Estimating Real Human Trajectories through Mobile Phone Data

Sahar Hoteit[†§], Stefano Secci[†], Stanislav Sobolevsky[§], Guy Pujolle[†], Carlo Ratti[§]
[†] LIP6/UPMC - University of Paris VI; 4 Place Jussieu, 75005 Paris, France - E-mail: Firstname.Familyname@lip6.fr
[§] MIT Senseable City Laboratory, 77 Massachusetts Avenue Cambridge, MA 02139 USA - E-mail : stanly@mit.edu, ratti@mit.edu

***Abstract*—Nowadays, the huge worldwide mobile-phone penetration is increasingly turning the mobile network into a gigantic ubiquitous sensing platform, enabling large-scale analysis and applications. In recent years, mobile data-based research reaches important conclusions about various aspects of human mobility patterns and trajectories. But how accurately do these conclusions reflect the reality? In order to evaluate the difference between the reality and the approximation methods, we study in this paper the error between real human trajectory and the one obtained through mobile phone data using different interpolation methods (linear, cubic, nearest, spline interpolations) and taking into consideration mobility parameters. From extensive evaluations based on real cellular network activity data of the Boston metropolitan area, we show that the linear interpolation offers the best estimation for sedentary people and the cubic one for commuters. Moreover, the nearest interpolation appears as the best one for "ordinary people" doing regular stops and standard displacements. Another important experimental finding described in this paper is that trajectory estimation methods show different error regimes whether used within or outside the "territory" of the user defined by the radius of gyration.**

***Index Terms*—Mobility pattern, interpolation methods, trajectory estimation, radius of gyration.**

I. Introduction

Human mobility and behavior pattern analysis has long been a prominent research topic for social scientists, urban planners, geographers and telecommunication researchers, but the pertinency of results has thus far been limited by the availability of quality data and suitable data mining techniques. Nowadays, the huge worldwide mobile-phone penetration is increasingly turning the mobile network into a gigantic ubiquitous sensing platform, enabling large-scale analysis and applications. In recent years, mobile data-based research reaches important conclusions about various aspects of human characteristics, such as human mobility and calling patterns [1] [2], virus spreading [3] [4], social networks [5] [6], content consumption cartography [7], urban and transport planning [8] and network design [9].

Nevertheless, in such user displacement sampling data, a high uncertainty is related to users movements, since available samples strongly depend on the user-network interaction frequency. For instance, we cannot determine the user positions between the calls with an acceptable accuracy. Some modeling techniques have been proposed in the literature to predict user movement between two places.

Authors in [11] propose a space-time prism approach, where the prism represents reachable positions as a space-time cube, given user's origin and destination points, time budget, and maximum speed. Spatial prisms allow for evaluation of binary statements, such as the possibility two moving users meet. However, the prism's maximum user speed limits the model applicability to a customer population using various transportation methods.

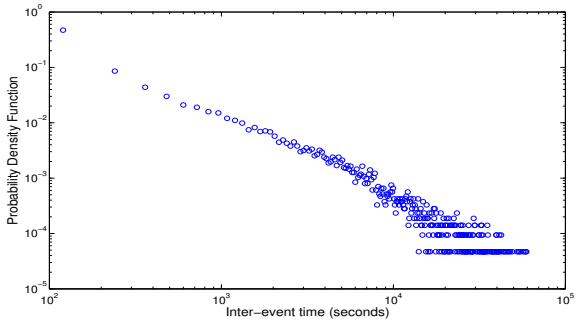


Fig. 1. PDF of the inter-call time empirical distribution

Authors in [12] propose a probabilistic extension of the space-time approach, applying a non-uniform probability distribution within the space-time prism. A strong assumption made therein is that users move linearly over time. This hypothesis is in a high contrast with the results obtained in [13] that show the tendency of users to stay in the vicinity of their call places. Authors in [13] propose a probabilistic inter-call mobility model that evaluates the density estimation of the spatio-temporal probability distribution of users position between calls, but it does not give an approximation of the fine-grained trajectory between calls. User displacements in many datasets have been analyzed in [14]; the authors find the displacement behavior show Levy walk properties (i.e., random walk with pause and flight lengths following truncated power laws). While very interesting in order to model inter-contact time distributions and general massive mobility, such random-based approaches cannot give precise approximations between given points on a per-user basis.

The objective of this paper is to assess the pertinence of different conceivable trajectory estimation approaches in terms of error from real available trajectories, via the analysis of real data from the Boston metropolitan area. By oversampling data-plan smart-phone user position samplings, and applying various interpolation methods, we assess the error between real human trajectories and estimated ones. We evaluate simple interpolation methods such as linear, nearest, cubic, and spline interpolations taking into consideration mobility parameters

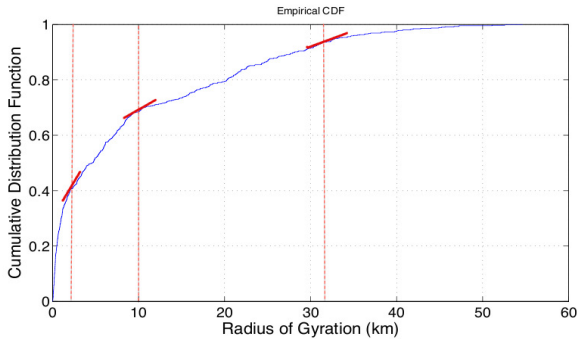


Fig. 2. Cumulative Distributive Function of the radius of gyration

the network operator may associate with each user. In particular, we highlight the dependance on the human mobility characteristic, with the user's radius of gyration as user mobility index. Our analysis proves that the linear interpolation shows the best performance for sedentary people (with a small radius of gyration) whereas the cubic one outperforms the others for commuters (having a big radius of gyration). On the other hand, the nearest interpolation presents the smallest error for a set of the population we identify as "ordinary people", who stop more often while moving during the day, whatever their radius of gyration is. Finally, we experimentally find that interpolations are more accurate when performed within the territory of the user, defined by the user's radius of gyration.

The paper is organized as follows. Section II presents the dataset used in our study and describes a user ranking with the radius of gyration as mobility pattern parameter. Section III presents the different interpolation methods evaluated in this paper. Section IV summarizes the results and findings. Finally, Section V draws some perspectives and discusses possible future work.

II. Dataset Description

We use a dataset consisting of anonymous cellular phone signaling data collected by AirSage [15], which converts the signaling data into anonymous locations over time for cellular devices. The dataset consists of location estimations - latitude and longitude - for about one million devices from July to October 2009 in the Boston metropolitan area. These data are generated each time the device connects to the cellular network including:

- when a call is placed or received;
- when a short message is sent or received;
- when the user connects to the Internet (e.g., to browse the web, or through email synch programs).

The location information is estimated through the AirSage's Wireless Signal Extraction (WiSE) technology [15], which aggregates, anonymizes and analyzes signaling data from cellular networks, and determines location information.

A. Trajectory Modeling

In order to qualify the precision of different interpolation methods, we have to determine the deviation of an estimated trajectory from the real one, being able to fix only a few real positions along the estimated trajectory.

We select anonymized signaling data of all users of a same operator during a single day (user identifiers change in different days in the available data). In order to determine real user trajectories, we fine-select data of those smartphones holders with a lot of samplings, typically those data-plan users with persistent Internet connectivity due to applications such as e-mail synch. By selecting users with more than 1000 connections (position samplings) during a given day, we can filter 707 smartphone users out of the whole dataset.

Then, in order to reproduce *artificial* "normal user" sampling, we subsample *real* data-plan smartphone quasi-continuous traces according to an experimental inter-event statistical distribution as given in Fig. 1. We determine it by analyzing *real* normal user samplings (for which the real quasi-continuous trajectory is unknown), available in the Airsage original dataset. Therefore, we extract, from the real trajectory, a first random position $P_i(longitude_i, latitude_i, time_i)$, then the corresponding next positions are extracted according to the inter-event time distribution values.

Hence, given a real trajectory with a high number of positions, and its oversampling that reproduces normal user's activity, we apply an interpolation method (see next section for the different interpolation methods) to estimate the trajectory across the given points. Given the real trajectory points $P_i(longitude_i, latitude_i, time_i)$, we estimate its corresponding position in time, in the estimated trajectory, P'_i $(longitude'_i, latitude'_i, time_i)$. Then we determine the deviation between the two points P_i and P'_i as the distance separating the exact position P_i to the estimated position P'_i in the interpolating curve joining the samples.

B. Mobility Ranking

People do not behave similarly, each person has different mobility habits in general and shows different mobility motifs during the particular day we consider in our study. Many studies have been conducted to find mobility patterns from network sampling, from very complex and complete ones able to determine precise motifs (e.g., [10]), to more aggregated and synthetic ones extracting a single parameter to characterize user mobility. A sufficiently precise, synthetic and easy to compute parameter is the radius of gyration, e.g., analyzed in [1], defined as the deviation of user positions from the corresponding centroid position. It is given by :

$r_g = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\vec{p_i} - \vec{p}_{centroid})^2}$; where $\vec{p_i}$ represents the i^{th} position recorded for the user and $\vec{p}_{centroid}$ is the center of mass of the user's recorded displacements obtained by $\vec{p}_{centroid} = \frac{1}{n}\sum_{i=1}^{n}\vec{p_i}$.

To explore the statistical properties of the population's mobility patterns, the cumulative distribution function (CDF) of the radius of gyration for the smartphone users is represented in Fig. 2. It is easy distinguish four main categories based on steep changes in the CDF slope.

- Users with $r_g \leq 3km$, who can be identified as the most sedentary people.
- Users with $3km \leq r_g \leq 10km$. They might be identified as urban mobile people as the diameter of the Boston urban area is very approximately around 10 km.

- Users with $10km \leq r_g \leq 32km$. They might be identified as peri-urban mobile people as the diameter of the Boston peri-urban area is very approximately around 32 km.
- Users with $r_g \geq 32km$, who can be identified as commuters spanning on the wide Boston metropolitan area.

III. Trajectory Interpolation Methods

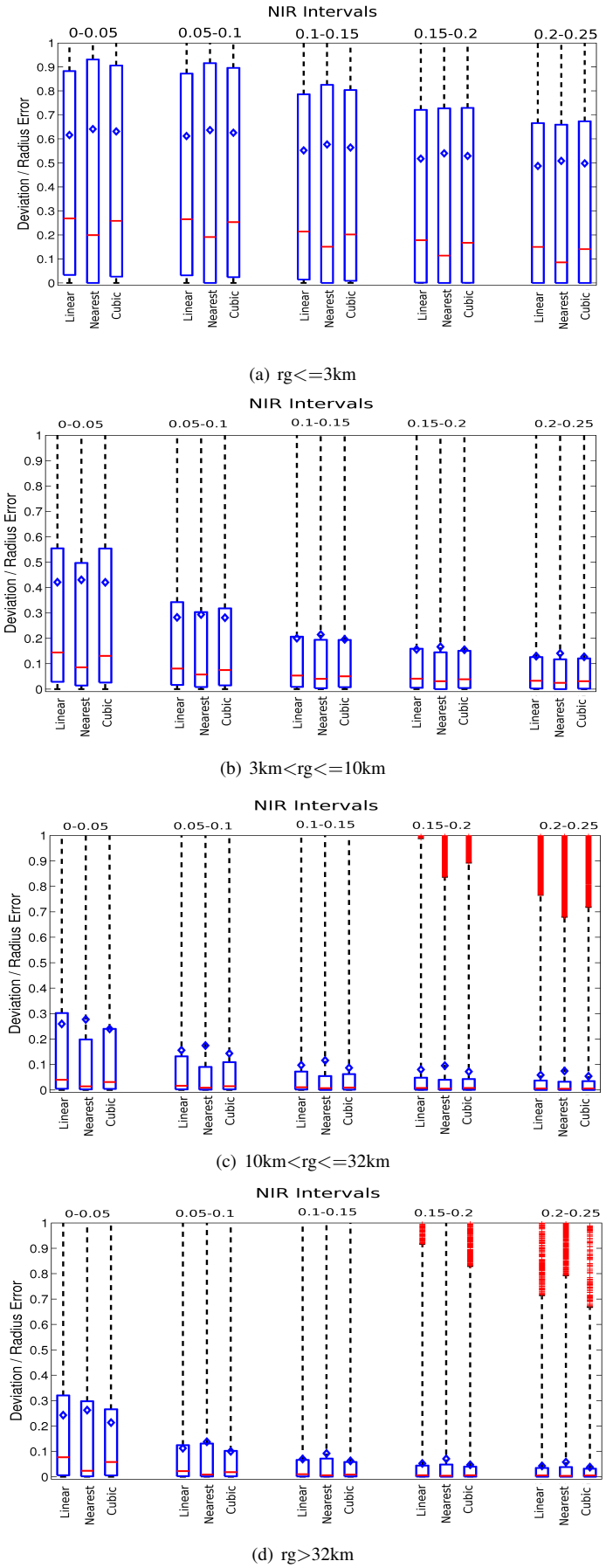
Different interpolation methods have been proposed in the literature to describe moving object trajectories. We present in the following a selection of classical ones, showing how they approximate the real trajectory (see an example in Fig. 3).

- the *Linear Interpolation*, is a popular interpolation used in movement objects databases [16]. It is obtained by joining straight interpolating lines between each pair of consecutive samples as shown in Fig. 3(b). Users are supposed to move at a constant speed along the straight lines.

- the *Nearest-neighbor Interpolation*, is an interpolation often used in mapping programs [17], also known as proximal interpolation. It consists of taking, for each position, the value of the nearest sampling position in time (not plotted because of the simplistic decision). Therefore, if we detect the same user in two different instants, at point A and point B respectively, the nearest interpolation attaches the user to position A for the first half period of time, and to position B for the second half.

- the *Piecewise Cubic Hermite Interpolation*, depicted in Fig. 3(c), is often used in image processing studies (see [18]). It is a third-degree spline that interpolates the function by a cubic polynomial using values of the function and its derivatives at the ends of each subinterval. This method interpolates the samples in such a way that the first derivative is continuous, but the second derivative is not necessary continuous. Suppose a subinterval $[x_1, x_2]$, with the function values: $y_1 = f(x_1)$, $y_2 = f(x_2)$ and the derivative values $d_1 = f'(x_1)$ and $d_2 = f'(x_2)$ are given. The cubic polynomial function in this subinterval is given by:
$C(x) = a + b(x-x_1) + c(x-x_1)^2 + d(x-x_1)^2(x-x_2)$
satisfying $C(x_1) = y_1$, $C(x_2) = y_2$, $C'(x_1) = d_1$ and $C'(x_2) = d_2$ This interpolation determines the coefficients a, b, c and d noting that:
$C'(x) = b + 2c(x-x_1) + d[(x-x_1)^2 + 2(x-x_1)(x-x_2)]$
is also continuous. The solution to this system is given by: $a = y_1$; $b = d_1$; $c = \frac{y_1' - d_1}{x_2 - x_1}$ and $d = \frac{d_1 + d_2 - 2y_1'}{(x_2 - x_1)^2}$, where $y_1' = \frac{y_2 - y_1}{x_2 - x_1}$.

- the *Spline Interpolation*, presented in Fig. 3(d), is a polynomial interpolation between samples, possessing a high degree of smoothness [19] [20]. It is often used in robot control and movement studies [21]. The function is constructed in exactly the same way as in Piecewise Cubic Hermite, but in this case the second derivative is also continuous.



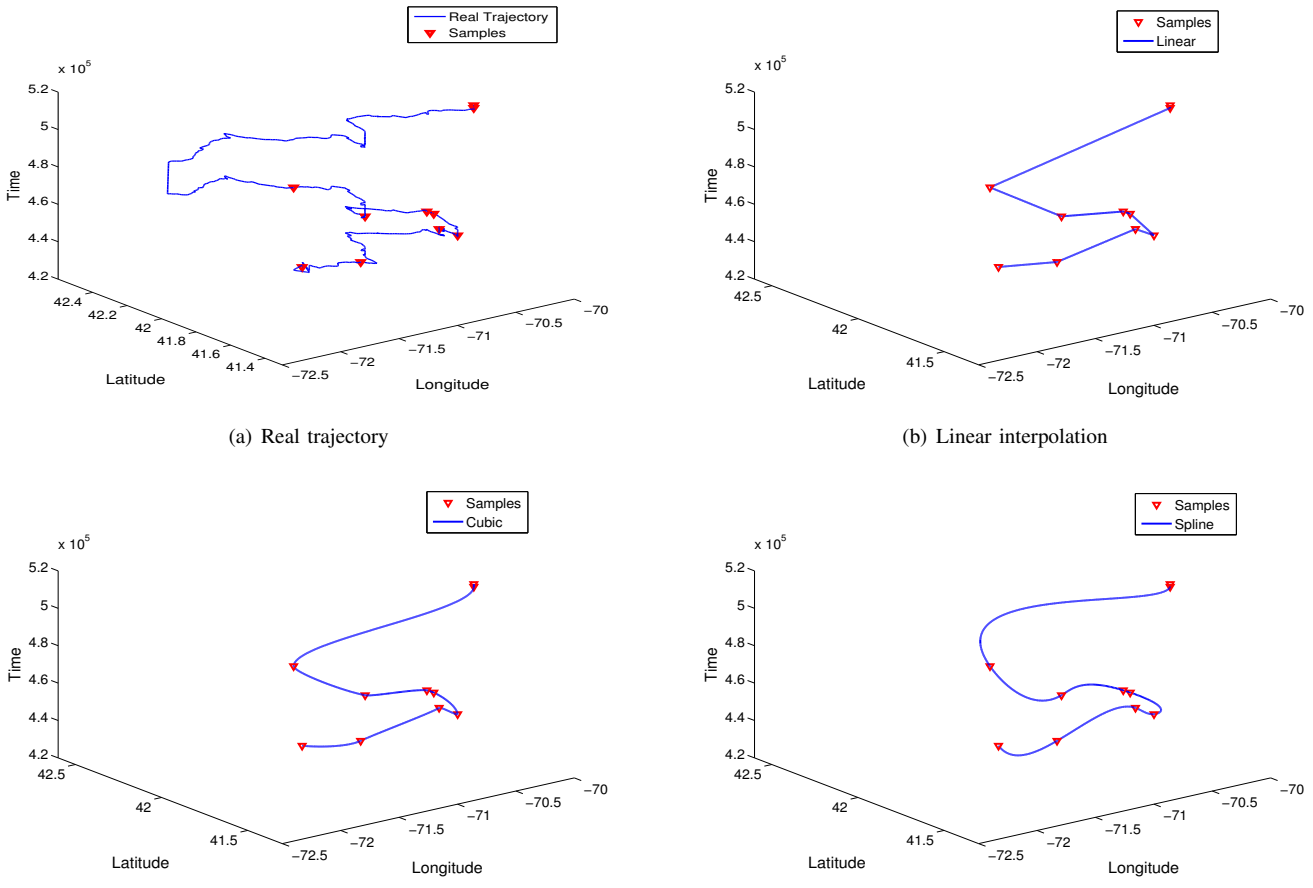
(a) rg<=3km

(b) 3km<rg<=10km

(c) 10km<rg<=32km

(d) rg>32km

Fig. 4. Boxplots of the deviation to the radius of gyration error for classical interpolation methods.



(a) Real trajectory

(b) Linear interpolation

(c) Cubic interpolation

(d) Spline interpolation

Fig. 3. Real and estimated trajectories

IV. Results

In this section, we present the main results obtained by applying the interpolation methods introduced in Section III. The spline interpolation is not included since its performance is by far worse than all other methods; the reason is that likely its large deviations from the straight lines do not approximate at all the majority of people paths.

First, we quantify the error, given by the ratio of the overall position deviation (computed as described in Section II-A) to the radius of gyration, for the different interpolation methods. Then, we further investigate the statistical distribution of the errors with respect to mobility parameters in order to understand what method performs better for each particular category of users.

A. Interpolation Error

Let us define the ratio between the number of over-sampled positions (normal user) to the total number of known positions (data-plan smartphone user), by the *"Network Interaction Ratio"* (NIR) parameter.

Fig. 4 reports boxplot[1] and average (the star) statistics about the interpolation error (trajectory deviation to the radius of gyration), for the linear, nearest and cubic interpolations. Boxplot statistics give a compact and rich enough view on the data to support the following analysis. At a first view, looking at the error averages, we can assess that:

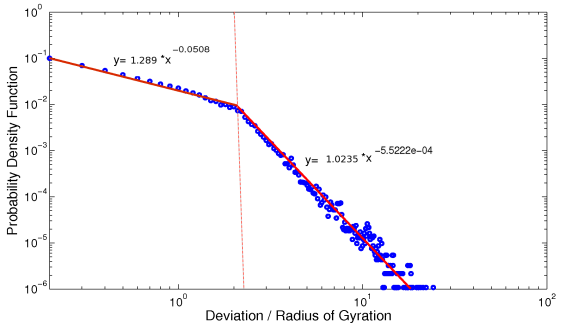
- The error is decreasing with the increase of the number of samples, for whatever interpolation, which is reasonable as one can get more accurate computations with more samples.
- The gap between the three methods decreases, especially for those with a radius of gyration higher than 10 km, i.e., for those who could be considered as peri-urban users and commuters (see Section II-B).
- The mean error of the linear interpolation is the smallest one comparing to other approaches, for users having a radius of gyration less than 3 km, i.e., sedentary users.
- The cubic interpolation presents the smallest mean error for higher radius of gyration, especially for commuters.
- For urban users, the linear and cubic interpolations show close performance.

Therefore, the trajectory deviation strongly depends on the mobility category, i.e., the user radius of gyration.

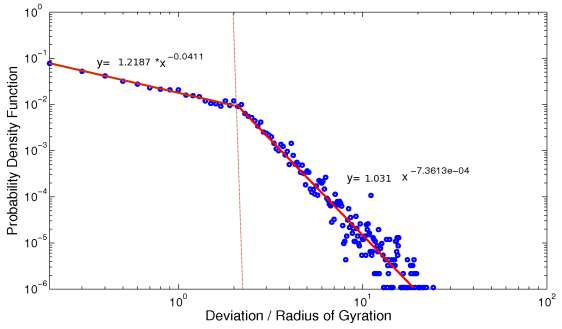
Finally, further looking into the whole statistics of the errors, including median and quartile lines, we can determine that:

- the median is always lower than the average, which indicates that the population contains an important part

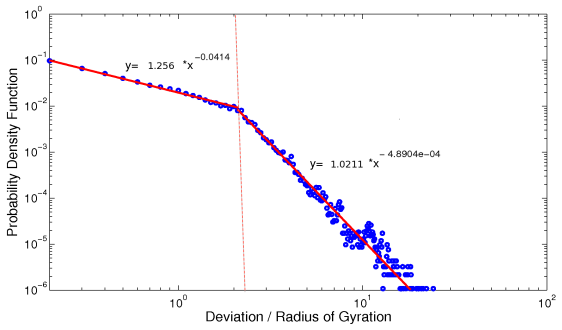
[1] i.e., first quartile, median, third quartile, maximum, minimum and outliers. It is worth noting that some maximum and outliers are cut in the figure for the sake of readability.



(a) linear



(b) nearest



(c) cubic

Fig. 5. Probability density function of error - (NIR: 0-0.05)

of users with much higher errors than the rest of the population.

- The medians and the averages show contradictory trends, when the average is higher for a method than for another, often the median is lower.
- The nearest interpolation shows better median statistics than all the other interpolations.
- The median error gets very low for NIRs higher than 0.1 for peri-urban and commuter users.

How to explain the huge gap between averages and medians, and the performance inversion indicating that nearest interpolation is on median the best interpolation, whatever the user category and the NIR are, is a matter of discussion. We interpret it with the fact that the median does not weight, as the average does, the error of those users for which a trajectory interpolation, whatever the type is, is not appropriate. That is, those "extraordinary" users that deviate too much from conventional paths. For example, users that have a backward path behavior (e.g., taxi drivers or similar users patrolling a zone, tourists coming back to already visited places, etc) can hardly be modeled by intuitive interpolations. The majority of "ordinary" users, typically moving forward, and regularly stopping at visited places, are instead captured by the median. For ordinary users, the nearest interpolation (introducing long stops at each sample and instantaneous displacement) is the best approximation, likely because ordinary users spend most of time without moving, even if they travel over long distances during the day.

The presence of a subset of the population which behaves very differently than the rest is confirmed by the fact that the average is often close and sometimes higher than the third quartiles (that indicate the upper bound for 75% of the population) in Fig. 4, and by the presence of many outliers especially for high NIRs. The ordinary users represent therefore more than 75% of the whole population, and the extraordinary ones have so high errors that the average is pushed close to the third quartile.

B. Interpolations' Probability Density Function

In order to further explore the statistical properties of the trajectory error, Fig. 5 shows the probability density function of the error for the linear, cubic and nearest interpolations.

It is easy to notice that there are two regimes. The distribution of errors over all users' positions is well approximated by a combination of two power law distributions joined by a breakpoint. It is surprising to notice that the breakpoint is the same (approximately equal to 2.2) for the different interpolation methods.

In practice, what does this power law breakpoint really mean? We interpret it as the point after which the interpolation error properties change abruptly, worsening. The value, around 2, corresponds to two times the user's radius of gyration, which in practice represents the user's "territory"(the circle of radius equal to the radius of gyration). This is a meaningful result: trajectory interpolations are more appropriate within the territory of a user than outside it.

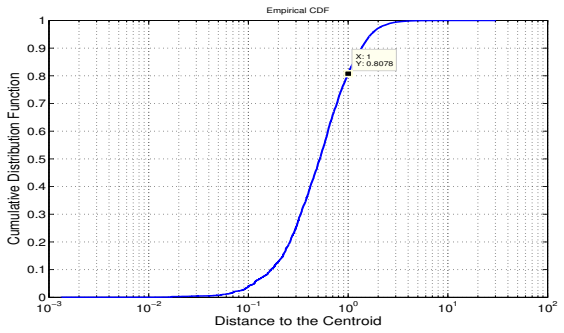
In order to further evaluate this dependency, we normalize the user position by the corresponding radius of gyration, and we plot in Fig. 6 the conditional cumulative density distribution of the two variables, error and the normalized distance to centroid. We can determine therein that:

- when small errors occur, we have a high probability (80.78%) that the user is inside the territory, and a low probability (19.22%) the user is outside it.
- When big errors occur, we have a probability of 40.25% that the user is inside its radius of gyration and a probability of 59.75% that the user is outside its radius.

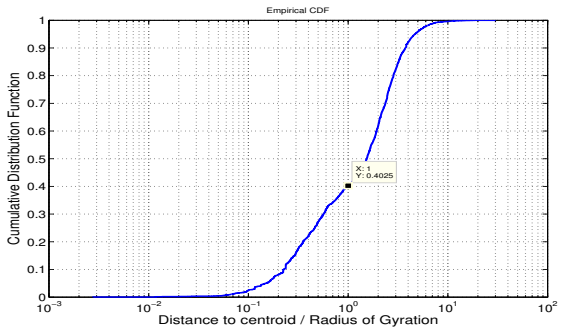
Therefore, we have an additional experimental proof that the trajectory error increases and its characteristics change when the user moves beyond the territory area roughly approximated by the radius of gyration.

V. Conclusion

Motivated by recent research on human mobility characterization based on cellular network log and probe data, we study in this paper the appropriateness of using such data in order to estimate the trajectory of people across metropolitan areas.



(a) Distributions of the normalized distance to the centroid when the deviation is less than 2.2 the radius of gyration



(b) Distributions of the normalized distance to the centroid when the deviation is more than 2.2 the radius of gyration

Fig. 6. Conditional cumulative density function

The applications are manyfold, ranging from content delivery network design to urban planning, yet our study is application independent and is of a fundamental nature.

Using Airsage data for millions of users from the Boston metropolitan area, we select data-plane smartphone users to get very precise localization data for a few hundreds of users. Then, we subsample these paths following the experimental normal user inter-event distribution, and apply to the sub-sampled position different interpolation methods. Finally, we finally analyze their errors to better understand the appropriateness of the different methods in detail, and of interpolation methods in general, for different mobility classes.

The major findings of our work can be summarized as follows.

- The radius of gyration is an appropriate, compact and easy to compute parameter to qualify user mobility in a metropolitan area network scope.
- The linear interpolation is the best approximation for sedentary users, linear and cubic interpolations work well for urban users, and the cubic interpolation is the best for peri-urban users and commuters.
- Separating ordinary users following conventional paths from the minority of users having unpredictable displacements, the nearest interpolation is by far the best approach whatever the mobility class is.
- Interpolation methods clearly work better when applied within the territory of the user defined by the radius of gyration.

As already mentioned, we believe the applications are manyfold. We are in particular interested in determining how content and Cloud delivery points in a urban and peri-urban environments can be identified and adapted online by inferring basic user mobility properties from big data log coming from cellular networks.

ACKNOWLEDGMENT

The authors would like to thank Airsage for providing the data used for the experiments, and the colleagues at MIT Senseable City Lab for their useful comments.

This work has been partially funded at the LIP6, UPMC, by the EIT KIC ICT Labs, Digital Cities of the Future chapter, Connected City project.

REFERENCES

[1] M. Gonzalez, CA . Hidalgo, Al. Barabasi "Understanding individual human mobility patterns", *Nature* 458, pp. 238-238, 2008.

[2] H. Hohwald, E. Frias-Martinez, and N. Oliver "User modeling for telecommunication applications: Experiences and practical implications", *in Proc. UMAP*, pp. 327-338, 2010.

[3] R. Huerta, L. Tsimring "Contact tracing and epidemics control in social networks", *Physical Review* E 66, 2002.

[4] P. Wang, MC. Gonzalez, CA . Hidalgo, Al. Barabasi "Understanding the spreading patterns of mobile phone viruses", *Science* 324, pp. 1071-1076, 2009.

[5] F. Calabrese, F. Pereira, G. Di Lorenzo, L. Liu, C. Ratti "The Geography of Taste: Analyzing Cell-Phone Mobility and Social Events", *In Proc. of 2010 IEEE Int. Conf. on Pervasive Computing (PerComp)*, 2010.

[6] M. Turner, S. Love, M. Howell, "Understanding emotions experienced when using a mobile phone in public: The social usability of mobile (cellular) telephones", *Telemat. Inf.* 25:3, pp. 201-215, 2008.

[7] S. Hoteit, S. Secci, Z. He, C. Ziemlicki, Z. Smoreda, C. Ratti, G. Pujolle "Content Consumption Cartography of the Paris Urban Region using Cellular Probe Data", *in Proc. of ACM URBANE 2012, CoNext 2012 Workshop*, 2012.

[8] M. R. Vieira, V. Frias-Martinez, N. Oliver and E. Frias-Martinez, "Characterizing dense urban areas from mobile phonecall data: Discovery and social dynamics", in *Proc. IEEE SocialCom*, pp. 241-248, 2010.

[9] H. Zang, J. Bolot, "Mining call and mobility data to improve paging efficiency in cellular networks", in *Proc. of 2007 ACM Int. Conf. on Mobile Computing and Networking (ACM MOBICOM 2007)*.

[10] C. Schneider, T. Couronne, Z. Smoreda, M. Gonzalez, "Are we in our travel decisions self-determined?", *Bulletin of the American Physical Society*, APS, 2012.

[11] T. Hagerstrand, "What about people in regional science?", *Papers in Regional Science* 24:1, pp. 6-21, December 1970.

[12] S. Winter and Z.C. Yin, "Directed movements in probabilistic time geography", *International Journal of Geographical Information Science* 24, pp. 1349-1365, 2010.

[13] M. Ficek and L. Kencl, "Inter-Call Mobility Model: A Spatio-temporal Refinement of Call Data Records Using a Gaussian Mixture Model", *In Proc. of IEEE INFOCOM*, 2012

[14] I.Rhee, M.Shin, S.Hong, K.Lee, S.J.Kim, S.Chong, "On the levy-walk nature of human mobility", in *Proc. of INFOCOM 2008*.

[15] Airsage: Airsage WISE technology, http://www.airsage.com.

[16] R. H. Guting and M. Schneider, *Moving Objects Databases*, Morgan Kaufmann, 2005.

[17] C. S. Yang, S. P. Kao, F. B. Lee and P . S . Hung, "Twelve different interpolation methods: A case study of Surfer 8.0", in *Proc. of XXth ISPRS*, 2004.

[18] F.N. Fritsch and R. E Carlson, "Monotone piecewise cubic interpolation", *SIAM Journal of Numerical Analysis* 17, 238-246, 1980.

[19] R.H. Bartels, J. C. Beatty and B.A. Barsky, "Hermite and Cubic Spline Interpolation", in *An Introduction to Splines for Use in Computer Graphics and Geometric Modelling*, pp. 9-17, Morgan Kaufmann, 1998.

[20] W.H. Press, B.P. Flannery, S.A. Teukolsky and W.T. Vetterling, "Cubic Spline Interpolation", *Numerical Recipes in FORTRAN: The Art of Scientific Computing*, pp. 107-110, 1992.

[21] S. Chand , B.J. Amis, W. Cartwright, A. Sadre and R. Kossey, "Robot control utilizing cubic spline interpolation", General Electric, US patent 4663726, 1987.