ITISE 2018

International Conference on Time Series and Forecasting

PROCEEDINGS OF PAPERS

Volumen 3

ITISE 2018
International Conference on Time Series and Forecasting

Proceedings of Papers
19-21 September 2018
Granada (Spain)

Editors and Chairs
Olga Valenzuela
Fernando Rojas
Héctor Pomares
Ignacio Rojas

I.S.B.N: 978-84-17293-57-4
Legal Deposit: Gr 1165-2018
Edit and Print: Godel Impresiones Digitales S.L.

A comparison of statistical methods for estimating individual location densities from smartphone data

Francesco Finazzi and Lucia Paci

Department of Management, Information and Production Engineering,
University of Bergamo, Italy
and
Department of Statistical Sciences,
Università Cattolica, Milan, Italy
francesco.finazzi@unibg.it
lucia.paci@unicatt.it

Abstract. In this work, the focus is on location data collected by smartphone applications. Specifically, we propose and compare a set of models of increasing complexity to estimate individual location at any time, uncertainty included. Unlike classic tracking for high spatio-temporal resolution data, the approaches are suitable when location data are sparse in time and are affected by non negligible errors. The approaches build upon mixtures of densities that describe past and future locations; the model parameters are estimated by maximum likelihood. The approaches are applied to smartphone location data collected by the Earthquake Network citizen science project.

Keywords: location-based applications; maximum likelihood; normal mixtures; spatio-temporal patterns

1 Introduction

The advent of GPS technology had a huge impact across a variety of fields. The main limit of GPS tracking is that the object to be tracked must carry a GPS receiver. When it comes to people, the limit is even more relevant as people are not supposed to carry around a GPS receiver. However, things have changed with the smartphone revolution. Nowadays, smartphones have a large number of built-in sensors and they are usually equipped with a GPS receiver. While this may suggest that tracking smartphones (and thus people) is an easy task, there are some issues to discuss.

Assuming that the smartphone user gave the permission to be tracked, the first problem is that the GPS receiver is likely off. Indeed, the receiver has an impact on battery consumption and it is usually on only when the smartphone is used for navigation purposes. Therefore, the smartphone location is usually given by the service provider's network infrastructure or wi-fi networks at lower accuracy than GPS. Additionally, independently of the way the location is obtained, acquiring and storing the smartphone location at high temporal frequency is considered a malpractice in smartphone application programming. Last but not least, the smartphone may stay off for long periods of times (from hours to days) and then no information on its location is collected. The above considerations suggests that, if the tracking does not have to negatively affect or alter the smartphone user experience, then the information on the smartphone location is available at low temporal resolution and with a lower precision if compared with GPS-based tracking.

This paper addresses the problem of estimating the smartphone/person location at a given point in time considering all the locations collected. Such problem is increasing relevant in many fields. For instance, if an area is affected by a natural disaster, it is useful to estimate the last location of missing people [1]. In epidemiological studies, pollution exposure can be dynamically assessed at population level if people locations are known at high temporal resolution [2]. Also, from a commercial perspective, individual recommendation or advertising can be provided to people whose location is known [3].

The analysis of individual location data has been explored by several authors to predict short-term trajectories. Customary, approaches used for human/animal tracking are based on interpolation methods, such as splines. [4] offered a comparison between interpolations methods at different temporal resolutions for animal tracking. Alternatively, [5] employed a Bayesian dynamic network for learning transportation routines between locations where the person spends a given amount of time and building personal map based on their behavior.

However, it is hard to recover the actual trajectory of a smartphone when the temporal resolution of the available locations is low and in the absence of additional information such as speed and acceleration in space. Secondly, when the smartphone user moves from a point $\mathbf{s}$ to another point $\mathbf{s}'$ in space, the followed path is rarely the shortest path in terms of Euclidean distance. This is because people are constrained by both natural topography (e.g., mountains) and man-made artifacts (streets, roads, etc.). For these reasons, tracking methods based on dynamic modeling are not suitable in this context.

A different approach to predict individual's location relies on the reproducibility of human patterns. Indeed, daily and weekly routines are well-established in human societies such that human activities are characterized by a certain degree of regularity and predictability [6, 7]. In this framework, [8] provided a spatio-temporal approach to predict arrival and residence times of users in their relevant places. [9] analyzed functional mobile data to identify subregions of the metropolitan area of Milan (Italy) sharing a similar pattern along time, and possibly related to activities taking place in specific locations and/or times within the city.

In principle, if we assume that people spend most of their time at few spatial locations (home, work, gym, etc.) than it is possible to group all the observed spatial locations into a small number of spatial clusters. As common in literature, clusters might be identified using finite mixture modeling. For instance, [10] employed a two-state mixture of Gaussian distributions centered at "home" and "work" locations to understand human motion from cell phone data and social networks. Also, [11] modeled human geolocation data from social networks by means of mixtures of kernel densities that allow to smooth individual's models towards an aggregate population model. However, the authors focused on user's spatial patterns ignoring the temporal dimension, i.e., assuming a time-invariant location density.

Mixture models can be extended to allow the model parameters to vary over time in order to represent complex dynamic distributions [12, 13]. For instance, we may want to give more weight to a given mixture component depending say on the time of the day or the day of the week. Then, a possible form for the mixture weights may depend on time-varying covariates or even dynamic processes. However, the resulting parameter space would be high dimensional and model estimation would become very challenging with existing algorithms. Moreover, the number of clusters should be estimated for each smartphone user, increasing the computational burden.

Our contribution is to propose flexible approaches that build upon mixtures of Gaussian distributions that describe past and future observed locations (with respect to each time). Time-varying mixing weights are introduced to estimate the location density at any time by exploiting temporal dependence as well as the fact that people follow reproducible patterns, such as daily and weekly patterns. Moreover, the precision of the mixture components depends upon the precision associated with each location; in other words, we also account for the positional error [14] arising in smartphone data. As a result, the number of model parameters is small and the number of clusters does not need to be estimated. A comparison among the proposed models shows the benefit of exploiting cyclical spatio-temporal patterns of people.

Our motivating data consist of smartphone locations collected by the Earthquake Network citizen science project (`www.earthquakenetwork.it`), which implements a world-wide early warning system based on smartphones. This is an instance of location data collected by a smartphone application (app) which makes use of geolocation but the primary role of which is not tracking. Clearly, the approach can be applied for modeling location data gathered by any location-aware app, including social networks.

The remainder of the paper is organized as follows. Section 2 introduces smartphone data and describes the inferential problem. Section 3 specifies the models proposed to describe the location density function, with estimation details provided in Section 3.4. An application to real data is illustrated in Section 4. We conclude with a brief discussion in Section 5.

2 Smartphone data and general set-up

Given a smartphone, a location-aware app is usually needed to acquire, store and send the smartphone location to a central server. Here, we assume that a smartphone app periodically acquires the smartphone location. If the Internet connection is available, the location is immediately sent to a central server otherwise it is stored and sent when the connection becomes available. The location provided by the smartphone at the generic time t is given as the probability density function of a bivariate Normal distribution centered on $\widetilde{\mathbf{s}}_t$ and with variance $\boldsymbol{\Sigma}_t$, i.e., $\varphi_t = \mathcal{N}_2(\mathbf{s}; \widetilde{\mathbf{s}}_t, \boldsymbol{\Sigma}_t)$, $\mathbf{s} \in \mathbb{R}^2$. In particular, $\widetilde{\mathbf{s}}_t$ is given by the easting and northing coordinates within Universal Transverse Mercator (UTM) zone, while $\boldsymbol{\Sigma}_t = \sigma_t^2 \mathbf{I}_2$ is a diagonal matrix, where $1/\sigma_t$ is the location precision and $\mathbf{I}_2$ is the 2×2 identity matrix. Here, we assume to have a collection of locations with associated precisions observed at irregularly times over a fixed period. In fact, the smartphone may stay off for long periods or the location may not be available when requested by the app.

In this work, we employ a classical measurement error model (MEM; [15]) by assuming that the location provided by the smartphone is a noisy version of the "true" unknown location $\mathbf{s}_t$, that is:

$$\widetilde{\mathbf{s}}_t = \mathbf{s}_t + \boldsymbol{\varepsilon}_t, \tag{1}$$

where $\boldsymbol{\varepsilon}_t \sim \mathcal{N}_2 (\mathbf{0}, \boldsymbol{\Sigma}_t)$. Therefore, the goal is to estimate the probability density function of the true location $\mathbf{s}_t$ over the projected geographic region $\mathcal{D} \in \mathbb{R}^2$ at the generic time t for any smartphone, that is

$$f_t \equiv f_t (\mathbf{s}; \mathcal{X}, \boldsymbol{\theta}), \tag{2}$$

where $\mathcal{X}$ denotes the information set and $\boldsymbol{\theta}$ is a vector of parameters. Given m observed locations for a generic smartphone, $\mathcal{X}$ contains all the observed locations with associated variances.

Using the density in (2), we can also provide a point estimate of the smartphone location at any time t. To accomplish that, different solutions can be considered such as the mean of f_t, the median, the mode and so on. Since, in general, f_t is multimodal, we estimate the smartphone location by the mode

$$\hat{\mathbf{s}}_t = \arg\max_{\mathbf{s} \in \mathcal{D}} f_t \left(\mathbf{s}; \mathcal{X}, \hat{\boldsymbol{\theta}} \right) \tag{3}$$

where $f_t(\mathbf{s}; \mathcal{X}, \hat{\boldsymbol{\theta}})$ denotes the density function of the smartphone location corresponding to parameter estimate $\hat{\boldsymbol{\theta}}$. By product, point estimator in (3) can be employed for online tracking when data are available at high temporal resolution. However, we stress that smartphone tracking is beyond our goal. Rather, we are interested in location density (2) and, potentially, its multiple modes. For instance, it might be useful to know all the locations where people are more likely to be if they are missing, say after a disaster.

3 Location density functions

In this section, parametric models of increasing complexity are introduced in order to describe f_t.

3.1 Tracking-like

Let $t' \leq t$ and $t'' \geq t$ be the nearest sampling times to a generic time t observed in the past and in the future, respectively. Hence, $\{\widetilde{\mathbf{s}}_{t'}, \boldsymbol{\Sigma}_{t'}\}$ and $\{\widetilde{\mathbf{s}}_{t''}, \boldsymbol{\Sigma}_{t''}\}$ denote the locations and associated variances at time t' and t'', respectively. The tracking-like model relies on the Gaussian assumption and assumes a Markovian structure such that the conditioning set $\mathcal{X}$ in (2) reduces to locations and variances of the nearest times from t. The resulting density is,

$$f_t = k_t \mathcal{N}_2\big(\mathbf{s};\ \boldsymbol{\mu}_t, \boldsymbol{\Lambda}_t\big)\ I\left(\mathbf{s} \in \mathcal{D}\right), \tag{4}$$

where k_t is the normalizing constant and

$$\boldsymbol{\mu}_t = w_t\, \widetilde{\mathbf{s}}_{t'} + (1 - w_t)\, \widetilde{\mathbf{s}}_{t''}, \tag{5}$$

$$\boldsymbol{\Lambda}_t = g(t;\alpha)\big[w_t \boldsymbol{\Sigma}_{t'} + (1 - w_t)\ \boldsymbol{\Sigma}_{t''}\big] \tag{6}$$

and

$$w_t = 1 - \frac{t - t'}{\Delta_t}, \tag{7}$$

$$g(t;\alpha) = \left(1 + \exp\left(\alpha\right) \Delta_t\right)^{w_t(1-w_t)}, \tag{8}$$

with $\Delta_t = t'' - t'$ the sampling interval, $I\left(\mathbf{s} \in \mathcal{D}\right)$ the indicator function equal to 1 if $\mathbf{s} \in \mathcal{D}$ and 0 otherwise. The density f_t has a maximum located over a straight line that connects $\widetilde{\mathbf{s}}_{t'}$ to $\widetilde{\mathbf{s}}_{t''}$ and depends on the temporal distance $t - t'$. The linear combinations in (5) - (6) are such that

$$f_{t=t'} = k_t\ \mathcal{N}_2\left(\mathbf{s};\ \widetilde{\mathbf{s}}_{t'}, \boldsymbol{\Sigma}_{t'}\right) I\left(\mathbf{s} \in \mathcal{D}\right), \tag{9}$$

$$f_{t=t''} = k_t\ \mathcal{N}_2\left(\mathbf{s};\ \widetilde{\mathbf{s}}_{t''}, \boldsymbol{\Sigma}_{t''}\right) I\left(\mathbf{s} \in \mathcal{D}\right), \tag{10}$$

namely, at the sampling times t' and t'', the density f_t is equal to the density φ_t sent by the smartphone and normalized over $\mathcal{D}$.

Assuming $\alpha > 0$, the term $g(t;\alpha)$ in (8) is equal to 1 when $t = t'$ and $t = t''$, while it is maximum when $t = \left(t' + t''\right)/2$. This model is suitable to describe a smartphone/person that moves from $\widetilde{\mathbf{s}}_{t'}$ to $\widetilde{\mathbf{s}}_{t''}$ on a straight path and at a constant speed, with precision on the starting and ending locations given by $1/\sigma_{t'}$ and $1/\sigma_{t''}$, respectively. The term $g(t;\alpha)$ modulates the location uncertainty along the path and it is high if Δ_t is high. Moreover, for $\Delta_t \longrightarrow \infty$, the density f_t converges to the density of the uniform distribution over the geographic region $\mathcal{D}$, since $f_t = 0$ outside $\mathcal{D}$. Finally, the higher α the more the straight path hypothesis is violated. In fact, a high value implies that f_t is more spread over $\mathcal{D}$ and thus the potential smartphone location is not restricted to lie on a straight path.

3.2 Bimodal density

The tracking-like model discussed above may be suitable when the sampling interval Δ_t is relatively small, say less than 5 minutes. In practice, if the consecutive sampling times are far apart (hours or days), the straight path and constant speed assumption would be unrealistic. Thus, when Δ_t tends to be large, we can assume that, for any $t' < t < t''$, there are no reasons to prefer any other location different from $\widetilde{\mathbf{s}}_{t'}$ or $\widetilde{\mathbf{s}}_{t''}$. Indeed, people tend to stay in the same location for long periods rather than constantly moving across space. Thus, assuming that the Markovian structure still holds, the following mixture model is proposed:

$$f_t = k_t \left[w_t \mathcal{N}_2 \left(\mathbf{s};\ \widetilde{\mathbf{s}}_{t'}, g(t;\alpha) \boldsymbol{\Sigma}_{t'} \right) + \big(1 - w_t\big) \mathcal{N}_2 \left(\mathbf{s};\ \widetilde{\mathbf{s}}_{t''}, g(t;\alpha) \boldsymbol{\Sigma}_{t''} \right) \right] I \left(\mathbf{s} \in \mathcal{D} \right), \tag{11}$$

where w_t and $g(t;\alpha)$ are defined in (7) and (8), respectively and k_t is the normalizing constant.

Contrary to (4) which is unimodal, for any $t' < t < t''$ the density f_t in (11) is bimodal with modes in $\widetilde{\mathbf{s}}_{t'}$ and $\widetilde{\mathbf{s}}_{t''}$. In this case, the parameter α describes how "fast" the locations $\widetilde{\mathbf{s}}_{t'}$ and $\widetilde{\mathbf{s}}_{t''}$ become unreliable when moving far in time from the sampling times t' and t'', respectively. Conditions (9) and (10) still hold for model (11).

3.3 Full-history mixture

In order to exploit all the available information on the smartphone location, we relax the Markovian structure and assume, at each time, a mixture of densities describing all past and future (with respect to t) locations. Therefore, the proposed density f_t is

$$f_t = k_t \left[\sum_{t' \in \mathcal{T}} v(t,t';\boldsymbol{\phi}) \mathcal{N}_2 \left(\mathbf{s};\ \widetilde{\mathbf{s}}_{t'}, \alpha \boldsymbol{\Sigma}_{t'} \right) \right] I \left(\mathbf{s} \in \mathcal{D} \right), \tag{12}$$

where, again, k_t is the normalizing constant, $\mathcal{T}$ is the set of observed times and $v(t,t';\boldsymbol{\phi})$ are the mixture weights defined as

$$v(t,t';\boldsymbol{\phi}) = \frac{l(t,t';\boldsymbol{\phi})}{\sum_{t' \in \mathcal{T}} l(t,t';\boldsymbol{\phi})}, \tag{13}$$

with

$$l(t,t';\boldsymbol{\phi}) = \exp\left(-\frac{|t-t'|}{\phi_1} \right) \exp\left(-\frac{|h(t,t')|}{\phi_2} \right) \exp\left(-\frac{1-d(t,t')}{\phi_3} \right). \tag{14}$$

The parameter α in (12) affects the variance of the components of the mixture such that the higher α, the more f_t is spread over the geographic region $\mathcal{D}$. When $\alpha \to \infty$ the density converges to the uniform density over the region.

The mixture weights (13) describe the daily and weekly cycle of smartphone users. Indeed, the function $h\,(t,t')$ returns the difference in time between t and t', independently of the calendar day, namely $|h\,(t,t')|$ is always less than 12 hours even when t and t' are more than 12 hours far apart. On the other hand, the function $d(t,t')$ is equal to 1 if t and t' are both working days or both weekends, otherwise it is equal to 0. In practice, $v(t,t';\boldsymbol{\phi})$ tends to be high when t and t' are close in time and/or when they are characterized by a similar time within the day and/or when they are days of the same type. The unknown parameters $\boldsymbol{\phi} > 0$ in the exponential terms of (14) describe the temporal persistence of the information carried by each mixture component. In particular, the persistence over time increases as ϕ_1 increases, while ϕ_2 modulates the intra-day persistence, i.e., the higher ϕ_2 the less the time within the day matters. Finally, the higher ϕ_3 the higher the weekend effect.

When $t \ll t_1$ or $t \gg t_m$, namely when f_t is used to predict the smartphone location far in time from the first or the last available observed location, the first exponential term in (14) approaches 0. However, the constraint on $v(t,t';\boldsymbol{\phi})$ implies that f_t still reflects the daily and the weekly cycles estimated from the information set. In this sense, the daily and weekly cycles characterize the smartphone location regardless of the position of t along the time line, and f_t is informative even if the smartphone has not sent its location for a long time.

3.4 Model estimation

For a smartphone i ($i = 1,\ldots,n$) at time $t_{i,j}$ ($j = 1,\ldots,m_i$), we observe its location $\widetilde{\mathbf{s}}_{i,j}$ with associated variance $\boldsymbol{\Sigma}_{i,j}$. Note that we suppressed the index t to simplify the notation. Let $\mathcal{X}_i = \left\{\widetilde{\mathbf{s}}_{i,1},\ldots,\widetilde{\mathbf{s}}_{i,m_i},\boldsymbol{\Sigma}_{i,1},\ldots,\boldsymbol{\Sigma}_{i,m_i}\right\}$ denote the information set of smartphone i; hence, $\mathcal{X} = \left\{\mathcal{X}_1,\ldots,\mathcal{X}_n\right\}$ is the full information set. Recall the MEM in (1); since $\boldsymbol{\varepsilon}_{i,j}$ is normally distributed and also $\mathbf{s}_{i,j}$ is assumed normally distributed or a mixture of normal distributions, then the likelihood function for smartphone i is given by the product of m_i normal densities or by the product of mixtures of normal density functions.

The information set $\mathcal{X}$ is exploited to estimate $\boldsymbol{\theta}_i$ for each smartphone i. Let $\mathcal{L}_h$ denote the log likelihood for model $h = 1,\ldots,3$; therefore, the maximum likelihood estimation is provided by

$$\hat{\boldsymbol{\theta}}_i = \arg\max_{\boldsymbol{\theta}_i} \log \mathcal{L}_h\left(\boldsymbol{\theta}_i;\mathcal{X}\right), \tag{15}$$

under the constraint that all the elements of $\boldsymbol{\theta}_i$ are positive.

4 Analysis of Earthquake Network data

Data used to illustrate models described above, come from the Earthquake Network project [16] which implements a world-wide early warning system based on smartphones. The earthquake detection is based on the signals sent by the smartphones every about 30 minutes to a central server, with information on the smartphone location and its precision. Additionally, the smartphone sends a signal when an acceleration above a threshold is detected. This implies that the sampling interval of the smartphone location is not regular, plus the smartphone does not send signals when it is switched off or Internet is not available.

Here, smartphones located in the metropolitan area of Rome, Italy, are considered. The dataset consists of 4.1 millions signals sent by 1336 smartphones to the server in the period January, 1st - April, 30th 2017 over the geographic box (41.75°N, 42.00°N, 12.35°E, 12.65°E). As an example, Figure 1 shows the locations observed for three different smartphones with associated standard deviations represented by the disks.

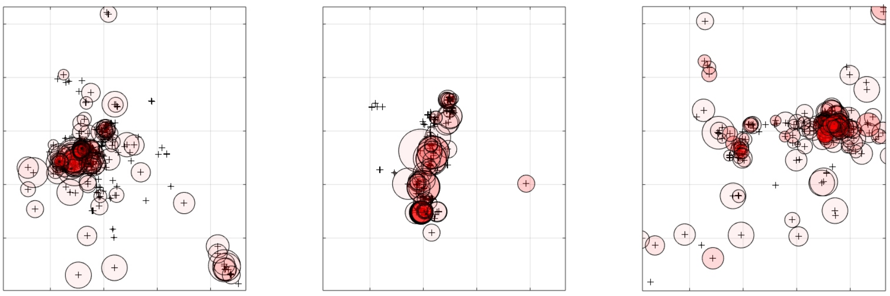

Fig. 1. Locations observed for three users; the center of the disk is the observed location while the radius of the disk is the associated standard deviation.

For each of the 1336 smartphones, the three models introduced in Section 3 are estimated adopting the maximum likelihood approach described in Section 3.4. Since the likelihood function may be characterized by multiple local maxima and the maximization in (15) is not guaranteed to converge to the global maximum, the maximization is carried out multiple times with different starting values for the parameter vector $\boldsymbol{\theta}_i$. The estimation $\hat{\boldsymbol{\theta}}_i$ with the highest likelihood is then retained.

Before presenting the estimation results, a flavor is given of the density f_t involved in the three models. Figure 2 shows $\log f_{t^*}\left(\mathbf{s};\mathcal{X},\hat{\boldsymbol{\theta}}_i\right)$ for smartphone $i = 1$ and t^* equal to February 5th, 2017 14:42:14 local time. Time t^* is a sampling time for the location of smartphone i but $f_{t^*}\left(\mathbf{s};\mathcal{X},\hat{\boldsymbol{\theta}}_i\right)$ is evaluated assuming that the smartphone location at t^* is unknown. In each panel of Figure 2, the diamond and the circular markers are the known locations of smartphone i at the nearest time from t^* in the past, say t^*_{-1}, and in the future, say t^*_{+1}, respectively. In particular, time t^*_{-1} is 13:09:10 while t^*_{+1} is 15:10:32. The dotted circles around each marker depict the smartphone location uncertainties, $\sigma_{t^*_{-1}} = 20.1\ m$ and $\sigma_{t^*_{+1}} = 900.0\ m$ (the circle around the diamond smartphone may not be visible since very small with respect to geographic area). The star marker is $\widetilde{\mathbf{s}}_{t^*}$, which is supposed to be unknown. Note that $f_{t^*}\left(\mathbf{s};\mathcal{X},\hat{\boldsymbol{\theta}}_i\right)$ has a different number of modes depending on the density model. In particular, only one mode for the tracking-like density, two modes for the bimodal model and multiple modes for the full-history. In all the cases, $f_{t^*}\left(\mathbf{s}; \mathcal{X}, \hat{\boldsymbol{\theta}}_i\right)$ is high at the locations $\mathbf{s}$ where the smartphone may be located at time t^*. The square marker is the maximum $\hat{\mathbf{s}}_{t^*}$ (see equation (3)). If we were to find the smartphone at time t^*, then $\hat{\mathbf{s}}_{t^*}$ would be the location we would search first.

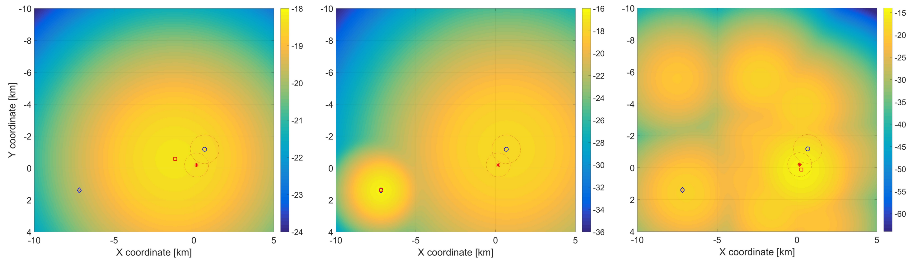

Fig. 2. Location densities under the three models at time t^* for a given smartphone: tracking-like density (left panel); bimodal density (mid panel) and full-history mixture (right panel). The star is the observed location at time t^*. Diamond and the circular markers are the locations at the nearest times from t^*; the dotted circles around each marker depicts their standard deviations.

The histograms in Figure 3 depict the estimated $\hat{\alpha}_i$ values for the 1336 smartphones when f_t is modeled using the tracking-like and the bimodal density. Note that an high $\hat{\alpha}_i$ value implies a large uncertainty on the smartphone location at time t. In this case, the average $\hat{\alpha}_i$ is equal to 48.4 and 32.9 for the tracking-like and the bimodal model, respectively. On the other hand, Figure 4 shows the parameter estimation results for the full-history density model. Note that $\hat{\alpha}_i$ is smaller since the model exploits all the past and future information on the smartphone behavior in space and time. Figure 5 shows the estimates of weights $v(t, t'; \boldsymbol{\phi})$ for the three smartphones the locations of which are plotted in Figure 1. These are the actual weights used to estimate the location density corresponding to February 14th, 2017 12:00:00 for the three smartphones.

The three density models are then compared using the AIC criterion. In Table 1, it is reported the percentage of times a model (row) is better than the other (column) when the comparison is based on AIC. For instance, the tracking-like density outperforms the other models only less than 1% of the times. The bimodal model is better than the tracking-like model around 99% of the times. The full-history model outperforms both the tracking-like and the bimodal model showing the benefit of accounting for the cyclical patterns.

5 Summary and concluding remarks

In this work, we addressed the problem of location density estimation over time using location data collected by smartphone apps. Specifically, smartphone loca-

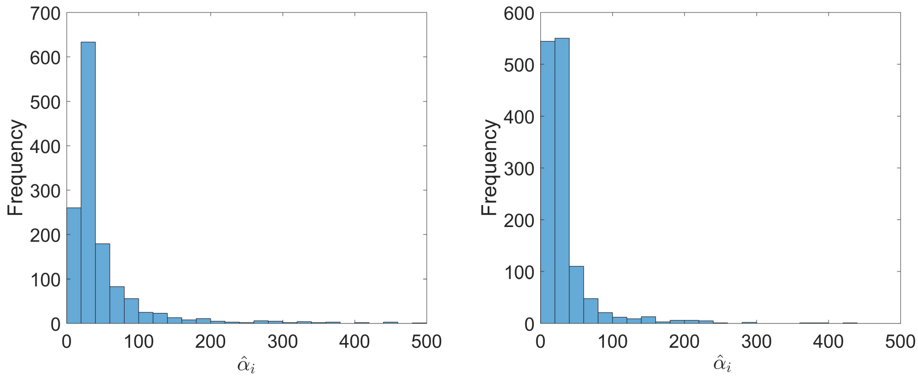

Fig. 3. Estimates of α_i for all smartphones when f_t is the tracking-like density (left panel) and the bimodal density (right panel), respectively.

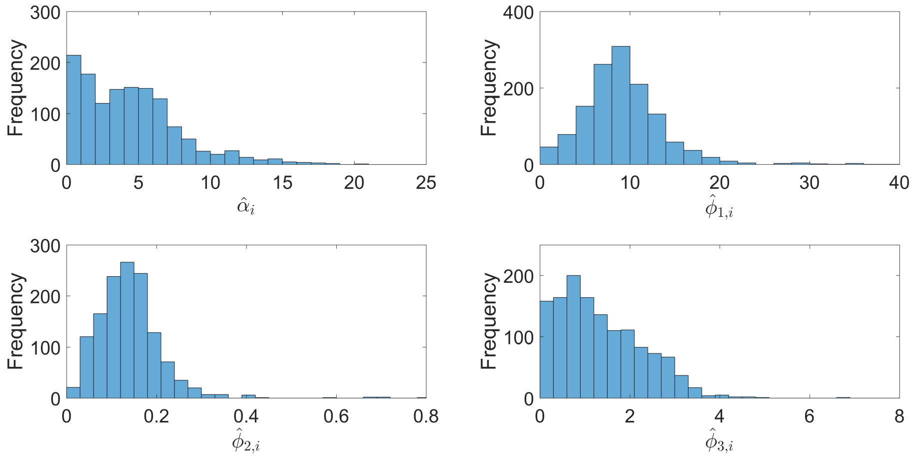

Fig. 4. Parameter estimates for all smartphones when f_t is a full-history mixture density.

Table 1. Percentage of times that a model (row) is better than the other (column) when the comparison is based on AIC.

	tracking-like	bimodal	full-history
tracking-like		0.67%	0.52%
bimodal	98.95%		2.40%
full-history	99.10%	97.60%	

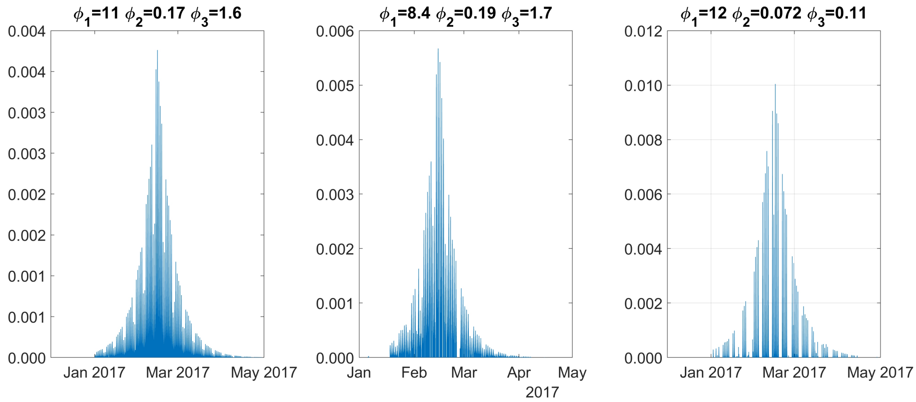

Fig. 5. Estimates of weights $v(t, t'; \hat{\boldsymbol{\phi}})$ when t is February 14th, 2017 12:00:00 for the smartphones shown in Figure 1.

tions and associated precisions are jointly used to estimate a spatial density for the smartphone location at any given time. The approaches are particularly suitable when the smartphone location is not observed at high temporal frequency and the sampling intervals are irregular, potentially with gaps of days or weeks. This is the case of smartphone locations collected by smartphone apps which make use of geolocation but the primary role of which is not tracking. The approaches are flexible and can be applied to analyze location data collected from any location-based app, including social networks.

Computationally, the model estimation time is linear in the number of observed locations and it is feasible on a laptop computer up to 10'000 locations. On the other hand, location prediction at any given point in time is almost real-time given the estimated model. The analysis is carried out using a MATLAB code.

References

1. Zorn, S., Rose, R., Goetz, A., Weigel, R.: A novel technique for mobile phone localization for search and rescue applications. In: 2010 International Conference on Indoor Positioning and Indoor Navigation. (2010) 1–4, DOI: 10.1109/IPIN.2010.5647107
2. Nyhan, M., Grauwin, S., Britter, R., Misstear, B., McNabola, A., Laden, F., Barrett, S.R.H., Ratti, C.: Exposure track: The impact of mobile-device-based mobility patterns on quantifying population exposure to air pollution. Environmental Science & Technology **50** (2016) 9671–9681
3. Do, T.M.T., Gatica-Perez, D.: Where and what: Using smartphones to predict next locations and applications in daily life. Pervasive and Mobile Computing **12** (2014) 79 – 91
4. Tremblay, Y., Shaffer, S.A., Fowler, S.L., Kuhn, C.E., McDonald, B.I., Weise, M.J., Bost, C.A., Weimerskirch, H., Crocker, D.E., Goebel, M.E., Costa, D.P.: Inter-

polation of animal tracking data in a fluid environment. Journal of Experimental Biology **209** (2005) 128–140
5. Liao, L., Patterson, D.J., Fox, D., Kautz, H.: Building personal maps from GPS data. Annals of the New York Academy of Sciences **1093** (2006) 249–265
6. González, M.C., Hidalgo, C.A., Barabási, A.: Understanding individual human mobility patterns. Nature **453** (2008) 779–782
7. Song, C., Qu, Z., Blumm, N., Barabási, A.L.: Limits of predictability in human mobility. Science **327** (2010) 1018–1021
8. Scellato, S., Musolesi, M., Mascolo, C., Latora, V., Campbell, A.T.: Nextplace: A spatio-temporal prediction framework for pervasive systems. In Lyons, K., Hightower, J., Huang, E.M., eds.: Pervasive Computing: 9th International Conference, Pervasive 2011, San Francisco, USA, June 12-15, 2011. Proceedings. Springer, Berlin, Heidelberg (2011) 152–169
9. Secchi, P., Vantini, S., Vitelli, V.: Analysis of spatio-temporal mobile phone data: a case study in the metropolitan area of Milan. Statistical Methods & Applications **24** (2015) 279–300
10. Cho, E., Myers, S.A., Leskovec, J.: Friendship and mobility: User movement in location-based social networks. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '11, New York, NY, USA, ACM (2011) 1082–1090
11. Lichman, M., Smyth, P.: Modeling human location data with mixtures of kernel densities. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '14, New York, NY, USA (2014) 35–44
12. Lawlor, S., Rabbat, M.G.: Time-varying mixtures of Markov chains: An application to road traffic modeling. Signal Processing IEEE Transactions **65** (2017) 3152–3167
13. Paci, L., Finazzi, F.: Dynamic model-based clustering for spatio-temporal data. Statistics and Computing, **28** (2017) 359–374
14. Cressie, N., Kornak, J.: Spatial statistics in the presence of location error with an application to remote sensing of the environment. Statistical Science **18** (2003) 436–456
15. Carroll, R.J., Ruppert, D., Stefanski, L.A., Crainiceanu, C.M.: Measurement Error in Nonlinear Models: A Modern Perspective. Chapman and Hall/CRC, Boca Raton (2006)
16. Finazzi, F.: The Earthquake Network project: Toward a crowdsourced smartphone-based earthquake early warning system. Bulletin of the Seismological Society of America **106** (2016) 1088–1099