Estimating human trajectories and hotspots through mobile phone data

Sahar Hoteit, Stefano Secci, Stanislav Sobolevsky, Carlo Ratti, Guy Pujolle

Estimating Human Trajectories and Hotspots through Mobile Phone Data

Sahar Hoteit[a], Stefano Secci[a], Stanislav Sobolevsky[b], Carlo Ratti[b], Guy Pujolle[a]

[a]*Sorbonne Universities, UPMC Univ. Paris 06, UMR 7606, LIP6, F-75005, Paris, France (e-mails: sahar.hoteit@upmc.fr, stefano.secci@upmc.fr, guy.pujolle@upmc.fr)*
[b]*MIT Senseable City Laboratory, 77 Massachusetts Avenue Cambridge, MA 02139, USA (e-mails: 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. Recently, mobile data-based research reached important conclusions about various aspects of human mobility patterns. But how accurately do these conclusions reflect the reality? To evaluate the difference between reality and 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 interpolations) taking into consideration mobility parameters. Moreover, we evaluate the error between real and estimated load using the proposed interpolation methods. From extensive evaluations based on real cellular network activity data of the state of Massachusetts, we show that, with respect to human trajectories, the linear interpolation offers the best estimation for sedentary people while the cubic one for commuters. Another important experimental finding 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. Regarding the load estimation error, we show that by using linear and cubic interpolation methods, we can find the positions of the most crowded regions ("hotspots") with a median error lower than 7%.

Keywords: Mobility patterns, interpolation methods, trajectory estimation, radius of gyration, hotspot estimation.

1. Introduction

Human mobility and behavior pattern analysis has long been a prominent research topic for social scientists, urban planners, geographers, transportation and telecommunication researchers, but the pertinence 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] [3], virus spreading [4] [5], social networks [6] [7] [8], content consumption cartography [9], urban and transport planning [10] [11], network design [12].

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, Call Data Records alone do not provide a sufficiently fine granularity and accuracy, exhibiting a vast uncertainty about the periods when the user is not active, i.e., not communicating. This represents an issue for applications or analyses assuming ubiquitous and continuous user-tracking capability.

Some modeling techniques have been proposed in the literature to predict user movement between two places.

Authors in [13] and [14] infer the top-k routes traversing a given location sequence within a specified travel time from uncertain trajectories; they use check-in datasets from mobile social applications[1]. Their proposed methods permit to identify the most popular travel routes in a city, but they do not allow constructing time-sensitive routes.

Authors in [15] propose a space-time prism approach, where the prism represents reachable positions as a space-time cube, given user's origin and destination points – i.e., the assumption of knowing the location of a user at one time and then again at another time fits well mobile phone data in which we only know users' position during their communication events – as

[1]In recent years, mobile social applications have become so popular that they generate huge volume of social media data, such as check-in records or geo-tagged photos. In a check-in service, users note their locations via a mobile phone to share photos, activities etc.

well as time budget and maximum speed. Spatial prisms so allow evaluating of binary statements, such as the potential of encounter between two moving users. However, the maximum speed cannot be set for all users in general, which limits the model applicability.

Similarly, the authors in [16] 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 [17] that show the tendency of users to stay in the vicinity of their call places. Authors in [17] propose a probabilistic inter-call mobility model, using a finite Gaussian mixture model to determine users' position between their consecutive communication events (call or SMS) using Call Data Records. The model 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 using GPS traces have been analyzed in [18]; 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 state of Massachusetts. These estimated trajectories are then used to determine cells load in the considered region. By subsampling 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 method such as linear, nearest and cubic interpolations taking into consideration mobility parameters the network operator may associate with each user.

In particular, we highlight the dependence 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 population movements we identify as "ordinary moves", with long stops. In

addition, we experimentally find that interpolations are more accurate when performed within the territory of the user, defined by the user's radius of gyration. Finally we show that the usage of linear and cubic interpolations for modeling human trajectories allows us to determine the hotspot positions with a median error of less than 7%.

The paper is organized as follows. Section 2 presents the dataset used in our study and describes a user ranking with the radius of gyration as mobility pattern parameter. Section 3 presents the different interpolation methods evaluated in this paper. Section 4 summarizes the results of the comparison between the different methods. Section 5 evaluates the load estimation error. Finally, Section 6 draws some perspectives and discusses possible future work.

2. Dataset Description

We use a dataset consisting of anonymous cellular phone signaling data collected by AirSage [19], 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 Massachusetts state.

These data are generated each time the device connects to the cellular network including:

- when a call is placed or received (both at the beginning and end of a call);
- 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 estimations[2] not only consist of ids of the mobile phone towers that the mobile phones are connected to, but an estimation of their positions generated through triangulation by means of the AirSage's Wireless Signal Extraction technology [19] that aggregates and analyzes wireless

[2]Each location measurement is characterized by a position expressed in latitude and longitude and a timestamp.

signaling data[3] from mobile phones to securely and privately monitor the location and movement of populations in real-time, while guaranteeing acceptable user anonymity and privacy.

In this paper, we select anonymized signaling data of all users during a single day (the observation period is limited to one day because the anonymized user identifiers change for day to another to ensure user privacy).

2.1. 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 few real positions along the estimated trajectory.

To determine real user trajectories, we fine-select data of those smartphone 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 out 707 smartphone users from the whole dataset.

In order to reproduce "normalphone user" sampling, we subsample[4][5] real trajectories (i.e. smartphone user trajectories) according to an experimental inter-event statistical distribution as given in Fig. 1. We determine it by analyzing *real* normalphone user samplings (for which the real 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 subsampling 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 subsampled 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 $(longitude'_i, latitude'_i, time_i)$.

[3]The location measurements are generated based on signaling events, i.e., when a cellphone communicates with the cellular network's elements through control channel messages.

[4]The ratio between the number of the sampled positions to the total number of known positions (data-plan smartphone user) is defined by the subsampling ratio.
We evaluate in the paper different subsampling ratios.

[5]The subsampling process is independent and identically distributed.

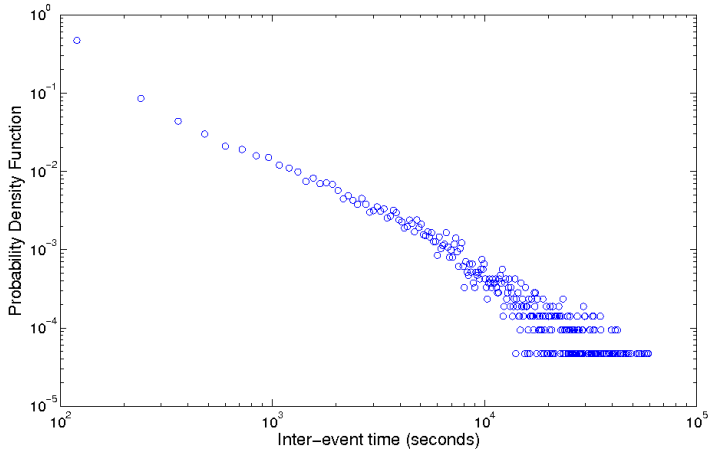


Figure 1: PDF of the inter-event time empirical distribution

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.

2.2. Mobility Ranking

People do not behave similarly, each person has different mobility habits in general and shows different mobility patterns 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., [20]), 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 [2]; it is defined as the deviation of user positions from the corresponding centroid position. More precisely, it is given by :

$$r_g = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\vec{p_i} - \vec{p}_{centroid})^2} \tag{1}$$

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 as:

$$\vec{p}_{centroid} = \frac{1}{n}\sum_{i=1}^{n}\vec{p_i}.$$

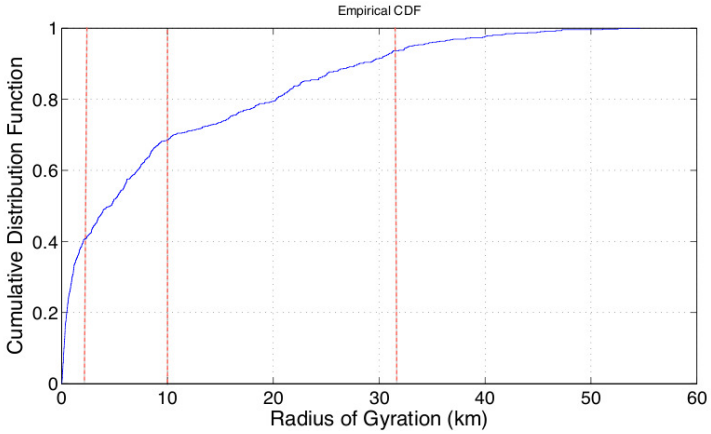


Figure 2: Cumulative Distributive Function of the radius of gyration

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[6] 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 spinning the whole Massachusetts state area.

[6]This categorization depends on city size, economic degree and other parameters. Comparing different sorts of human settlements on different levels of social and economical development, might be an interesting objective for the further studies but unfortunately, for now, we have access only to data covering Massachusetts' state in USA and not elsewhere.

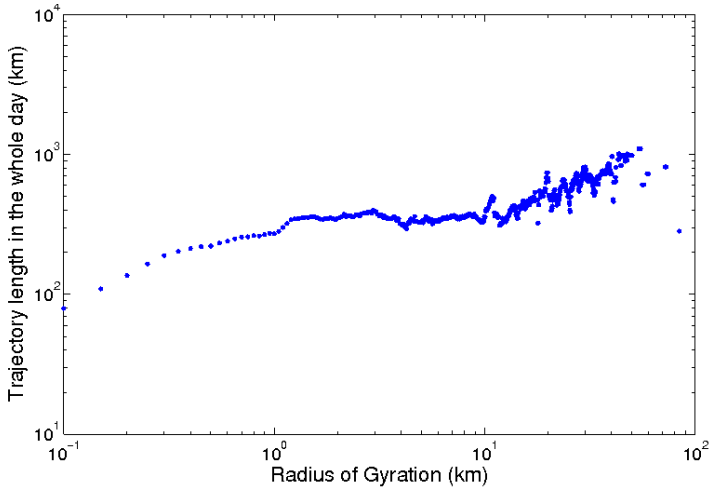


Figure 3: Total trajectory length with respect to the radius of gyration

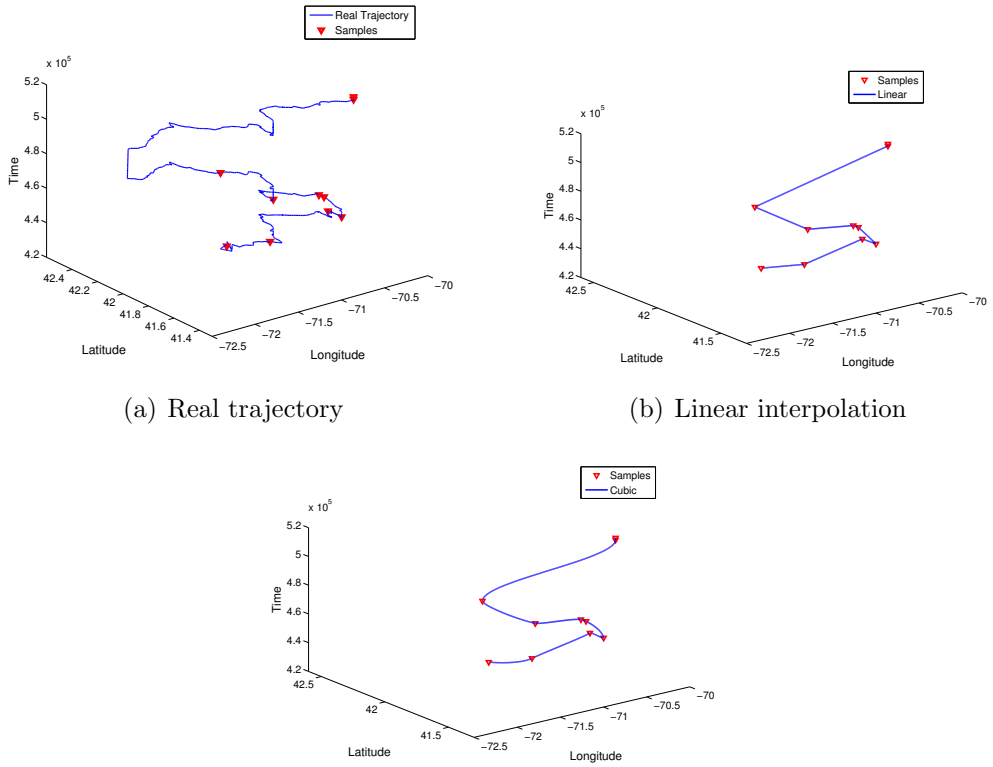
This ranking seems appropriate as the total traveled length increases with the radius of gyration[7], as displayed in Fig. 3. Moreover, this correlation may be interpreted by the fact that the radius of gyration can be viewed as a proper "territory"of each user, and thus increasing the territory area means that the person is able to move over longer distances.

3. 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. 4).

- the *Linear Interpolation*, is a popular interpolation used in movement objects databases [21]. It is presented in Fig. 4(b). It is obtained by joining straight interpolating lines between each pair of consecutive samples. Users are supposed to move at a constant speed along the straight lines. One limitation of the linear interpolation is that it can

[7]The absolute length is of course overestimated with respect to the real one. After looking into details, we discover that this is due to handover flipping among close antennas. The important aspect here remains the relative (and not the absolute) increasing trend.



(a) Real trajectory

(b) Linear interpolation

(c) Cubic interpolation

Figure 4: Real and estimated trajectories

fail in some situations where the interval of time between interpolated points is high. For example, suppose there are two points A and B in the road network with a curved path connecting them while with the linear interpolation we always assume the user drives on a straight line.

- the *Nearest-neighbor Interpolation*, is an interpolation often used in mapping programs [22], 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* is often used in image processing studies (see [23]). It is depicted in Fig. 4(c). 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. The slopes are chosen in a way that the function is "shape preserving" and respects monotonicity. 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:

$$C(x) = a + b(x - x_1) + c(x - x_1)^2 + d(x - x_1)^2(x - x_2) \qquad (2)$$

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:

$$C'(x) = b + 2c(x - x_1) + d[(x - x_1)^2 + 2(x - x_1)(x - x_2)] \qquad (3)$$

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 $y'_1 = \frac{y_2 - y_1}{x_2 - x_1}$.

4. Results

In this section, we present the main results obtained by applying the interpolation methods introduced in Section 3.

First, we quantify the error, given by the ratio of the overall position deviation (computed as described in Section 2.1) 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.

4.1. Interpolation Error

Fig. 5 reports boxplot[8] and average (the star) statistics about the interpolation error (trajectory deviation to the radius of gyration), for the linear,

[8] 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.

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 subsampling ratio, for whatever interpolation, which is reasonable as one can get more accurate computations with more samples.

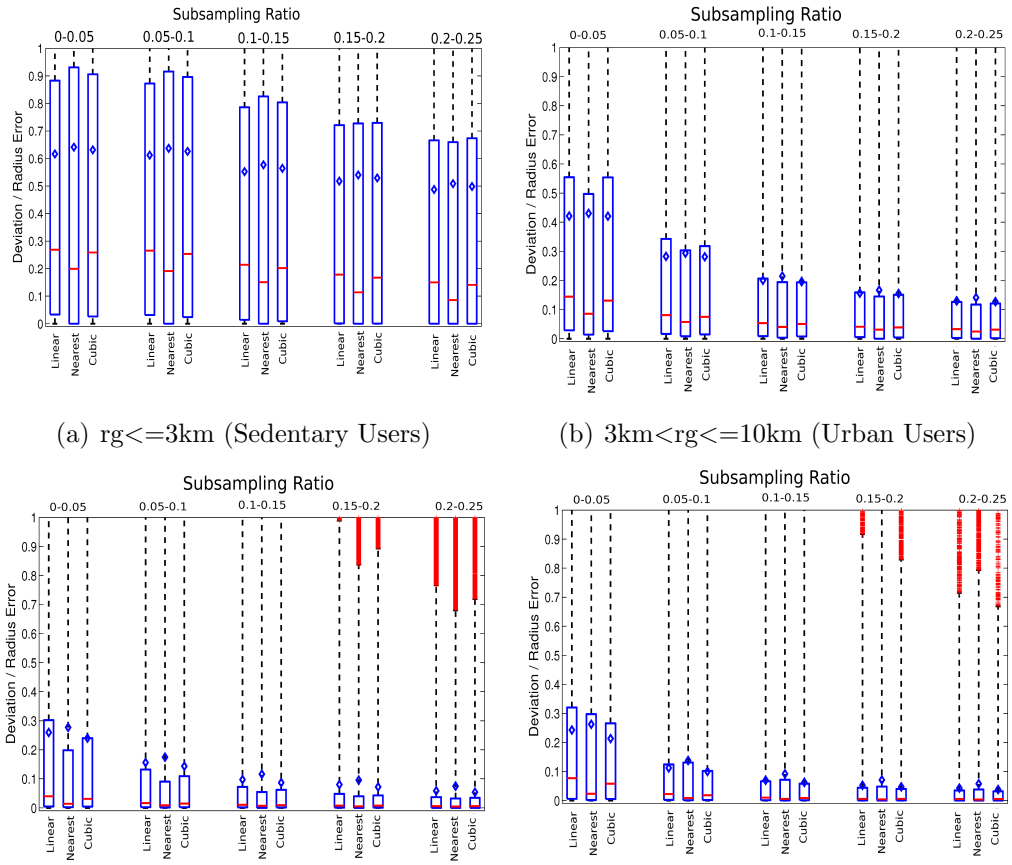
- The gap between the three interpolation methods decreases with the increase of the radius of gyration, especially for those users with a radius of gyration higher than 10 km, i.e., those who could be considered as peri-urban users and commuters (see Section 2.2).

- The lowest mean error among different interpolation methods depends on the category to which the user belongs. Indeed, for those users having a radius of gyration less than 3 km, i.e., sedentary users, the linear interpolation method presents the smallest mean error when compared to other methods. Instead, for those users having a higher radius of gyration, especially for commuters (i.e., those with a radius of gyration of more than 32 km), the cubic interpolation presents the smallest mean error. Finally, for urban users with a radius of gyration between 3 and 10 km, the linear and cubic interpolations show close performance.

Therefore, we can confirm that the trajectory deviation strongly depends on the mobility category, i.e., the user radius of gyration. In order to determine the correlation function between the deviation and the radius of gyration, Fig. 6 shows for each user (one point), given by its radius of gyration, the trajectory deviation (just for the linear interpolation, knowing that other interpolation methods give a very similar trend). The trend being generally increasing, we have positive correlation. Indeed, with the increase of the radius of gyration, users are able to move over longer distances, the distance between two samples increases, hence finding a good interpolation method that accurately approximates the real trajectory traversed by the user gets more challenging.

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



(a) rg<=3km (Sedentary Users)

(b) 3km<rg<=10km (Urban Users)

(c) 10km<rg<=32km (Peri-urban Users)

(d) rg>32km (Commuters)

Figure 5: Boxplots of the deviation to the radius of gyration error for classical interpolation methods.

- Overall, the nearest interpolation shows better median statistics than all the other interpolations for all user categories with different radius of gyration.
- The median error becomes very low for subsampling ratio of more than 0.1 for peri-urban and commuter users.

4.2. Interpolations' Probability Density Function

How to explain the huge gap between averages and medians, and the performance inversion indicating that nearest interpolation is on median the

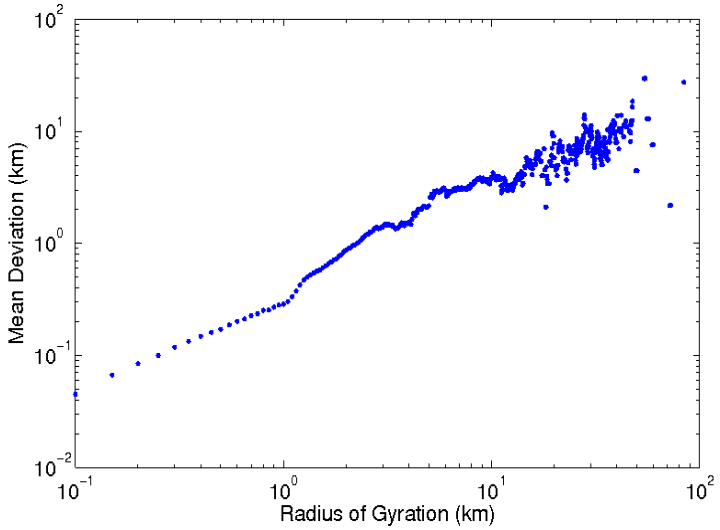


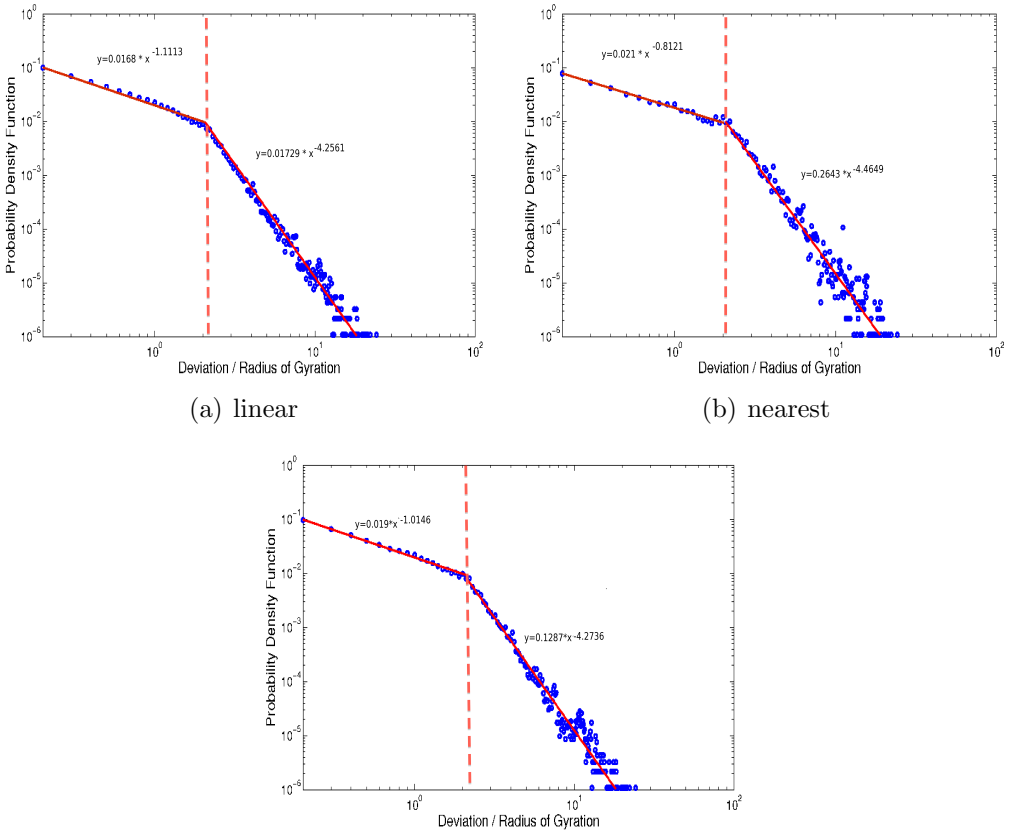
Figure 6: Mean deviation with respect to the Radius of Gyration

best interpolation, whatever the user category and the subsampling ratio 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' moves for which a trajectory interpolation, whatever the type is, is not appropriate; that is, those extraordinary moves that deviate too much from conventional paths. For example, the moves of users having a backward path behavior (e.g., tourist moves coming back to already visited places, etc) can hardly be modeled by intuitive interpolations. The majority of ordinary moves, with long stops at visited places, are instead captured by the median. For ordinary moves, the nearest interpolation (introducing long stops at each sample and instantaneous displacement) is the best approximation.

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 in Fig. 5, and by the presence of many outliers especially for high subsampling ratios. The ordinary moves represent therefore more than 75% of the whole moves, and the extraordinary ones (around 25% of the whole moves) have so high errors that the average is pushed close to the third quartile.

In order to further explore the statistical properties of the trajectory error, Fig. 7 shows the probability density function (PDF) of the error for



(a) linear

(b) nearest

(c) cubic

Figure 7: Probability density function of error - (subsampling ratio: 0-0.05)

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. 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

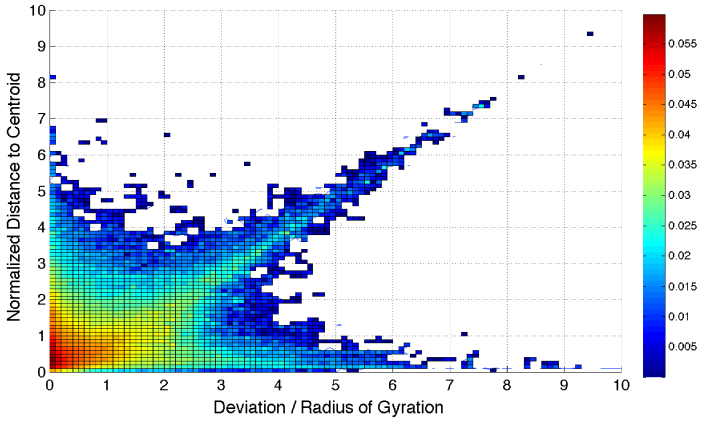


Figure 8: Joint Probability of the Deviation to the radius of gyration Error with the normalized distance to the centroid

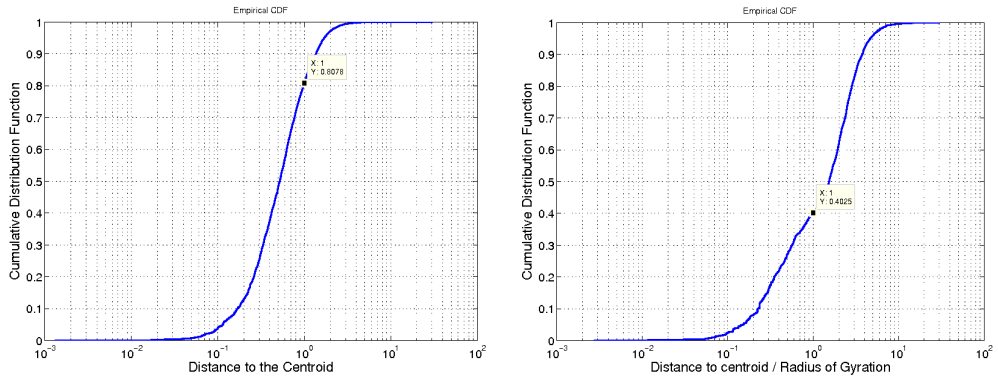
a user than outside it.

In order to further evaluate this dependency, we normalize the user position by the corresponding radius of gyration, and we plot in Fig. 8 the joint PDF of the normalized distance of users' positions to the centroid of the trajectory with the trajectory error. The figure shows that when the smallest errors occur, it is highly probable that the user is within the radius of gyration (when the normalized distance to the centroid is less than 1), i.e., the user's territory; when the highest errors occur, it is highly probable that the user is outside the territory.

These values can alternatively be analyzed by the conditional cumulative density distribution of the two variables, error and the normalized distance to centroid, as presented in Fig. 9. 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.



(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

Figure 9: Conditional cumulative density function

5. Estimation of Hotspot Positions

A fundamental issue to be taken into account for the management of broadband mobile cellular networks is finding the best location for the deployment of adaptive content and cloud distribution solutions at the base station and backhauling network level. Intuitively, an adaptive placement of content and computing resources in the most crowded regions can grant important traffic offloading, improve network efficiency and user quality of experience. We use thereafter the term "hotspots" to denote these regions. A limited amount of work exists in the literature for the estimation of hotspots and rendez-vous points in wireless access networks. E.g., in [24] vehicular data is exploited to determine accident-risk points. Many other works, such as [25], [26], [27] and [12], while assuming the availability of mobility information, focus in user-profile aware QoS provisioning, load balancing and network signaling improving techniques.

Traffic load forecasting has also been investigated from an analytical and mathematical modeling perspective. For example, authors in [28] show how under certain conditions periodic sinusoidal functions can be used as cellular traffic profile. Unfortunately the simplicity and the too theoretical approaches fail from precisely matching with the actual real traffic load, which is a strict requirement of our investigation.

Motivated by the usage of signaling mobile phone data that give real tra-

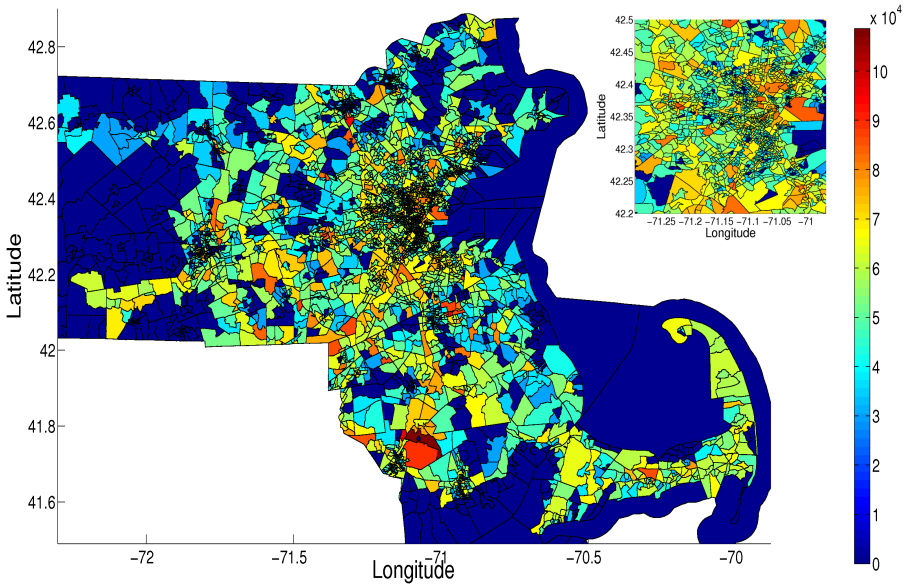


Figure 10: Real Block Load

jectories of smartphone users, we extract in this section real hotspot positions and compare them with the estimated positions that one can get by applying the interpolation methods defined above.

Decomposing the state of Massachusetts into census blocks[9][29] , we compute the real load of each block in the region (i.e., expressed as the users' number of visits to each block) as shown in Fig. 10.

The small map in the upper right corner is a zoom in of the Boston urban area, the state's largest city where small blocks exist. The figure clearly shows the load difference among the blocks and the existence of crowded blocks that define the most visited places where large masses of people usually visit.

Then, we estimate the load of each of these blocks by choosing for each user category the best interpolation method obtained in the results before (i.e. for sedentary and urban mobile users, we use the linear interpolation method to join the samples, while for peri-urban mobile users and the commuters we follow the cubic interpolation).

[9]A census block is the smallest geographic unit used by the United States Census Bureau. Blocks are typically bounded by streets, roads or creeks.

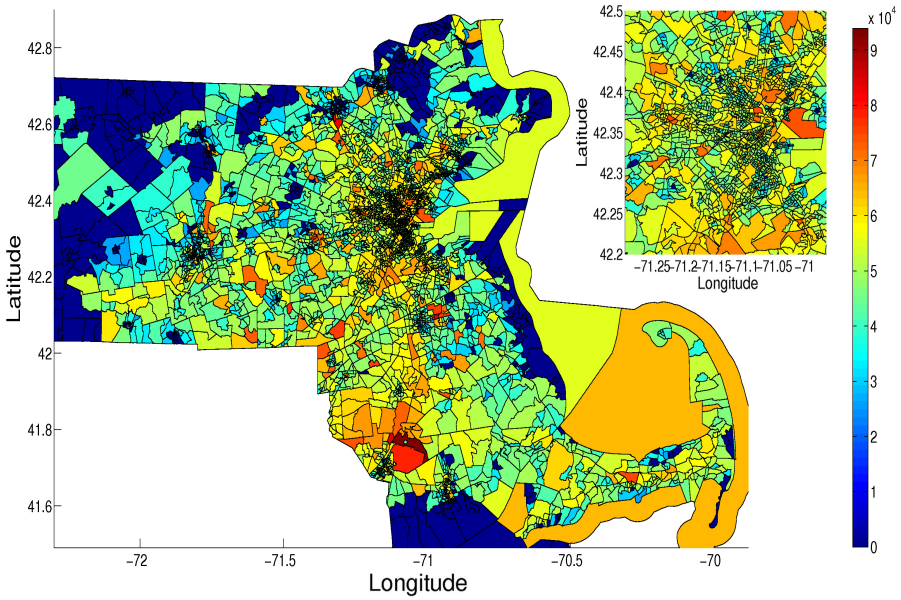


Figure 11: Estimated Block Load

After these, we compute the estimated block load. The results are obtained in Fig. 11, one can notice that the load is overestimated especially for the less crowded blocks. But what about the hotspots? How does the estimation error vary for the most crowded places?

Fig. 12 represents the variation of the estimation error with respect to the real load. In-line with ones exceptional for a statistically good estimation, we can state that:

- The estimation error is very high for the less visited blocks in the region.
- The estimation error rapidly decreases with the increase of the real load.
- For the most crowded blocks, we notice that the estimation error is significantly smaller.

By choosing different thresholds beyond which we identify the hotspot blocks (i.e., if a block has a load, expressed by the total number of users' visits during the day, that exceeds the chosen threshold it is considered as a hotspot block.), we plot for each case the cumulative distribution function

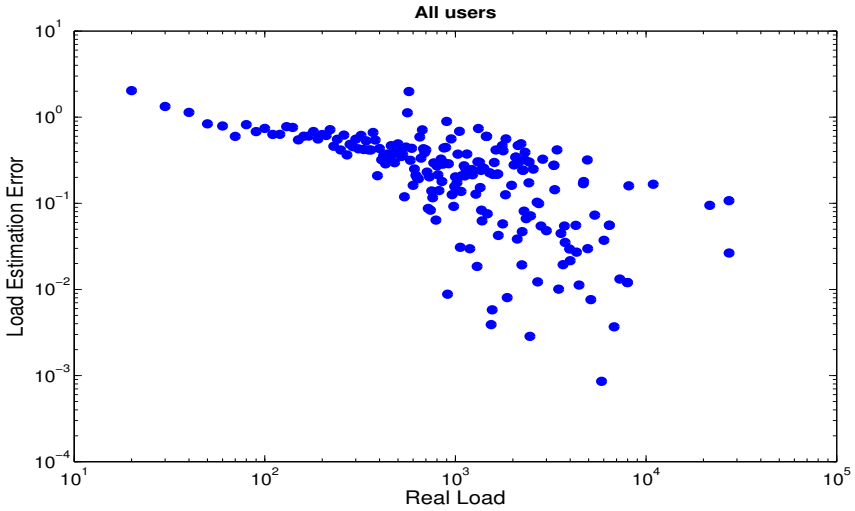


Figure 12: Load Estimation Error

of the block load estimation error. The results are shown in Fig. 13. We can state that:

- The median estimation error decreases with the increase of the real block load.
- The median estimation error reaches 7% for blocks of more than 2000 visits per day while for those with more than 100 visits per day, we get an error of 36%.

As a conclusion we can clearly confirm that the interpolation methods we have evaluated in this paper are able to find the hotspot positions with a small median error. We should note here that the proposed hotspot estimation method is scalable in a way that, taking a sample of users instead of the whole population enables us to find the hotspot positions in a relatively accurate way.

The online estimation of hotspot positions we propose is therefore very accurate and shows interesting properties in support of advanced urban computing services. A context of application could be that of content offloading [30] or Cloud offloading in mobile access networks: detecting hotspot positions in the backhauling network can allow adaptively allocating content caches or dimensioning CloudLet resources [31] for location-sensitive services.

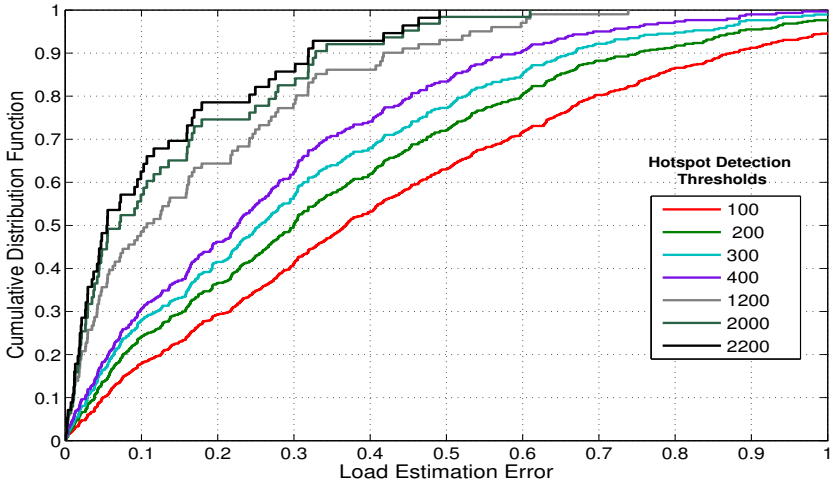


Figure 13: Cumulative Distribution Functions of the Load Estimation Error for different Hotspot Detection Thresholds

The availability of such an adaptive urban network sensing and related network management techniques can pave the way to advanced mobile application design, for example performing adaptive mobile Cloud computation offloading [32].

6. 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. 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 data for millions of users from the Massachusetts state, 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 subsampled position different interpolation methods. Finally, we 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 moves following conventional paths from the minority of user moves with 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.
- Interpolation methods are able to find the hotspot positions of the most crowded places with a very high precision.

As already mentioned, we believe the applications are manyfold. We are in particular interested in determining how content and Cloud delivery points in an 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.

We further thank Ericsson, the MIT SMART Program, the Center for Complex Engineering Systems (CCES) at KACST and MIT CCES program, the National Science Foundation, the MIT Portugal Program, the AT&T Foundation, Audi Volkswagen, BBVA, The Coca Cola Company, Expo 2015, Ferrovial, The Regional Municipality of Wood Buffalo and all the members of the MIT Senseable City Lab Consortium for supporting the research.

This work was partially supported by the ANR ABCD project (Grant No: ANR-13-INFR-005), and by the EU FP7 IRSES MobileCloud Project (Grant No. 612212).

[1] S. Hoteit, S . Secci, S. Sobolevsky, G. Pujolle and C. Ratti "Estimating Human Trajectories through Mobile Phone Data", *in Proc. of 2013 IEEE Int. Conference on Mobile Data Management (IEEE MDM 2013), Human Mobility Computing Workshop,* 3-6 June, 2013, Milan, Italy.

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

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

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

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

[6] 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.

[7] 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.

[8] R.C. Nickerson, H. Isaac, B. Mak "A multi-national study of attitudes about mobile phone use in social settings", *Int. J. Mob. Commun.* 6:5, 541-563, 2008.

[9] 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.

[10] 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.

[11] H. Wang, F. Calabrese, G. Di Lorenzo and C. Ratti, "Transportation mode inference from anonymized and aggregated mobile phone call detail records", *Proc. IEEE ITSC*, pp. 318-323, 2010.

[12] 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).*

[13] L. Wei, Y. Zheng, W. Peng "Constructing Popular Routes from Uncertain Trajectories", *18th SIGKDD conference on Knowledge Discovery and Data Mining*, KDD 2012.

[14] K. Zheng, Y. Zheng, X. Xie, and X. Zhou "Reducing Uncertainty of LowSampling-Rate trajectories", *In IEEE International Conference on Data Engineering, ICDE, 2012.*

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

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

[17] 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

[18] 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.*

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

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

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

[22] 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.

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

[24] TK. Anderson "Kernel density estimation and K-means clustering to profile road accident hotspots", Accident Analysis and Prevention, Vol. 41, No. 3, 2009.

[25] K. Seada "Rendezvous regions: a scalable architecture for service location and data-centric storage in large-scale wireless networks", in Proc. of 2004 Parallel and Distributed Processing Symposium.

[26] S.K. Das, S.K.S. Jayaram "A novel load balancing scheme for the teletraffic hot spot problem in cellular networks", Wireless Networks, Vol. 4, No. 4, 2004.

[27] D. Ghosal, B. Mukherjee "Exploiting user profiles to support differentiated services in next-generation wireless networks", IEEE Networks, Vol. 18, No. 5, 2004.

[28] E. Oh and B. Krishnamachari, "Energy Savings through Dynamic Base Station Switching in Cellular Wireless Access Networks", *In Proc. of IEEE Globecom 2010.*

[29] US census Bureau, http://www2.census.gov/

[30] V. Jacobson, D.K. Smetters, J.D. Thornton, M. Plass, N. Briggs and R. Braynard "Networking Named Content", *CoNEXT '09* .

[31] M. Satyanarayanan, P. Bahl, R. Caceres and N. Davies "The Case for VM-based Cloudlets in Mobile Computing", *IEEEPervasive Computing, 8(4), 2009..*

[32] L. Jiao, R. Friedman, X. Fu, S. Secci, Z. Smoreda and Hannes Tschofenig "Challenges and Opportunities for Cloud-based Computation Offloading for Mobile Devices", *in Proc. of Future Network and Mobile Summit, 2013.*