Kernel-based estimation of individual location densities from smartphone data

Francesco Finazzi[1] **and Lucia Paci**[2]

[1] Department of Management, Information and Production Engineering, University of Bergamo, Bergamo, Italy

[2] Department of Statistical Sciences, Università Cattolica del Sacro Cuore, Milan, Italy

---

Address for correspondence: Lucia Paci, Department of Statistical Sciences, Università Cattolica del Sacro Cuore, Largo Gemelli, 1, Milano, Italy.

E-mail: `lucia.paci@unicatt.it`.

Phone: (+39) 02 72342491.

Fax: (+39) 02 72343064.

Abstract: Localising people across space and over time is a relevant and challenging problem in many modern applications. Smartphone ubiquity gives the opportunity to collect useful individual data as never before. In this work, the focus is on location data collected by smartphone applications. We propose a kernel-based density estimation approach that exploits cyclical spatio-temporal patterns of people to estimate the individual location density at any time, uncertainty included. Model parameters are estimated by maximum likelihood cross-validation. Unlike classic tracking methods designed for high spatio-temporal resolution data, the approach is suitable when location data are sparse in time and are affected by non negligible errors. The approach is applied to location data collected by the Earthquake Network citizen science project which carries out a world-wide earthquake early warning system based on smartphones. The approach is parsimonious and is suitable to model location data gathered by any location-aware smartphone application.

Key words: location-based applications; location error; likelihood cross-validation; kernel-based estimator; spatio-temporal patterns

1 Introduction

Nowadays, an increasing number of applications benefit from learning people location across space and over time. If an area is affected by a natural disaster, it is useful to know the last location of missing people (Zorn et al., 2010) . In epidemiological studies, pollution exposure can be dynamically assessed at individual level if people locations are known at high temporal resolution (Nyhan et al., 2016; Finazzi and Paci, 2019). From a commercial perspective, individual recommendation or advertising can be provided to people whose location is known (Do and Gatica-Perez, 2014). More generally, it can be useful to know people location at high temporal resolution for studying population dynamics and examining mobility patterns (Kelly et al., 2013).

Ideally, Global Positioning System (GPS) technology is suitable for tracking people at high spatial and temporal resolution. However, the main limit of GPS tracking is that a GPS receiver is needed and people are usually not equipped with it. Things have changed with the smartphone revolution. Smartphones, in fact, have a large number of built-in sensors including a GPS receiver. In 2018, smartphone penetration ranged from 50% to 80% in developed countries[1]. While this may suggest that tracking smartphones (and thus people) is an easy task, some issues are worth to be discussed.

Assuming that the smartphone user gives the permission to be tracked, a main 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 available through the service provider's network infrastructure or Wi-Fi networks at lower accuracy than GPS. Additionally, independently of the way location is obtained, acquiring and storing the smartphone location at high temporal frequency is considered a malpractice in smartphone application (app hereafter) programming. Quoting the Android guide for developers[2]: "For your app to be a good citizen, it should seek to limit its impact on the battery life of its device". 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 available. These considerations suggest that, in many circumstances, the information on the smartphone location is available at lower temporal resolution and precision when comparing to GPS-based tracking.

The analysis of individual location data has been explored by several authors to predict short-term trajectories. Since the seminal work of Kalman (Kalman, 1960), state space models (Durbin and Koopman, 2001) have been extensively used for trajectory estimation and tracking objects; see e.g. Jonsen et al. (2005) and Breed et al. (2012) for recent applications of dynamic linear models used to analyse animal track data. Also, Liao et al. (2006) employed a Bayesian dynamic network for learning

[1] `https://newzoo.com/insights/rankings/top-50-countries-by-smartphone-penetration-and-users/`

[2] `https://developer.android.com/training/monitoring-device-state/index.html`

people transportation routines and to build personal map based on their behaviour. Katenka et al. (2013) introduced a penalized maximum likelihood framework for tracking people and animals over time using binary decisions collected by a wireless sensor network.

Tracking methods based on dynamic modelling, however, do not appear suitable in our framework. Indeed, it is hard to recover the actual trajectory of a smartphone when its location is sampled at low temporal frequency and in the absence of additional information, such as speed and acceleration in space. Secondly, when the smartphone user moves from a point in space to another, the 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 artefacts (streets, roads, etc.).

A different approach to predict individual location relies on the reproducibility of human patterns. In fact, daily and weekly routines are well-established in human society, i.e., human activities are characterised by a certain degree of regularity and predictability (González et al., 2008; Song et al., 2010). In this framework, Scellato et al. (2011) provided a spatio-temporal approach to predict arrival and residence times of users in their relevant places. Secchi et al. (2015) analysed functional mobile data to identify sub-regions 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.), then it is possible to group all the observed spatial locations into a small number of clusters. As common in literature, clusters might be identified using mixture modelling. For instance, Cho et al. (2011) employed a two-state mixture of Gaussian distributions centred at "home" and "work" locations to understand human motion from cell phone data and social networks. Lichman and Smyth (2014) modelled 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 spatial patterns ignoring the temporal dimension, i.e., assuming a time-invariant location density.

Our contribution is to address the challenge of estimating the smartphone/person location density at any given point in time. When the density is thought to change over time, it might be estimated nonparametrically by using a kernel and by weighting the observations through a scheme derived from time series modelling (Harvey and Oryshchenko, 2012). Following this track, a kernel-based approach (Silverman, 1986; Wand and Jones, 1995) is tailored to fit our smartphone location data in order to provide a time-varying location density of the smartphone/person using all the locations collected by the smartphone app over a given period. In particular, starting from a time-invariant location density, we introduce a new weighting scheme that exploits the reproducible patterns of people location, such as daily and weekly patterns. Weights are time-varying and they depend on a small number of unknown parameters estimated through maximum likelihood cross-validation. We also account for the positional error (Cressie and Kornak, 2003) arising in smartphone data.

The approach has several advantages. Contrary to classic state space models adopted in trajectory estimation, our estimator is able to describe spatial densities characterised by multiple modes. Secondly, when estimating the person location, the flexible form of the mixing weights allows to use the most recent locations when available, otherwise the daily and weekly patterns come into play. Finally, the approach is parsimonious and computationally feasible.

Our motivating data are smartphone locations collected by Earthquake Network citizen science project (www.earthquakenetwork.it), which implements a world-wide earthquake early warning system based on smartphones. This is a case of location data collected by a smartphone app which make use of geolocation but the primary role of which is not tracking. Similarly, the approach can be applied for modelling location data gathered by any location-based app, including social networks. We also offer a comparison of our approach with a state space model, customary employed for people tracking; results show how our kernel-based estimator outperforms the state space model in terms of accuracy in locating people.

The remainder of the paper is organized as follows. Section 2 introduces smartphone location data used in this study. In Section 3 we first describe a time-invariant location density estimator and then move to a time-varying density by weighting the observations according to a scheme that accounts for the cyclical pattern of people. Details about model estimation are given in Section 4. Section 5 illustrates the results of the real data analysis. We conclude with a brief discussion in Section 6.

2 Earthquake Network location data

An example of location-aware app is the Android app developed by the Earthquake Network project (Finazzi, 2016), which implements a world-wide early warning system based on smartphones. Smartphones with the app installed take part to the project and they are used to detect earthquakes in real-time. This allows to alert in advance the population living above a given distance from the epicentre. As of today, almost

5 million of people world-wide took part to the project. The earthquake detection is based on the signals sent from the smartphones to a central server, with information on the smartphone location and its precision. A statistical algorithm is used to understand in real-time if an earthquake is occurring, controlling the probability of false alarm (Finazzi and Fassò, 2017).

Each smartphone in the network periodically sends a signal to the server, roughly every 30 minutes. The smartphone also sends a signal when an acceleration above a threshold is detected. However, the smartphone does not send signals when it is switched off or the Internet is not available. As a result, the sampling interval of the smartphone location is not regular.

A useful service offered by Earthquake Network is the geolocation alert. When an earthquake is detected in real-time, an estimate of the user location is sent to a list of trusted contacts. This is helpful for search and rescue operations in the case the person is missing after the quake. In this context, it is useful to provide an accurate estimate of the smartphone location (equipped with uncertainty information) even when the smartphone did not send its coordinates for a long time or it was off during the earthquake. Therefore, location data used for earthquake detection are also exploited for estimating the smartphone/person location when the quake strikes.

Specifically, locations provided by the smartphone are expressed in terms of coordinates and precision. The smartphone location at time t' ($t' = 1, \ldots, T$) is given as the probability density function of a bivariate Normal distribution centred on $\mathbf{s}_{t'}$ and with covariance $\mathbf{\Sigma}_{t'}$, i.e., $\varphi\left(\mathbf{s}; \mathbf{s}_{t'}, \mathbf{\Sigma}_{t'}\right)$, $\mathbf{s} \in S^2$ with S^2 the sphere in $\mathbb{R}^3$. Namely, the mean vector $\mathbf{s}_{t'} = (\mathrm{lat}_{t'}, \mathrm{lon}_{t'})'$ is the vector of latitude and longitude coordinates and represents the location measured by the smartphone, while $\mathbf{\Sigma}_{t'} = \sigma^2_{t'}\mathbf{I}_2$ is a diagonal matrix, i.e., $1/\sigma_{t'}$ is the location precision also measured by the smartphone. For instance, $\sigma_{t'}$ is usually equal to 5 meters when the GPS receiver of the smartphone is on, while it is equal to 20 meters when it is off and the smartphone localizes itself using cell towers. Note that $\varphi\left(\mathbf{s};\, \mathbf{s}_{t'}, \boldsymbol{\Sigma}_{t'}\right)$ is not a valid density function when the support is the sphere. In this work, however, the focus is on small geographic regions if compared with the Earth surface. Moreover, $\sigma^2_{t'}$ is also small when compared with the Earth circumference. Thus, it is advisable to adopt a map projection and a 2-dimensional Cartesian coordinate system such as the Universal Transverse Mercator (UTM) coordinate system. If UTM is adopted, a location $\mathbf{s}_{t'}$ on the Earth is given by a UTM zone number and by the easting and northing coordinates within that zone. It follows that $\mathbf{s} \in \mathbb{R}^2$ and $\varphi\left(\mathbf{s};\, \mathbf{s}_{t'}, \boldsymbol{\Sigma}_{t'}\right)$ is a valid density function.

In this work, smartphones located in the municipality of Rome, Italy, are available. The data set consists of signals sent by the smartphones to the central 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) corresponding to an area of roughly 700 km^2 that covers the municipality of Rome. The number of smartphones available over the selected period is $1,188$, with a number of signals per smartphone ranging from around 100 to around $5,300$ and an average of around $3,000$ signals. This shows that the number of available locations, per smartphone, is relatively small if compared with data from tracking apps. Indeed, assuming 12 hours of smartphone activity per day, GPS tracking would have collected around 1 million of locations over the same period.

As an example, Figure 1 shows the locations provided by three different smartphones over the period with associated standard deviations represented by disks; note that the three users exhibit quite different patterns, both in terms of location and preci-

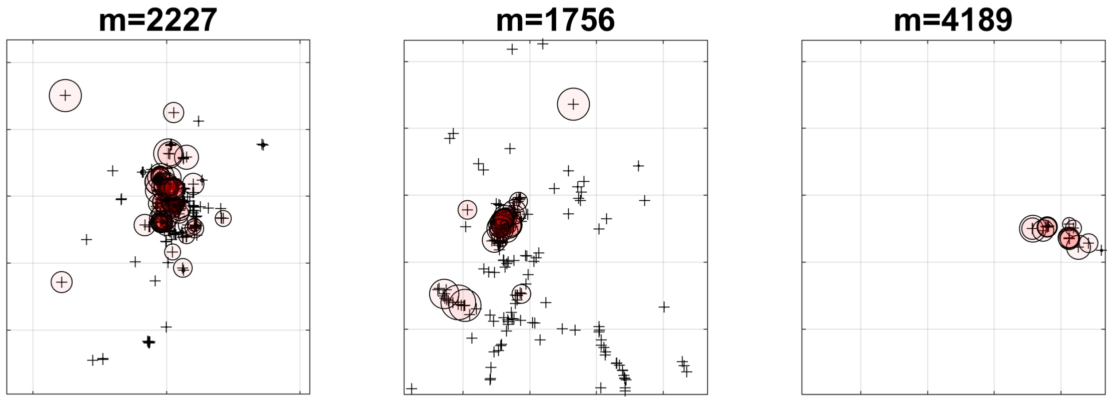

Figure 1: Spatial locations provided by three smartphones within the same geographic area (axis labels removed for privacy reason). The centre of the disk is the observed location while the radius of the disk is the associated standard deviation; disks corresponding to standard deviations of roughly 5 meters are not visible in the plots. The area in the plot covers the municipality of Rome.

sion. Figure 2 displays the distributions of the sampling interval (left panel) and the standard deviation $\sigma_{t'}$ (right panel) related to the 3.8 million total signals sent by the $1,188$ smartphones. The sampling interval has a median equal to around 30 minutes while the 95th and the 99th percentiles are around 66 and 484 minutes, respectively. The median of the distribution of σ_t is around 20 meters while the 95th and the 99th percentiles are around 860 and 1300 meters, respectively. Moreover, the distribution of $\sigma_{t'}$ is multi-modal with modes at some integer values. Relevant statistics about the signals aggregated at the smartphone level are provided in the Supplementary materials available at `http://www.statmod.org/smij/archive.html`.

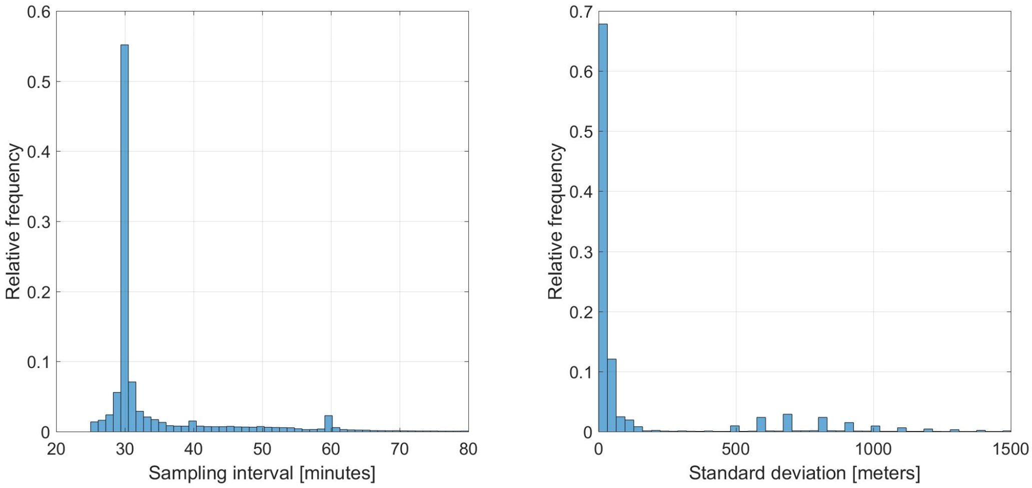

Figure 2: Histogram of the smartphone location sampling intervals (left panel) and histogram of the smartphone location standard deviations (right panel) for around 3.8 million signals received by the server of the Earthquake Network project.

3 Location density functions

To facilitate the understanding of location densities, we start by discussing a kernel-based estimator (Silverman, 1986; Wand and Jones, 1995) for the time-invariant location density of a given individual using all the locations registered by his/her smartphone. Then, we enrich the estimator by introducing a new weighting scheme that leads to a time-varying location density at any generic time $t \in \mathbb{R}^+$.

Let $F(\mathbf{s})$ be the time-invariant distribution of the individual location. Given a sample of T observations drawn from $F(\mathbf{s})$ with corresponding probability density function $f(\mathbf{s})$, we look for an estimator of $f(\mathbf{s})$. We recall that our location data are augmented by a positional error that we employ for building an adaptive kernel-based estimator of $f(\mathbf{s})$, that is

$$\hat{f}_T(\mathbf{s};\alpha) = \frac{1}{T}\sum_{t'=1}^{T} \varphi\left(\mathbf{s};\, \mathbf{s}_{t'}, (1+\alpha)\mathbf{\Sigma}_{t'}\right), \tag{3.1}$$

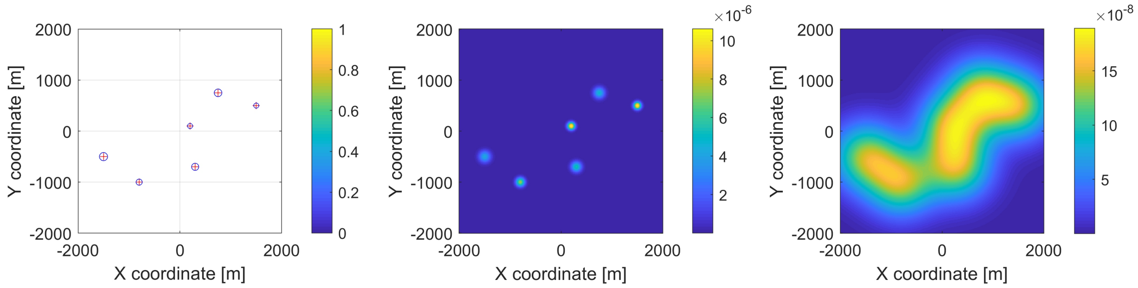

Figure 3: Example to clarify the role of α in (3.1). Left panel: simulated locations (cross markers) with associated standard deviations (disks). Middle panel: location density estimated by (3.1) when $\alpha = 0$. Right panel: location density estimated by (3.1) when $\alpha = 500$. Different legends are used to highlight the differences between the two plots.

where $\varphi\left(\mathbf{s}; \mathbf{s}_{t'}, (1+\alpha)\mathbf{\Sigma}_{t'}\right)$ denotes the bivariate normal density centred on $\mathbf{s}_{t'}$ and with covariance $(1+\alpha)\mathbf{\Sigma}_{t'}$.

The parameter α in (3.1) mimics a global bandwidth in adaptive kernel-based estimators while $\mathbf{\Sigma}_{t'}$ plays the role of a local bandwidth factor. To clarify, we consider the scenario depicted in Figure 3. The cross markers in the left panel are simulated locations $\mathbf{s}_{t'}$ while the corresponding circles represent the standard deviation $\sigma_{t'}$ of the normal density centred on $\mathbf{s}_{t'}$. The middle panel displays the location density estimated by (3.1) when $\alpha = 0$. Conversely, the right panel shows the estimated location density when $\alpha = 500$. Hence, the parameter α controls the overall degree of smoothing of the estimated density such that the spread of the density increases as α increases. The parameter α is unknown and must be estimated; here, maximum likelihood cross-validation is employed to estimate α. Finally, note that $(1+\alpha)$ in (3.1) reflects the fact that an observation cannot take part to the density estimator with a variance smaller than the one provided by the smartphone.

3.1 Time-varying location density

The main drawback of estimator (3.1) is that the estimated density is constant over time. Rather, our goal is to estimate the probability density function of the individual location at any time t, i.e., a time-varying density function. To accomplish this, we introduce a weighting scheme in (3.1) that discounts far-in-time observations and exploits the fact that people usually follow cyclical patterns. We propose a smoothing kernel-based estimator of the time-varying location density (MIX), that is

$$\hat{f}_{t|T} = \hat{f}_{t|T}(\mathbf{s};\boldsymbol{\theta}) = \sum_{t'=1}^{T} w(t,t';\boldsymbol{\phi})\varphi\left(\mathbf{s};\,\mathbf{s}_{t'},(1+\alpha)\boldsymbol{\Sigma}_{t'}\right), \tag{3.2}$$

where $\boldsymbol{\theta} = \{\alpha,\boldsymbol{\phi}\}$ and $\sum_{t'}^{T} w(t,t';\boldsymbol{\phi}) = 1$. Here, α has similar interpretation as above; it modulates the covariance of all observed locations such that the higher α the more $\hat{f}_{t|T}$ is spread. In particular, when $\alpha \to \infty$ the density converges to the uniform density over $\mathbb{R}^2$, meaning that the estimated density is not informative on the smartphone location.

The mixing weights $w(t,t';\boldsymbol{\phi})$ are defined as

$$w(t,t';\boldsymbol{\phi}) = \frac{u(t,t';\boldsymbol{\phi})}{\sum_{t'=1}^{T} u(t,t';\boldsymbol{\phi})}, \tag{3.3}$$

with

$$u(t,t';\boldsymbol{\phi}) = \exp\left(-\frac{|t-t'|}{\phi_1}\right)\exp\left(-\frac{h(t,t')}{\phi_2}\right)\exp\left(-\frac{1-d(t,t')}{\phi_3}\right). \tag{3.4}$$

Mixing weights (3.3) - (3.4) depend on temporal interval $|t-t'|$ as well as they describe daily and weekly cyclical patterns of smartphone users. Indeed, the function $h\left(t,t'\right) = \min\left(\left(t-t'\right) - \lfloor t-t' \rfloor, 1-\left(t-t'\right) - \lfloor t-t' \rfloor\right)$ returns the difference in time between t and t' considering the circularity of the day and independently of the calendar day, namely $h\left(t, t'\right)$ is always less than 0.5 (12 hours) even when t and t' are more than 12 hours far apart. On the other hand, the function $d\left(t, t'\right)$ is equal to 1 if t and t' are of the same type, namely both working days or both weekend, otherwise it is equal to zero. In principle, the function $d(t, t')$ can be specified to distinguish all days of the week, leading to an increased number of corresponding parameters to estimate. Here, $w(t, t'; \boldsymbol{\phi})$ tends to be small when t and t' are far in time and/or when they are characterised by a different time within the day and/or when they are related to days which are not of the same type.

The unknown parameters $\boldsymbol{\phi} > 0$ in the exponential terms of (3.4) describe the temporal persistence of the information carried by each 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 smaller ϕ_3 the higher the weekend effect, meaning that it is important to use smartphone locations belonging to the same day type to predict the smartphone location at time t. The simulation study in the Supplementary materials clarifies the behaviour of the weighting scheme (3.3) - (3.4).

It is worth noting that when $\hat{f}_{t|T}$ is used to predict the smartphone location far in time from the last available observed location, then the first exponential term in (3.4) approaches zero. However, the normalized exponential function in (3.3) implies that $\hat{f}_{t|T}$ still reflects the daily and the weekly cycles estimated from the information set. In this sense, daily and weekly cycles characterise the smartphone location regardless of the position of t along the time line, and $\hat{f}_{t|T}$ is informative even if the smartphone has not sent its location for a long time.

3.2 Trimming

Although our approach is suitable to deal with locations that are sparse in time, the number of observations can substantially increase when the observational time frame increases, leading to a higher computational cost. In this situations, trimming the number of components entering in (3.2) provides a natural solution.

To expedite the computation when the number of observed locations is large, the number of components can be reduced to a subset of the observed locations. Similar to Lawlor and Rabbat (2017), we can trim the weak components in (3.2) that do not contribute to describe $\hat{f}_{t|T}$. In other words, we can remove observations that are associated with small weights $w(t, t'; \boldsymbol{\phi})$. For instance, a fixed threshold strategy can be adopted such that, at each time t, the components entering in (3.2) are only those related to the temporal indexes

$$\mathcal{T}_t = \left\{ t', t' = 1, \ldots, T : w\left(t, t'; \boldsymbol{\phi}\right) > \max\bigl(w\left(t, t'; \boldsymbol{\phi}\right)\bigr)/\delta \right\}, \tag{3.5}$$

where δ is a constant value. We refer the reader to Crouse et al. (2011) for a review of pruning strategies and mixture reduction algorithms. More details about trimming and a strategy for choosing δ are given in the Supplementary material.

4 Estimation

The mixture parameters are estimated by maximizing the likelihood cross-validation (LCV) criterion (Silverman, 1986; Harvey and Oryshchenko, 2012; Wu, 2018). To derive the LCV, it is helpful to think about the observed location as the noisy version of the "true" unknown location $\widetilde{\mathbf{s}}_t$, that is,

$$\mathbf{s}_t = \widetilde{\mathbf{s}}_t + \boldsymbol{\varepsilon}_t, \tag{4.1}$$

where $\boldsymbol{\varepsilon}_t$ are independent normally distributed errors with covariance $\boldsymbol{\Sigma}_t$, i.e., $\boldsymbol{\varepsilon}_t \sim \mathcal{N}_2\left(\mathbf{0}, \boldsymbol{\Sigma}_t\right)$. If $\hat{f}_{t|T}(\mathbf{s};\boldsymbol{\theta})$ is the estimator of the density of the true location $\widetilde{\mathbf{s}}_t$, then the estimator of the density of the observed location $\mathbf{s}_t$ becomes

$$\hat{g}_{t|T}(\mathbf{s};\boldsymbol{\theta}) = \sum_{t'=1}^{T} w(t,t';\boldsymbol{\phi})\varphi\left(\mathbf{s};\, \mathbf{s}_{t'}, (1+\alpha)\boldsymbol{\Sigma}_{t'} + \boldsymbol{\Sigma}_t\right). \tag{4.2}$$

In other words, given (4.1), the covariance of all components in (4.2) is "increased" by the covariance $\boldsymbol{\Sigma}_t$ associated with the location $\mathbf{s}_t$. Hence, the log LCV is given by

$$\begin{aligned} \log \mathcal{LCV}\left(\boldsymbol{\theta};\mathcal{S}\right) &= \frac{1}{T}\sum_{t=1}^{T} \hat{g}_{(-t)|T}(\mathbf{s}_t;\boldsymbol{\theta}) \\ &= \frac{1}{T}\sum_{t=1}^{T}\log\left[\sum_{t'\neq t;\, t'=1}^{T} w(t,t';\boldsymbol{\phi})\,\varphi\left(\mathbf{s}_t;\, \mathbf{s}_{t'}, (1+\alpha)\boldsymbol{\Sigma}_{t'} + \boldsymbol{\Sigma}_t\right)\right], \end{aligned} \tag{4.3}$$

where $\mathcal{S}$ is the set of observed locations and variances. Note that, for each t, all the observed locations take part into the sum with the exception of the location $\mathbf{s}_t$, which is the location where the density is evaluated. This suggests that maximising (4.3) corresponds to find the best model parameters that guarantee high value of $\hat{f}_{t|T}$ at the smartphone location when all the locations are used to estimate the smartphone location at time t.

Note also that, when the covariance $\boldsymbol{\Sigma}_t$ goes to infinity, the density collapses to a uniform density over $\mathbb{R}^2$. In this case, the contribution to the likelihood is constant and does not depend on $\mathbf{s}_t$ nor on the values of the model parameters. Conversely, a covariance close to zero implies that the observed location is very informative on parameter estimation.

As a final remark, we notice that the log LCV has the same form of (4.3) in the case of trimming, but the inner sum involves only the observed locations related to weights higher than the trimming threshold.

The log LCV is maximized with respect to $\boldsymbol{\theta}$ under the constraint that all elements of $\boldsymbol{\theta}$ are positive. Maximisation is performed using numerical optimization algorithms (e.g., the Quasi-Newton algorithm) and the approximated Hessian matrix $\hat{\mathbf{H}}$ at the maximum is computed. Model estimation is performed separately for each smartphone.

The simulation study presented in the Supplementary materials shows the capability of the LCV estimation procedure to recover model parameters. Results suggest that the model is identifiable and that the estimator of $\boldsymbol{\theta}$ is not biased for most of the scenarios considered in the simulation study.

Finally, a Monte Carlo (MC) estimate of the standard deviation of the location density estimator in (3.2) is given as follows. We draw a sample of size B from the asymptotic distribution of the estimated parameter set, say $\boldsymbol{\theta}_l^* \sim \mathcal{N}\left(\hat{\boldsymbol{\theta}}, \hat{\mathbf{H}}^{-1}\right)$, with $l = 1, \dots, L$. Then, we compute the location density for each of the L parameter set over a discrete spatial domain. The empirical standard deviation at each spatial point provides an estimate of the desired uncertainty associated with the estimated location density.

5 Analysis of smartphone location data

We illustrate our approach by modelling smartphone location data collected by Earthquake Network as described in Section 2.

5.1 Estimation results

For each of the $1,188$ smartphones, the location density parameters are estimated using the LCV described in Section 4 without any trimming. Since the likelihood function may be characterised by multiple local maxima and the maximisation of (4.3) is not guaranteed to converge to the global maximum, the maximisation is carried out multiple times with different randomly chosen starting values for the parameter vector $\boldsymbol{\theta}$. The estimate $\hat{\boldsymbol{\theta}}$ with the highest likelihood is then retained. Although this approach does not guarantee that the global maximum of the likelihood function is found, changing the starting values reduces the possibility of providing an estimate $\hat{\boldsymbol{\theta}}$ related to a "bad" local maximum of the likelihood function.

The histograms of the estimated $\hat{\alpha}$ values and the estimated $\hat{\boldsymbol{\phi}}$ values for the $1,188$ smartphones are available in the Supplementary materials. As an instance, Figure 4 shows the estimated weights $w(t, t'; \hat{\boldsymbol{\phi}})$ when t is February 20th, 2017 05:30 PM corresponding to the three smartphones plotted in Figure 1. For smartphone in the left panel, the nearest observed locations are five days far apart from February February 20th, 2017 05:30 PM; this explains the absence of components entering in the estimator around that time.

5.2 Model comparison

We evaluate the accuracy of our approach using a cross-validation study, that is, for each smartphone, we hold out the 10% of time points and we compare the estimated density with the density location provided by the smartphone. The simulation study presented in the Supplementary materials showed that time-invariant estimator (3.1)

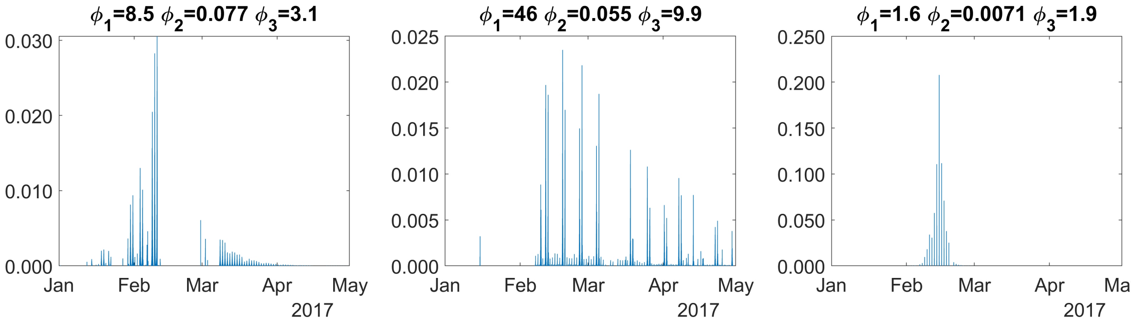

Figure 4: Estimates of weights $w(t, t'; \hat{\boldsymbol{\phi}})$ when t is February 20th, 2017 05:30 PM for the three smartphones shown in Figure 1.

does not provide accurate predictions of the true locations. Rather, a natural benchmark for model comparison is given by the state space model (Durbin and Koopman, 2001), usually employed for tracking.

In particular, let $\mathbf{x}_t$ be a $p \times 1$ vector containing p predictors. A dynamic linear model (DLM) is given by

$$\begin{aligned} \mathbf{s}_t &= \mathbf{X}_t \boldsymbol{\beta} + \widetilde{\mathbf{s}}_t + \boldsymbol{\epsilon}_t \\ \widetilde{\mathbf{s}}_t &= \mathbf{G}\, \widetilde{\mathbf{s}}_{t-1} + \boldsymbol{\eta}_t, \end{aligned} \tag{5.1}$$

where $\mathbf{X}_t = \mathbf{I}_2 \otimes \mathbf{x}_t'$, $\boldsymbol{\beta}$ is the $2p \times 1$ vector of coefficients and $\mathbf{G}$ is the 2×2 transition matrix. Finally, $\boldsymbol{\eta}_t \sim N(\mathbf{0}, \boldsymbol{\Lambda})$ is the innovation error and $\boldsymbol{\epsilon}_t \sim N\left(\mathbf{0}, \sigma_t^2 \mathbf{I}_2\right)$ is the measurement error with known variance σ_t^2.

To capture the cyclical pattern of individual locations, $\mathbf{x}_t$ includes 48 dummy variables to distinct between working days and weekend days and all the 24 hours in each day. Parameters $\boldsymbol{\Psi} = (\boldsymbol{\beta}, \mathbf{G}, \boldsymbol{\Lambda})$ of model (5.1) are estimated via maximum likelihood and the EM algorithm using the D-STEM software (Finazzi and Fassò, 2014). Hence, the estimated density $\hat{f}_{t|T}\left(\mathbf{s};\, \boldsymbol{\Psi}\right)$ at validation times is a normal density with mean $E_{\boldsymbol{\Psi}}\left(\mathbf{s}; \mathcal{S}\right)$ and variance $Var_{\boldsymbol{\Psi}}\left(\mathbf{s}; \mathcal{S}\right)$ given by the Kalman smoother. Finally, in order to provide a fair comparison, we discretise the data set in regular time steps of 30 minutes.

As an illustration, Figure 5 shows the estimated location densities at a given time obtained from the DLM (left panel) and our approach (right panel) for the smartphone displayed in the left panel of Figure 1. The star represents the location provided by the smartphone at validation time February 20th, 2017 05:30 PM and the disk is the associated standard deviation. The region plotted in the right panel covers the region within the square box on the left panel. Clearly, the location density estimated by MIX is more concentrated on the observed smartphone location than the estimates from the DLM. Indeed, the nearest observed location are five days far apart from the validation time, as we argued in Section 5.1. Also, note that the density has multiple local maxima, which correspond to places visited by the person at around the same time of the day.

The left panel of Figure 6 displays the MC standard deviation of the location density estimator which exhibits a spatial pattern similar to the estimated density but with a lower magnitude. The right panel of Figure 6 presents a relative measure of uncertainty given by the ratio (in percentage) between the MC standard deviation and the estimated location density. Depicting the ratio allows to easily highlight areas of large relative uncertainty.

Recall that the true user location is unknown, even within the cross-validation setting. Indeed, at validation times, we only have the location density provided by the smartphone, i.e., $\varphi\left(\mathbf{s}; \mathbf{s}_t, \mathbf{\Sigma}_t\right)$. In other words, model comparison cannot be provided by evaluating the estimated density at the real user location, since such location is always unknown. Rather, we propose to use $\varphi\left(\mathbf{s}; \mathbf{s}_t, \mathbf{\Sigma}_t\right)$ as a spatial weight in the

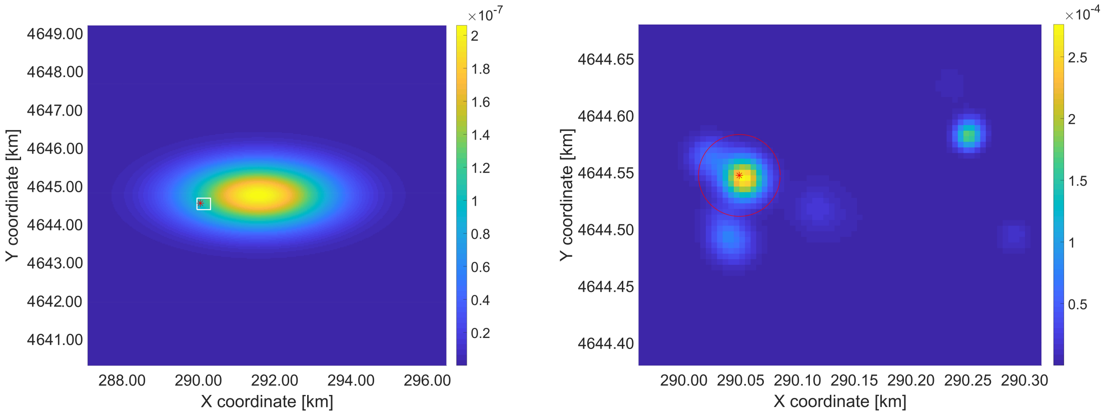

Figure 5: Estimated location densities obtained from the DLM (left panel) and our proposed mixture (right panel) for the smartphone user displayed in the left panel of Figure 1. The star represents the location provided by the smartphone at validation time February 20th, 2017 05:30 PM and the disk is the associated standard deviation. The region plotted in the right panels covers the region within the square box on the left panel.

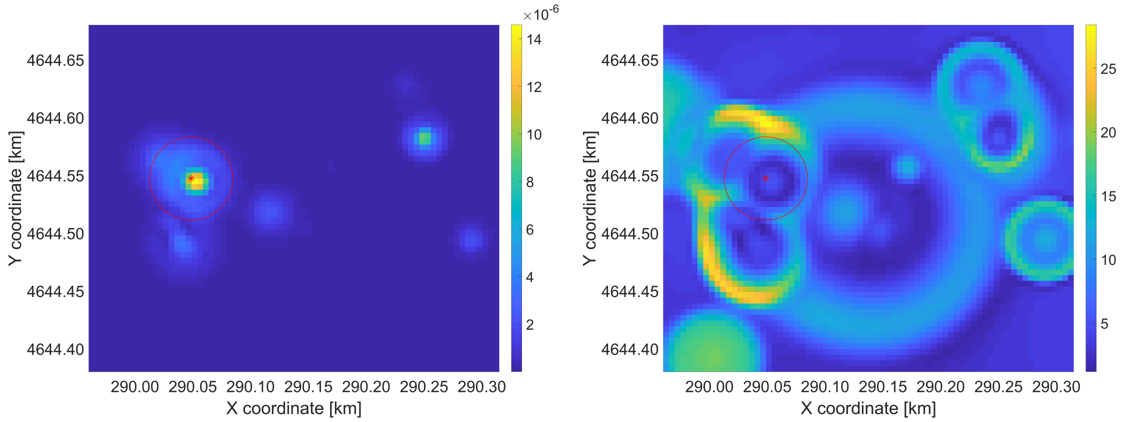

Figure 6: Left panel: MC standard deviation associated to the estimated location density $\hat{f}_{t|T}$ presented in the right panel of Figure 5. Right panel: relative standard deviation. The star represents the location provided by the smartphone at validation time February 20th, 2017 05:30 PM and the disk is the associated standard deviation.

following criteria,

$$d_t^{MIX} = \int \hat{f}_{t|T}\left(\mathbf{s}; \hat{\boldsymbol{\theta}}\right) \varphi\left(\mathbf{s};\, \mathbf{s}_t, \boldsymbol{\Sigma}_t\right) d\mathbf{s}, \tag{5.2}$$

$$d_t^{DLM} = \int \hat{f}_{t|T}\left(\mathbf{s}; \hat{\boldsymbol{\Psi}}\right) \varphi\left(\mathbf{s};\, \mathbf{s}_t, \boldsymbol{\Sigma}_t\right) d\mathbf{s}. \tag{5.3}$$

For a given validation time t, the model with the highest d_t is preferred. To clarify, when the uncertainty on the user location is very high, $\varphi\left(\mathbf{s};\, \mathbf{s}_t, \boldsymbol{\Sigma}_t\right)$ tends to be uniform across space so that d_t^{MIX} and d_t^{DLM} tend to be equal (since the integral of $\hat{f}_{t|T}$ is 1). It follows that neither of the two models is favoured. On the other hand, when $\varphi\left(\mathbf{s};\, \mathbf{s}_t, \boldsymbol{\Sigma}_t\right)$ is very concentrated, the comparison is essentially based on $\hat{f}_{t|T}\left(\mathbf{s}; \hat{\boldsymbol{\theta}}\right)$ and $\hat{f}_{t|T}\left(\mathbf{s}; \hat{\boldsymbol{\Psi}}\right)$ evaluated in $\mathbf{s}_t$.

Figure 7 shows the comparison between our approach and the DLM based on criteria (5.2) and (5.3) over all validation times. In particular, the figure displays the histogram of the proportion of validation times for which $d_t^{MIX} > d_t^{DLM}$ over the $1,188$ smartphones. Clearly, our MIX estimator outperforms the DLM most of the times and for most of the users.

6 Summary and future works

In this work, we addressed the problem of estimating the location of a smartphone/person using historical location data collected by a smartphone app. Specifically, smartphone locations and associated precisions are jointly used to estimate a spatial density for the person location at any given time, assuming that smartphone and person are together.

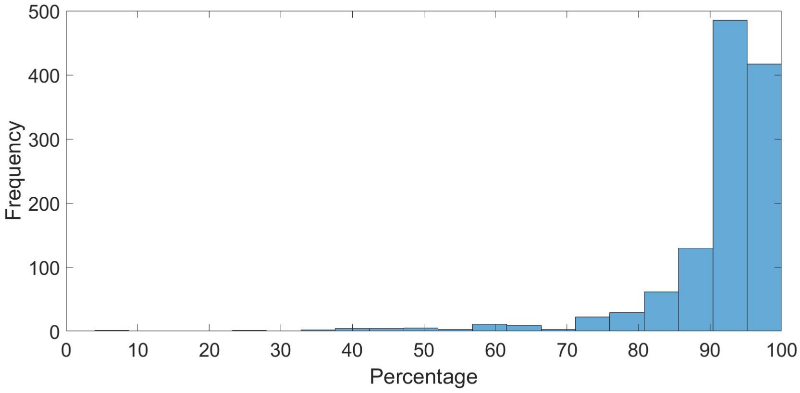

Figure 7: Proportion of validation times for which $d_t^{MIX} > d_t^{DLM}$ over $1,188$ smartphones, i.e, the proportion of times that our approach (MIX) is better than the dynamic linear model (DLM).

The approach is suitable to situations when smartphone locations are sparse in time and sampling intervals are irregular, potentially with gaps of days or weeks. This is the case, for instance, of smartphone locations collected by apps which make use of geolocation but the primary role of which is not tracking. Hence, the approach can be employed to analyse location data collected by any location-based app, including social networks. Conversely, our estimator is not appropriate for analysing high frequency data, e.g., for real-time people tracking.

Our proposed kernel-based estimator is parsimonious in terms of parameters and is able to describe a multi-modal density, uncertainty included. Moreover, the positional error arising in smartphone location data is appropriately accounted for and employed to build the adaptive estimator. When compared to classic state space modelling, the estimated location density is less spread and it is concentrated around locations where the person is expected to be found, even when the smartphone has not sent its location for a long period.

Computationally, the model estimation time is linear in the number of observed locations and it is feasible on laptop computers. 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 available at http://www.statmod.org/smij/archive.html.

Attention has been focused on the cyclical behaviour of people movements by distinguishing between weekdays and weekends. However, alternative weighting schemes can be proposed and explored. For instance, the weighting scheme can benefit from a higher resolution, say the different days of the week. On the other side, our approach it is not suitable to capture sudden and radical changes in smartphone/person behaviour, since it is designed to exploit people cyclical routines.

Another avenue of development for our work concerns how to include potential covariate information in the density estimation. Indeed, some apps might have further location information about the user that could be highly relevant, e.g., home address, work environment, commuting plans. Semi-parametric mixtures (Wang and Chee, 2012) models offer a suitable framework to move towards this extension.

Future work will also find us at investigating models that allow to borrow strength information among smartphones, relaxing the independence assumption for groups of people. For instance, parameters shared by multiple smartphones can be introduced in (3.2). The inference of such models is computational challenging and the topic is currently in progress.

Acknowledgements

We thank the Editor, the anonymous Associated Editor and reviewers for their useful comments which have improved the paper.

Conflict of interest statement

The data set used in this article was collected via the commercial Android Earthquake Network app developed by the first author.

References

Breed, G. A., Costa, D. P., Jonsen, I. D., Robinson, P. W., and Mills-Flemming, J. (2012). State-space methods for more completely capturing behavioral dynamics from animal tracks. *Ecological Modelling*, **235-236**, 49 – 58.

Cho, E., Myers, S. A., and Leskovec, J. (2011). 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, pages 1082–1090, New York, NY, USA, 2011. ACM.

Cressie, N. and Kornak, J. (2003). Spatial statistics in the presence of location error with an application to remote sensing of the environment. *Statistical Science*, **18**, 436–456.

Crouse, D. F., Willett, P., Pattipati, K., and Svensson, L. (2011). A look at Gaussian mixture reduction algorithms. In *14th International Conference on Information Fusion*, pages 1–8.

Do, T. M. T. and Gatica-Perez, D. (2014). Where and what: Using smartphones to predict next locations and applications in daily life. *Pervasive and Mobile Computing*, **12**, 79 – 91.

Durbin, J. and Koopman, S. (2001). *Time Series Analysis by State Space Methods.* Oxford University Press, Oxford.

Finazzi, F. and Paci, L. (2019). Quantifying personal exposure to air pollution from smartphone-based location data. *Biometrics.* doi: 10.1111/biom.13100.

Finazzi, F. (2016). The Earthquake Network project: Toward a crowdsourced smartphone-based earthquake early warning system. *Bulletin of the Seismological Society of America*, **106**, 1088–1099.

Finazzi, F. and Fassò, A. (2014). D-STEM: A software for the analysis and mapping of environmental space-time variables. *Journal of Statistical Software, Articles*, **62**, 1–29.

Finazzi, F. and Fassò, A. (2017). A statistical approach to crowdsourced smartphone-based earthquake early warning systems. *Stochastic Environmental Reseaerch Risk Assessment*, **31**, 1649–1658.

González, M. C., Hidalgo, C. A., and Barabási, A. (2008). Understanding individual human mobility patterns. *Nature*, **453**, 779–782.

Harvey, A. and Oryshchenko, V. (2012). Kernel density estimation for time series data. *International Journal of Forecasting*, **28**, 3 – 14.

Jonsen, I. D., Flemming, J. M., and Myers, R. A. (2005). Robust state-space modeling of animal movement data. *Ecology*, **86**, 2874–2880.

Kalman, R. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, **82**, 35–45.

Katenka, N., Levina, E., and Michailidis, G. (2013). Tracking multiple targets using binary decisions from Wireless sensor networks. *Journal of the American Statistical Association*, **108**, 398–410.

Kelly, D., Smyth, B., and Caulfield, B. (2013). Uncovering measurements of social and demographic behavior from smartphone location data. *IEEE Transactions on Human-Machine Systems*, **43**, 188–198.

Lawlor, S. and Rabbat, M. G. (2017). Time-varying mixtures of Markov chains: An application to road traffic modeling. *Signal Processing IEEE Transactions*, **65**, 3152–3167.

Liao, L., Patterson, D. J., Fox, D., and Kautz, H. (2006). Building personal maps from GPS data. *Annals of the New York Academy of Sciences*, **1093**, 249–265.

Lichman, M. and Smyth, P. (2014). 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, pages 35–44, New York, NY, USA, 2014.

Nyhan, M., Grauwin, S., Britter, R., Misstear, B., McNabola, A., Laden, F., Barrett, S. R. H., and Ratti, C. (2016). Exposure track: The impact of mobile-device-based mobility patterns on quantifying population exposure to air pollution. *Environmental Science & Technology*, **50**, 9671–9681.

Scellato, S., Musolesi, M., Mascolo, C., Latora, V., and Campbell, A. T. (2011). Nextplace: A spatio-temporal prediction framework for pervasive systems. In Lyons, K., Hightower, J., and Huang, E. M., editors, *Pervasive Computing: 9th International Conference, Pervasive 2011, San Francisco, USA, June 12-15, 2011. Proceedings*, pages 152–169. Springer, Berlin, Heidelberg.

Secchi, P., Vantini, S., and Vitelli, V. (2015). Analysis of spatio-temporal mobile phone data: a case study in the metropolitan area of Milan. *Statistical Methods & Applications*, **24**, 279–300.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Chapman & Hall, London, United Kingdom.

Song, C., Qu, Z., Blumm, N., and Barabási, A.-L. (2010). Limits of predictability in human mobility. *Science*, **327**, 1018–1021.

Wand, M. and Jones, M. (1995). *Kernel Smoothing*. Chapman and Hall, London.

Wang, Y. and Chee, C.-S. (2012). Density estimation using non-parametric and semi-parametric mixtures. *Statistical Modelling*, **12**, 67–92.

Wu, X. (2018). Robust likelihood cross-validation for kernel density estimation. *Journal of Business & Economic Statistics*, pages 1–10. doi: 10.1080/07350015.2018.1424633.

Zorn, S., Rose, R., Goetz, A., and Weigel, R. (2010). A novel technique for mobile phone localization for search and rescue applications. In *2010 International Conference on Indoor Positioning and Indoor Navigation*, pages 1–4.