measures that qualify mobile phone data and then to use one or to combine different measures to sample users, depending upon the nature and objectives of the study.

4.2 Local and Global Measures

As stated in Sect. 3, individual trajectories are defined for each day and for each user. After performing location sampling, for each location-point composing a user's valid trajectory, local quality measures (Sect. 4.2.1) are first defined and further global ones (Sect. 4.2.2) are introduced. Finally, according to the application purposes these measures can be combined to estimate reliability of user's data.

4.2.1 Local Measures

In this section we define some measures that allow estimating position accuracy and precision for each user at local (location point) level. To do this, we propose to characterize pairwise consecutive points.

First, we define a speed index (named theta and noted θ) which describes user mobility between two sequential records, $n-1$ and n, considering laps-time $\Delta t = t(n) - t(n-1)$ and distance between the location of two sequential records far from each other from $d = \mathrm{Dist}(n-1,\ n)$. Note that measures defined in Eqs. 7 and 8 were proposed and tested in (Couronné et al. 2011), but they are explained here for better understanding.

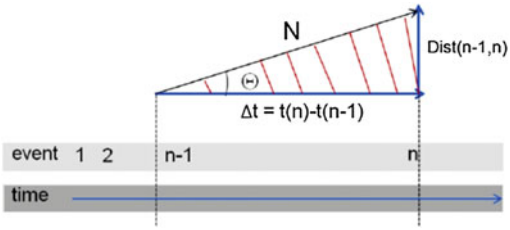

Fig. 10 Speed and uncertainty estimation

As we show in Fig. 10, θ is computed by the following equation:

$$\theta = \arctan \frac{d}{\Delta t} \tag{7}$$

The second index named "uncertainty" and noted U assesses whether θ estimation is confident or if this value has only a mathematical ground. Thus, the "uncertainty" reflects how confident the measured mobility state is; it is defined by the norm's vector having θ angle:

$$U = \frac{\Delta t}{\cos(\theta)} \tag{8}$$

The uncertainty estimation is related to the entropy concept. Thus, considering an entropy approach, we define a quality indicator, noted Q which measures the spatial accuracy. The more we are confident about our measurement, the less the entropy increases: the more probable it will be that we will find the user in a given location.

$$Q = e^{-\frac{\theta}{\frac{\pi}{2}} * U} \tag{9}$$

If uncertainty or theta increases then the quality Q of the measurement decreases: the more Q is close to one the better the confidence is. Q is defined as an exponential so that it spans between 0 and 1.

Q describes the probability distribution function to find a user in a spherical 3 dimensional space (2 geographical and 1 temporal), knowing two spatiotemporal measurements. The more the distance in space and or time is large, the larger the probability function will be widely distributed.

From the speed estimation and its confidence, we can derive the probable successions of position of the user during the time between two records; the quality of the user mobility increases with sampling frequency.

4.2.2 Global Measures

While local measures qualify data at point-level for each user, global measures addressed in this section are meant to qualify data at a user's trajectory level. A trajectory is hence made of successive points (locations) visited by a user during a

time interval. The aim of the measures proposed hereafter is to evaluate the confidence in a users' trajectory created from mobility/communication mobile phone data. These measures allow selection of users according to the applications.

(a) **Number of events**

As a user's trajectory is made of location points, one of the most intuitive ways to qualify a user's trajectory is to count the number of events generating the trajectory. A threshold can be applied in order to select users having the number of points greater than the threshold. As we discussed in Sect. 4.1, this selection can introduce some biases in final results when the threshold is important. We notice that this measure has meaning if the threshold is small, for example less than three. Indeed, users having less than three points (per day or during the time period of study) are not considered as relevant.

(b) **Temporal activity spread**

During a day, a user's communication or mobility habits gives more or less mobile network records (see Fig. 11). The aim of this second measure is to find the period of the day when the user is most active in terms of number of mobility/communication records. This measure is of highest importance as it gives a hint on the temporal resolution related to a user's mobility: a user having a uniformly distributed temporal spread is more reliable when analyzing its trajectory than one having few recordings (for which the loss of information is high).

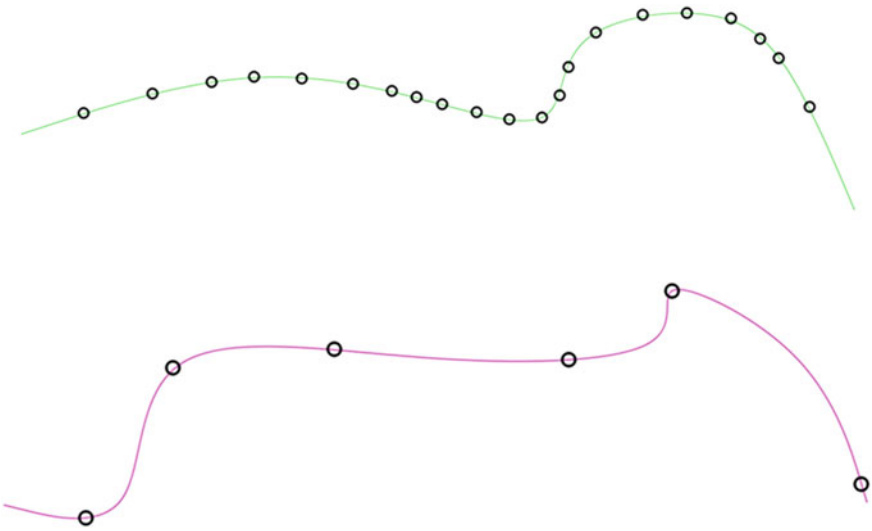

Fig. 11 Types of user trajectories. The *upper image* illustrates a reliable trajectory for mobility analysis as it is built upon a high number of points p_k (*illustrated by circles*) belonging to a user's trajectory. On the contrary, the trajectory depicted in the *lower image* is less reliable as it is built upon fewer points

We define the temporal activity spread of a user as the number of communication/mobility events recorded during the time span between the first and the last activity events generated by the user.

The number of recordings for a user is either given by communication or mobility events. To qualify a user's trajectory, it is crucial to establish if the recordings are given by the communication or itinerancy events. We introduce the event-rate (denoted R_e) as the number of events recorded for a user during the observation period. Let e be the number of events recorder for a user and Δt be the time interval during the first and the last recording hours considered for each user. Then the event-rate is given by the Eq. 10:

$$R_e = \frac{e}{\Delta t} \tag{10}$$

Similarly, we also define the voice activity rate (R_{va}) and the itinerancy rate (R_{vi}). The R_{va} is defined as the number of user-generated activity events (voice calls, SMS) during the time interval when the user is active while the R_{vi} is computed as the number of user-itinerancy events during the time span when the user is observed.

(c) **User based spatio temporal entropy measure**

Using the local measure defined in Sect. 4.1, we define a global one by computing for each user based on Shannon entropy and named "user based spatio temporal entropy" (UBSTE).

Let n be the number of records of the user u_i. The entropy $h(u_i)$ of the user u_i on a given number n of records is computed as follows:

$$h(u_i) = -\frac{1}{n}\sum_{j=1}^{n-1} Q_j * \log(Q_j) \tag{11}$$

where Q_j is the quality indicator, normalized w.r.t. $\sum_{j=1}^{n-1} Q_j = 1$.

The entropy describes the quantity of information we have about the state of the observed system. It increases when the quality of measurement decreases.

This indicator can be used as a filter to exclude sequences of records (user trajectories) having high entropy, so weak global spatiotemporal measurement quality. Traditionally, to reject users, the mean and the standard deviation of the entropy computed for the total number of users are used.

Thus, using this measure a user u_i is rejected if the next condition is true:

$$h(u_i) > \bar{h} + \sqrt{\frac{1}{N}\sum_{k=1}^{N} (h(u_k) - \bar{h})^2} \tag{12}$$

where $\bar{h} = \frac{1}{N}\sum_{k=1}^{N} h(u_k)$ represents the mean value of the entropy, and N represents the total number of users.

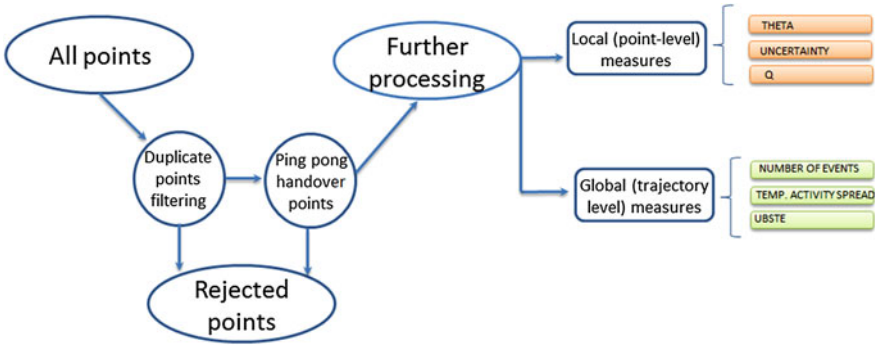

Fig. 12 Decision system for points filtering and user's selection

4.3 Decision Making

As the use of mobile network data reveals caveats for human mobility analysis, the measures proposed in this article can be handy when it comes to quantifying the bias they introduce. Such measures can be combined in cascade of decision systems taking into account the proposed measures depending on the nature and objectives of the studies considered, as illustrated in Fig. 12.

5 Conclusion

This chapter addresses location sampling and user sampling in mobile phone data. We propose two methods to handle location sampling: the first one consists of removing redundant location points (generally caused by rejected communications or SMS exchange in a very small time interval) and the second one allows the detection prior to discarding of erroneous locations points (caused by the ping-pong effect). Furthermore, we propose global and local quality measures meant to qualify mobile phone data for user sampling, a process which depends upon the nature and objectives of the study. Results of our study show a correlation between mobility span and communication frequency. This means that when working with CDR data, user selection has to be carefully performed, as random sampling is not efficient. Moreover, the mobility behavior seems to be associated with increased mobile communication, and billing data are generated only when a communication event is available. Thus, the data of more active users are of better quality (more points), but those users' mobility is also different. This obviously can cause serious problems in the mobile phone data based analysis of human mobility.

The measures proposed here should be seen as complementary measures to other existing state of the art measures. These measures can be integrated in a decision system aiming at defining the most exhaustive set of measures.

Future work in this area first includes testing our methods on different mobile phone datasets (more than one day, and covering different spatial areas such as dense urban areas, less dense urban areas, rural areas). To validate the proposed methods we would like to experimentally test these methods on a long period of time for different applications (O/D matrix, tourists behavior analysis, detection of the mean of transportation) and compare the obtained results to previous results (without using location and user sampling).

Second we wish to explore more ways to combine such measures in order to get the most reliable information out of the data. One of the improvements could be the design of a reliable decision system including the proposed measures.

Acknowledgments We would like to thank our colleague, Cezary Ziemlicki, who preprocessed data and has discussed with us many technical issues related to this chapter.

References

Andrienko G, Andrienko N, Bak P, Bremm S et al (2010) A framework for using self-organising maps to analyse spatio-temporal patterns, exemplified by analysis of mobile phone usage. J Locat Based Serv 4(3–4):200–221

Andrienko G, Andrienko N, Bak P, Keim D, Kisilevich S, Wrobel S (2011) A conceptual framework and taxonomy of techniques for analyzing movement. J Visual Lang Comput 22(3):213–232

Andrienko G, Andrienko N, Hurter C, Rinzivillo S, Wrobel S (2012) Scalable analysis of movement data for extracting and exploring significant places. In: Proceedings of IEEE transactions on visualization and computer graphics

Becker R, Cáceres R, Hanson H, Isaacman S, Loh JM et al (2013) Anonymous location data from cellular phone networks sheds light on how people move around on a large scale. Commun ACM 56(1):74–82

Caceres N, Wideberg J, Benitez F (2007) Deriving origin-destination data from a mobile phone network. Intel Trans Syst IET 1(1):15–26

Calabrese F, Di Lorenzo G, Liu L, Ratti C (2011a) Estimating origin-destination flows using opportunistically collected mobile phone location data from one million users in Boston metropolitan area. IEEE Pervasive Comput 10(4):36–44

Calabrese F, Smoreda Z, Blondel V, Ratti C (2011b) Interplay between telecommunications and face-to-face interactions—a study using mobile phone data. PLoS One 6(7):e208814

Couronné T, Smoreda Z, Olteanu AM (2011a) Chatty mobiles: individual mobility and communication patterns. NetMob, Boston

Couronné T, Olteanu AM, Smoreda Z (2011) Urban mobility: velocity and uncertainty in mobile phone data. In: Proceedings of Privacy, security, risk and trust (passat), 2011 IEEE third international conference on and 2011 IEEE third international conference on social computing (socialcom), pp 1425–30

Csáji BC, Browet A, Traag VA, Delvenne JC, Huens E et al (2012) Exploring the mobility of mobile phone users. Phys A 392(6):1459–1473

Diminescu D, Licoppe C, Smoreda Z, Ziemlicki C (2009) Tailing untethered mobile user: studying urban mobilities and communication practices. In: Ling R, Campbell SW (eds) The reconstruction of space and time. Mobile communication practices. Transaction Publishers, New Brunswick, NJ, pp 17–37

Ekiz N, Salih T, Kucukoner S, Fidanboylu K (2005) An overview of handoff techniques in cellular networks. Int J Inf Technol 2(3):132–136

Feher Z, Veres A, Heszberger Z (2012) Ping-pong reduction using sub cell movement detection. In: Proceedings of vehicular technology conference (VTC Spring)

González MC, Hidalgo CA, Barabási AL (2008) Understanding individual human mobility patterns. Nature 453:779–782

Gudmundson M (1991) Analysis of handover algorithms. In: Proceedings of 4th IEEE vehicular technology conference, gateway to the future technology in motion

Haoyi X, Zhang D, Zhang D, Gauthier V (2012) Predicting mobile phone user locations by exploiting collective behavioral patterns. In: Proceedings of the 9th IEEE conference on ubiquitous intelligence and computing (UIC'12), Fukuoka, Japan, 2012

Hasan S, Schneider CM, Ukkusuri SV, González MC (2012) Spatiotemporal patterns of urban human mobility. J Stat Phys 151:304–318

He D, Chi C, Chan S, Chen C, Bu J, Yin M (2010) A simple and robust vertical handoff algorithm for heterogeneous wireless mobile networks. Wireless Pers Commun 59(2):361–373

Kang C, Liu Y, Mei Y, Xu L (2012) Evaluating the representativeness of mobile positioning data for human mobility patterns. GIScience, Columbus

Olteanu Raimond AM, Trasarti R, Couronne T, Giannotti F, Nanni M et al.(2011) GSM data analysis for tourism application. In: Proceedings of 7th international symposium on spatial accuracy assessment in natural resources and environmental sciences

Olteanu Raimond AM, Couronne T, Fen-Chong J, Smoreda Z (2012) Le Paris des visiteurs, qu'en disent les téléphones mobiles? Inférence des pratiques spatiales et fréquentations des sites touristiques en Ile-de-France. Revue Internationale de la Géomantique 3:413–437

Onnela JP, Arbesman S, González MC, Barabási AL, Christakis NA (2011) Geographic constraints on social network groups. PLoS One 6(4):e16939

Phithakkitnukoon S, Horanont T, Di Lorenzo G, Shibasaki R, Ratti C (2010) Activity-aware map: identifying human daily activity pattern using mobile phone data. In: Proceedings of international conference on pattern recognition, Workshop on human behavior understanding, pp 14–25

Pollini GP (1996) Trends in handover design. IEEE Commun Mag 34(3):82–90

Ranjan G, Zang H, Zhang Z, Bolot J (2012) Are call detail records biased for sampling human mobility? ACM SIGMOBILE Mob Comput Commun Rev 16(3):33–44

Schulz D, Bothe S, Körner C (2012) Human mobility from GSM data-a valid alternative to GPS? Mobile data challenge 2012 workshop, June 18–19, Newcastle, UK

Sevtsuk A, Ratti C (2010) Does urban mobility have a daily routine? Learning from the aggregate data of mobile networks. J Urban Technol 17(1):41–60

Smoreda Z, Olteanu-Raimond AM, Couronné T (2013) Spatiotemporal data from mobile phones for personal mobility assessment. In: Zmud J, Lee-Gosselin M, Carrasco JA, Munizaga MA (eds) Transport survey methods: best practice for decision making. Emerald Group Publishing, London

Song C, Qu Z, Blumm N, Barabási AL (2010) Limits of predictability in human mobility. Science 327:1018–1021

Steenbruggen J, Borzacchiello MT, Nijkamp P, Scholten H (2011) Mobile phone data from GSM networks for traffic parameter and urban spatial pattern assessment: a review of applications and opportunities. GeoJournal 78:223–243

Tiru M, Ahas R (2012) Passive anonymous mobile positioning data for tourism statistics. In: Proceedings of 11th global forum on tourism statistics, Iceland

Vieira MR, Frias-Martinez E, Bakalov P, Frias-Martinez V, Tsotras VJ (2010) Querying spatio-temporal patterns in mobile phone-call databases. In: Proceedings of eleventh international conference on mobile data management (MDM), pp 239–248

Yuan Y, Raubal M, Liu Y (2011) Correlating mobile phone usage and travel behavior—a case study of Harbin, China. Comput Environ Urban Syst 36(2):118–130

Zhang Y, Qin X, Dong S, Ran B (2010) Daily O-D matrix estimation using cellular probe data. In: Proceedings of 89th annual meeting transportation research board

Zhao N, Huang W, Song G, Xie K (2011) Discrete trajectory prediction on mobile data. In: APWeb'11 Proceedings of the 13th Asia-Pacific web conference on web technologies and applications, pp 77–88

Zheng Y, Li Q, Chen Y, Xie X, Ma WY (2008) Understanding mobility based on GPS data. In: Proceedings of the 10th international conference on ubiquitous computing