UNIVERSITY OF TARTU
Institute of Computer Science
Computer Science Curriculum

Salijona Dyrmishi

Creating a novel approach for mobile positioning based on CDR data

Master's Thesis (30 ECTS)

Supervisor: Amnir Hadachi, PhD.

Tartu 2019-2020

Creating a novel approach for mobile positioning based on CDR data

Abstract:

User geographical positioning is important for many fields that rely on passive geo-location analytics, like targeted marketing, urban and rural transportation planning, public health, etc. A new popular type of data that is commonly used for passive mobility analysis is mobile data or the so-called Call Detail Records (CDR). The CDR events are stored by mobile operators for the primary purpose of billing. They are generated every time we use SMS, call, or internet services. CDR data events are becoming more frequent due to the lower costs of using mobile services and smartphones becoming a necessary tool in our daily life. However, CDR data has two major drawbacks: temporal and spatial uncertainties. Although the first problem is widely covered by trajectory reconstruction techniques, the second problem still remains challenging. Hence, in this thesis, we propose the usage of a new method based on the Sequential Monte Carlo algorithms called particle filtering. The particle filtering application implemented in this thesis models the trajectory movement to predict the user's position in a given area. This method uses CDR data and solely the information related to the area of the coverage from mobile towers. Our goal is to evaluate if this nonlinear method can out-perform the existent linear methods like Switching Kalman Filter. Therefore, the model performance and the effects of the parameters on accuracy were evaluated in controlled experimental settings. Additionally, experiments were performed on a dataset from a real case study and compared with the results achieved by existing methods. Finally, the usability of the method and future work is discussed.

Keywords: *mobile data, particle filtering, location prediction, trajectory prediction*

CERCS: P170 (Computer science, numerical analysis, systems, control)

Uue lähenemisviisi loomine mobiilse positsioneerimise jaoks CDR-andmete põhjal

Lühikokkuvõte:

Kasutaja geograafiline positsioneerimine on oluline paljudes valdkondades, kus kasutatakse inimeste asukohapõhiseid andmeid, näiteks turunduses, linna- ja maatranspordi kavadamisel, rahvatervise uurimisel jne. Uut tüüpi andmed, mida kasutatakse passiivse liikuvuse analüüsimisel, on mobiilse andmeside kasutamisel salvestunud kirjed (CDR - Call Data Records). Tavaliselt salvestavad mobiilioperaatorid CDR-sissekandeid arvelduse eesmärgil. Neid genereeritakse iga kord, kui kasutame SMS-, kõne- või Interneti-teenuseid. Logitud CDR-andmete maht on järjest kasvanud, kuna mobiilsed teenused on läinud odavamaks ja nutitelefonide kasutamine on muutunud vajalikuks tööriistaks meie igapäevaelus. CDR-andmetel on siiski kaks suurt puudust: ajaline ja ruumiline ebatäpsus. Ehkki esimest probleemi käsitlevad trajektoori rekonstrueerimise tehnikad laialdaselt, on teine probleem endiselt väljakutsuv. Seetõttu teeme käesolevas lõputöös ettepaneku kasutada uut meetodit, mis põhineb järjestikuste Monte Carlo algoritmidel ja mida nimetatakse osakeste filtreerimiseks. Selles lõputöös rakendatud osakeste filtreerimise rakendus modelleerib trajektoori liikumist, et ennustada kasutaja asukohta antud piirkonnas. See meetod kasutab CDR-andmeid ja ainult mobiilside tornide levialaga seotud teavet. Meie eesmärk on hinnata, kas see mittelineaarne meetod suudab ületada olemasolevaid lineaarseid meetodeid nagu näiteks Kalmani filtri kasutamine. Seetõttu hindasime kontrollitud katseseadistustes mudeli jõudlust ja parameetrite mõju täpsusele. Lisaks tehti katseid reaalajas uuringu andmestikuga ja võrreldi olemasolevate meetodite abil saadud tulemustega. Lõpuks arutasime meetodi kasutatavuse ja edasise töö üle.

Võtmesõnad: mobiilne andmeside, osakeste filtreerimine, asukoha ennustamine, trajektoori ennustamine

CERCS: P170: Arvutiteadus, arvutusmeetodid, süsteemid, juhtimine (automaatjuhtimisteooria)

Acknowledgement

First, I would like to express my gratitude to my supervisor Dr. Amnir Hadachi for proposing such an interesting topic, always having a positive attitude and for encouraging me to continue further even when I felt like it was not worth it. I would like to thank Artjom Lind for providing the necessary technical help and always being available for some extra explanations and questions. Additionally, I would like to acknowledge the work of Dr. Amnir Hadachi and Oleg Batrashev on providing a starting point for the code implementation.

Many thanks to my friends whose physical or virtual presence and support has kept me sane during this intensive period. Additionally, I could not finish the writing process in time without daily encouragement from Sri. Thank you for always believing in my academic abilities, proofreading this document, and your love. Most importantly I would like to thank my family, my brother Ledio, and my parents Eqerem and Alime, for always considering my education a priority. I always have felt their genuine interest through questions related to my university and work projects. They have supported me unconditionally in every decision and without their sacrifices I would not have achieved this milestone.

Contents

1 Introduction

Geo-localization as a service is the base of many applications in today's digital world. Active mobile positioning is important in order to serve customers in a timely manner but as well in critical cases like emergency calls. The source of localization data is different mainly to different technologies used (GSM, CDMA). Often the choice of the technologies is a trade-off between the costs and the information gain from the data. Today, most of the geo-localization services base their estimations on GPS data which has high accuracy especially when the line of sight is clear. Additionally, the sample rate starting as low as one event per second makes GPS data really dense.

On the other hand, passive mobile positioning gives access to understanding the spatio-temporal behavior of the users. Lately, this type of data is used in a vast pool of fields like mining human mobility patterns, urban planning, tourism estimation, vaccination planning, public transport rescheduling, etc. The reason for the popularity in so many fields is mainly because passive mobile positioning serves as a great replacement for traditional methods that build models upon samples rather than data. The common passive mobile positioning data are the logs from Mobile Network Operators (MNO). Every time we use the mobile phone to make a phone call, send an SMS, or use 4G/5G internet packages they store what is called Call Detail Records (CDR) for the primary purpose of billing. Nowadays, there is no doubt that the mobile penetration rate is increasing even in the most rural/ unreachable places. Additionally, the cost reduction of telecom services has contributed to a wider geographical spread of the CDR data, as well as an increase in its frequency. This situation is promising for an increase in importance for CDR data and probably an expansion to other fields.

1.1 Problem statement

GPS data is highly desirable but it is not the most appropriate choice for passive analyses mainly due to the limitations related to its availability. It is not possible to have GPS records for every mobile phone user, simply because many are concerned about privacy and they are not comfortable sharing their location. Additionally, the GPS-enabled receiver performs complex calculations, and their processing power requirements drains the battery of mobile devices really fast. On the contrary, CDR data is easily available considering the fact that the Mobile Network Operators (MNO) always store them and can be retrieved for all users of a network. This feature makes CDR highly desirable for passive spatio-temporal analysis, however, they have some disadvantages.

The first disadvantage comes from the fact that the radius of the tower coverage introduces

difficulties for the exact user positioning. Every CDR record is coupled with the coverage area of the tower where the user is connected. These areas, called cells, do not have unique shapes. They vary from the height of the tower location, the urban environment around the tower, population density, the obstacles, etc. The radius of the cell ranges from several meters in urban areas to tens of kilometers in rural areas. Within this radius, the location of the mobile phone user is actually unknown. The second disadvantage is related to temporal uncertainties. Being that CDR data is not collected in constant predefined intervals, the gaps between two consecutive events can be really considerable. They might start from one minute to several hours. Unfortunately, the last one is a common case. The high level of spatial and temporal uncertainties can have a negative effect on the reliability of the studies related to human mobility patterns.

In order to reduce the uncertainties, it is necessary to apply models that will firstly pre-process CDR data like eliminating undesirable effects related to network functionalities. Secondly, it is necessary to reduce temporal uncertainties by building models that will fill in the gaps between consecutive events. Thirdly, estimating the user location within the cell in order to reduce spatial uncertainties. From the study of related literature, we have noticed that the majority of studies apply the first technique before performing aggregated analysis in fields like tourism, healthcare, transport planning, etc. Fewer studies deal with trajectory reconstruction for the gaps. Lastly, only a couple of studies tries to localize the users within the coverage areas of the towers.

1.2 Contributions

This thesis aims to introduce a new non-linear technique, based on Sequential Monte Carlo Methods named Particle Filtering, to predict the position of the mobile user within a given area. It is in this research's interest to evaluate if mobile positioning can occur with decent accuracy using only the minimum required CDR information which includes the timestamp of the generated event and the coverage area of the tower. The main contribution of this thesis will be exploring if it is possible by using a nonlinear approach to enhance the mobile positing compared to the existing linear approaches.

We will try to answer this question following the steps as below:

- Propose an application of Particle Filtering to estimate the mobile user position using CDR data.
- Systematically evaluate the proposed model in experimental settings and real-world data.

- Compare the non-linear Particle Filtering method with existing linear methods.

This is a challenging task and not a largely explored area. This work is among the few that study the problem of the user positioning using only the minimum required information from CDR data. Additionally, previous studies consider only linear approaches. We recognize that human mobility does not follow a linear model, therefore, we are proposing Particle Filtering. Improvements in user positioning will reduce CDR uncertainties and can bring CDR data closer to GPS data regarding accuracy.

1.3 Road-map

The rest of this thesis is organized as follows:

Chapter 2 (Background): This chapter introduces the user with all the necessary background terms, history, and context to understand the source of the problem. In addition, it gives an overview of already present mobile data-related research.

Chapter 3 (Methodology): This chapter presents the basic concepts of the particle filtering as an introduction to the work done for this thesis. Afterward, it demonstrates step by step the adaption of the particle filter for user positioning based on CDR data.

Chapter 4 (Experiments): This chapter describes the experimental setup and the datasets used to evaluate the model. It is followed by a short description of the evaluation metrics. At the end of every experiment done, it displays the outcomes and results in the form of tables and graphs.

Chapter 4: (Discussions) In this chapter we will discuss the outcomes of the experiment and the main lessons we have derived. Based on these discussions a road-map for future work will be provided.

2 Background

This chapter will provide the necessary information for understanding the scope and the terminologies of this project. The first section will describe how the mobile network is functioning. It is necessary to understand its components and interaction with mobile terminals in order to understand the challenges and opportunities that arise while using mobile data for user localization. The second part describes the structure of the logs collected by mobile operators. And lastly, the third section gives an overview of the current research that uses CDR data.

2.1 Global System for Mobile communication

Almost all Mobile Network Operators (MNO) in the world use two main mobile technologies: Global System for Mobiles(GSM) and Code Division Multiple Access (CDMA). GSM technology was first launched in Finland in 1991. It was developed by European Telecommunications Standards Institute (ETSI) to describe the protocols for 2G mobile communications. Presently, GSM comprises approximately 90% of mobile connections worldwide, being the most widely used digital mobile telephony system nowadays [1]. Therefore, in our study, we will concentrate only on understanding the functionality of GSM networks.

Definition 1. *The global system for mobile communication (GSM) is a globally accepted standard for digital cellular communication. GSM is the name of a standardization group established in 1982 to create a common European mobile telephone standard that would formulate specifications for a pan-European mobile cellular radio system operating at 900 MHz [2].*

2.1.1 GSM network components

The GSM network is based on four separate components: the mobile device itself, the base station subsystem (BSS), the network switching subsystem (NSS), and the support subsystem (OSS) [2]. The Subscriber Identity Module (SIM) card is tied to a specific network within GSM network. Every part of the infrastructure of this sub-network makes sure to serve to the device where the SIM card is integrated. The functionalities of each component are described below:

- Mobile device

  The mobile device connects to the network via hardware. The SIM card provides the network with identifying information about the mobile user.

- Base station subsystem (BSS)

  The BSS handles traffic between the cell phone and the NSS. It consists of two main components: the base transceiver station (BTS) and the base station controller (BSC). The BTS contains the equipment that communicates with the mobile phones, largely the radio transmitter-receivers and antennas, while the BSC, is the intelligence behind it. The BSC communicates with and controls a group of BTSs.

- Network switching subsystem (NSS)

  The NSS portion of the GSM network architecture, often called the core network, tracks the location of callers to enable the delivery of cellular services. The NSS is owned by mobile carriers.

- Support subsystem (OSS)

  The OSS is the functional entity from which the network operator monitors and controls the system. An important function of OSS is to provide a network overview and support the maintenance activities of different operation and maintenance organizations.

The GSM network is made up of geographic areas. As shown in Figure 1, every BTS has a radius of coverage represented by the hexagons. In the MNO terminology, they are identified as cells. Several cells together are part of a Local Area Connector (LACs).

Definition 2. *A cell is the area given radio coverage by one BTS.*

The GSM network identifies each cell via a unique assigned number named cell global identity (CGI). The cell size can vary from less than 1 km^2 to several km^2. In areas with dense population, the cell size tends to be smaller compared to rural areas. There are four different types of cells: pico, micro, macro, and umbrella. The coverage area for each type depends on their location environment and the height where they are positioned. Typically, pico cells are used mostly indoor with a radius of 4 - 200 meters. Micro cells are usually used in urban areas with coverage between 200 m and 2 km. They serve to malls or other important hubs. Macro cells make about 90% of GSM networks and their radius varies somewhere from 1 km to 30 km. Umbrella cells cover the gaps between several micro cells and serve often as a solution for network overloading. In reality, the shape of the cells does not follow the regular hexagon pattern that we illustrated in Figure 1. Additionally, there are often overlaps between two cells and sometimes there are gaps. [3][4]

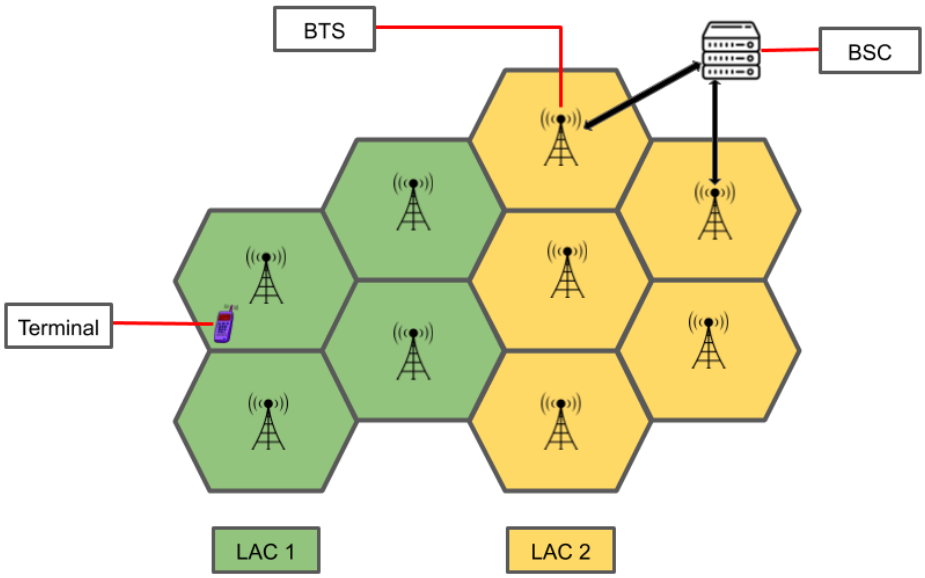


Figure 1. Overview of the mobile network

2.1.2 GSM cell localization

At an initial stage when a terminal (mobile device) is switched on, it goes into IDLE mode. When the terminal is in IDLE mode it will connect to the BTS with the strongest signal. Due to noise, this is not necessary the nearest base station. After the initial connection is established, the connection switches to the Dedicated mode where everything is synchronized and information can be exchanged. When the terminal is in Dedicated mode the GSM network algorithm decides about handover between cells. The decision is based on the load of the cells and the motion pattern of the terminal. The goal is to serve the terminal with high quality but as well to equally distribute the load of the BTSs. Due to this algorithm, the selected cell might not always be the nearest. When the handover happens frequently in a couple of seconds, we are dealing with the ping pong effect. In reality, the user has not moved but due to the ping pong effect, the CDR data reflects back and forth movements. When moving between different cells, a terminal sends a location update only when it enters a cell that is not in the previous LAC. [3]

2.2 Call Detail Records (CDR)

Today 66.77% of the world's population has a mobile device, which equals to 5.17 billion people. In 2017, the percentage of people owning at least one mobile device was 53%. And by 2023, the prediction is that the number of mobile device users will increase to 7.33 billion [5]. The percentage is considerably high considering that a part of the population by default can not have a mobile device i.e. underage children or people in remote areas with no access to electricity. With the increasing number of mobile users and the time we spend using our devices, there is an expected increase on the scale and frequency of mobile data as well. Mobile data is considered as an interchangeable term with CDR which are logs generated by the activity of using mobile devices.

Fiadino et al. [6] have taken under study CDR datasets describing the activity of customers from all mobile operators in Spain, from 2014 and 2016. Their comparison of the data quality showed that the number of daily data connection has increased in 2016 from 10.9 to 50.1 and overall the daily action count grew by 4 times. The data connection customer share had the largest growth from 50% to 85%. Several factors were considered like Days of Visibility, Hourly Action Rate, Average Lag Time, Total Inactive Time, and Entropy on this study. The authors revealed that all of them had marked improvements from 2014 to 2016. This study is a good indicator that the user patterns have changed and the CDR data is becoming more fine-grained for mobility studies.

Telecommunication companies store mobile data logs or CDR mainly for billing purposes. The term might be misleading because it dates from the time when the only service you could use were mobile calls but nowadays they are triggered on the event of calls, sending/receiving SMS, and using Internet services. In particular cases, the configuration allows the generation of mobile logs additionally every time the user switches cells or LAC.

Typically, a CDR record has the following fields[7]:

- **Unique sequence number**: The Sequence number identifying the record
- **Calling party**: The phone number of the caller
- **Called party**: The phone number of the receiver
- **Billed number**: The billing phone number that is charged for the call
- **Call duration**: The duration of the call in minutes
- **Stat time**: The starting time of the call(date and time)
- **Call type**: The type of call that was made(VoIP (Voice over Internet Protocol), voice, or raw data)
- **Cell global identity (CGI)**: Unique identifier for each cell

The minimum required information for geo-spatial analysis are the timestamp and the CGI paired with the cell coverage area. For mobile positioning, the geographical shape of each cell plays an important role. The shape depends on factors such as antenna radiation pattern and height, network load, signal attenuation on the landscape and indoors, signal reflections, radio interference and noise, network configuration parameters etc [8].

2.3 CDR based localization and human mobility analytics: A literature review

CDR data does not seem to be an unknown source for researchers. With the increase of its availability, the CDR is becoming complementary data for many passive mobility analyses, especially in humanities studies. And lately, the application fields are expanding. However, we have found really few studies which focus on actual user positioning and path reconstruction.

We can separate the research focusing on CDR in two major groups:

- Context aware analysis
- Trajectory reconstruction and localisation

The largest volume of studies uses methods from the first group. The range of applications includes tourism statistics, urban planning, human mobility patterns, disaster recovery, invective disease epidemics, commuting patterns, etc. For example authors in [9] use rule-based filters into the database of CDR records to detect tourists from non-tourists based on the last five user locations. Similarly, Saluveer et al. [10] filtered CDR data in order to separate tourist and non-tourists, producing country-wise tourism statistics. Another study from Qin et al. [11] modeled tourist flows only in scenic touristic areas. The authors in [12] built an approach on estimating static population densities based on aggregated mobile network traffic metadata. Another important application is building Origin-Destination (OD) matrices which try to quantify the movements between an origin and destination zone in different aggregation levels. Traditionally travel surveys have been used for this purpose. However, they require high cost and do not generalize enough due to the low sample size. Other methods include using cameras in the entrance and exit points of road segments but this is not a scalable solution. Friedrich et al.[13] proposed a new method of generating OD matrices based on floating mobile data (CDR). This method is more efficient compared to the traditional method of generating OD using a travel demand model. They apply counting to each link compromised of origin and

destination point and afterward perform clustering in order to build OD matrices. [14] Pourmoradnasseri et al. uses a similar technique but apply a trajectory reconstruction step before generating OD matrix.

In the second group, there is research done to deal with the uncertainty of mobile phone data. Firstly, the researchers have tried to deal with sparsity. The most common technique to do so is to fill the gaps with new cell IDs based on trajectory reconstruction. Trajectory reconstruction aims to infer information from aggregated user movements and then fill the missing values for unknown individual user position. [15] Hoteit et al. categorized mobile phone users in 4 categories: sedentary people, urban mobile people, peri-urban mobile people, and commuters by analyzing the cumulative distribution of the radius of gyration. They select users with more than 1000 data points during a given day and sub-sample from those trajectories to produce normal user behaviors. On the sub-sampled data, they use three classical interpolation techniques to do trajectory reconstruction. Finally, they evaluate their methods by comparing them with data before sub-sampling. In their study, the authors are not considering the cell surface but rather the BTS longitude and latitude.

Another tool to complete individual CDR-based trajectories is proposed by Chen et al. in [16] based on tensor factorization. On the first iteration, the authors consider the users whose home locations are known with 80% confidence. For these users, they fill in the night gaps depending on time granularity. In a second step, the data is organized in 3D tensors which represent the daily, weekly, and instantaneous mobility of the user. They use the tensor factorization method, which has been useful in other contexts, to fill in the tensors which on average were only 0.63% complete. Tensor factorization decomposes the tensor in hyperparameters and then uses them to generate new data to fill the voids. Lastly, the positions generated are mapped back to the coordinates of the closest BTS. Their results show that they can locate users with a median displacement between 1 and 2 network cells, meaning the user location is predicted on the exact cell or in the cell next to it. In another study[17], the authors worked on trajectory reconstruction based on the path edge level. They select for every two consecutive events in the CDR trajectory a random point from the uniform distribution and map it to the closest road segment. The fastest path between two points is considered as a possible trajectory. We can consider the random points as a location estimation within the cell.

The last domain of the second group, the least explored one, is the reduction of spatial uncertainties by estimating the position of the user within the cell-plan. The only study we have found in this area is from Lind et al. [18] firstly published in 2017 and later extended in a second version in Hadachi et Lind [19] where the authors use a version of Kalman Filter called the Switching Kalman filter with smoothing. The model tries to extract the movement patterns from the data-driven exploration phase and label each

record as a Stay or Jump position. Simultaneously the authors estimated the user position within the cell and mapped it to the closest road segment or building. The results were quite promising especially for the stay model with an RMSE of 2106.8 meters. However, the Switching Kalman Filter falls into the category of linear models.

Regarding the usage of non-linear Monte Carlo Sequential models for positioning, like particle filtering, the research dates as early as the 2000s but applied to other domains. For example in [20] the authors proposed to use particle filtering for car positioning. The measurement is taken from wheel speed sensors in the car and the velocity vector is considered as the input signal. The authors argued that it is possible to use mobile phone measurement data as an input vector. In another study by Hu et al. [21], the authors are using a modified version of particle filtering to localize moving nodes in a range-free wireless network. Their search space is divided into equal rectangles. As part of their work, they analyze the effects of sample size and motion control into accuracy, but they do not compare different sizes of the coverage. The technique was tested on a simulated environment and resulted to improve the accuracy of localization compared to the case where the mobility information is not taken under consideration. Later Dil et al. [22] used the same model, as the previous authors, for range-based wireless networks. Similarly, they used fixed-size areas and evaluated the effects of the sample size and speed. The reported localization improvement ranged from 12% to 49%.

3 Methodology

This section will give an overview of the Particle Filtering method. Additionally, we will describe step by step the proposed adaptation for the task of user localization.

It is common for us to try to quantify and model certain phenomena in the world. However, we need to take into account the noise level related to measurements that introduce uncertainties to our model. User positioning can be considered as one of these problems. The signal is the Cell ID plus the timestamp and the noise is related to effects like ping-pong or spatial uncertainty. On Bayesian modeling, we estimate how likely is each observation given knowledge of prior distribution. And as more data flows in, we have to update our posterior distributions. The literature describes three methods of modeling natural phenomena according to Bayesian statistical inference. Firstly, when the data is modeled according to a linear model in a state-space environment and has a Gaussian noise then it is possible to use Kalman Filters. Kalman filters try to estimate the state via posterior distribution and update their beliefs after receiving new measurements. The main steps in a Kalman Filter are performed via matrix multiplication. In other cases when the data is modeled as discrete states it is possible to build estimations using Hidden Markov Models. When none of these restrictions apply i.e. the noise is not normally distributed or the measurements are not linear then the above solutions do not generalize well.

To solve this issue in 1993 Gordon et al.[23] introduced what they called a bootstrap filter, extended today to Particle Filter. Particle Filters fall into the category of Monte Carlo algorithms and provide approximate solutions. They are used for dynamic systems modeled by a Bayesian network where the states are observed only partially. It can be used as well into linear models with normally distributed noise but the Kalman filter is most preferred on these cases due to its simplicity compared to Particle Filter. The most common applications of Particle Filtering are tracking, SLAM, localization, etc. In this thesis, we recognize the fact that humans do not move linearly and the noise in CDR trajectories is not normally distributed. Therefore, we are taking understudy the application of Particle Filtering for user positioning using CDR data. Before presenting the adapted application we are going to present in the next session the theory behind particle filters.

3.1 Particle filtering

The idea behind this method is to approximate the posterior density or the belief state using a weighted set of particles sampled from the entire space of the state trajectories:

$$p(z_{1:t}|y_{1:t}) \approx \sum_{s=1}^{S} \hat{w}_t^s \delta_{z_{1:t}^s}(z_{1:t}) \tag{1}$$

In the formula above $y_{1:t}$ represents the signal during t timestamps and $z_{1:t}$ are the drawn samples from our proposal distribution. Meanwhile, $\hat{w}_t^s$ is the normalized weight of sample s at time t and $\delta_{z_{1:t}^s}$ is the Dirac function [24].

You will find below in Figure 2, a visual representation of the main steps of the algorithm described in the rest of this section.

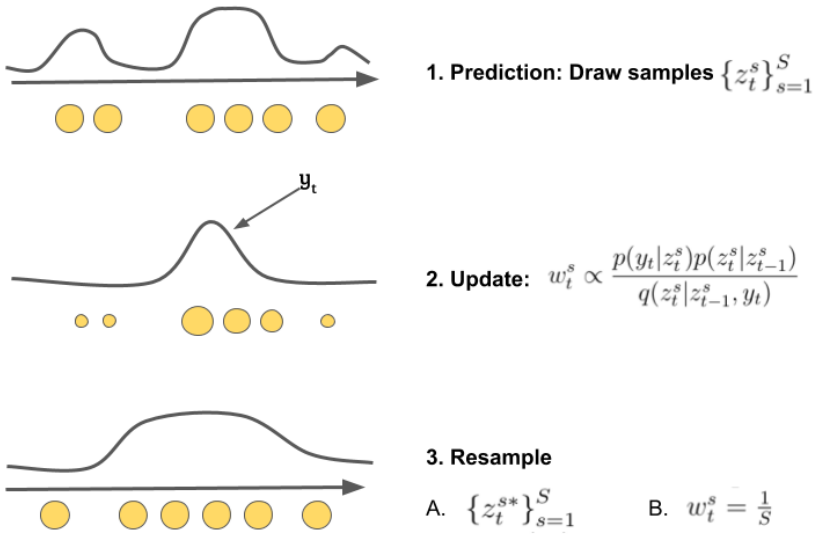


Figure 2. Particle filter schematic view

To achieve the approximation in equation (1) we can ignore the previous parts of the trajectory and compute the marginal distribution over the most recent state $p(z_t|y_{1:t})$. From the proposal distribution $q(z_{1:t}^s|y_{1:t}) = q(z_t|z_{1:t-1}, y_{1:t})q(z_{1:t-1}|y_{1:t-1})$ we draw S samples at timestamp t. This is the first step or prediction phase of the particle filtering represented in Figure 2. Each of this samples get an importance weight equal to:

$$w_t^s = w_{t-1}^s \frac{p(y_t|z_t^s)p(z_t^s|z_{t-1}^s)}{q(z_t^s|z_{1:t-1}^s, y_{1:t})} \tag{2}$$

The weights need to be normalized as follows:

$$\hat{w}_t^s = \frac{w_t^s}{\sum_{s'} w_t^{s'}} \tag{3}$$

If we make the assumption that $q(z_t^s|z_{1:t-1}^s, y_{1:t}) = q(z_t|z_{t-1}, y_t)$, then we do not need the history of the trajectory in our calculations. We can only consider the previous step observation to compute the new distribution. Now, the weight for each sampled particle is calculated as:

$$w_t^s \propto \frac{p(y_t|z_t^s)p(z_t^s|z_{t-1}^s)}{q(z_t^s|z_{t-1}^s, y_t)} \tag{4}$$

From there we can approximate the posterior filtered density using the formula:

$$p(z_t|y_{1:t}) \approx \sum_{s=1}^{S} \hat{w}_t^s \delta_{z_t^s}(z_t) \tag{5}$$

This is represented in Figure 2 as step number 2, the update.

Until this step, the algorithm has been named Sequential Importance Sampling (SIS). Unfortunately, it suffers from the degeneracy problem as after some steps only a few of the particles will have a considerable weight. For the rest of them, the weight decreases with a tendency to go towards zero. When the variance of the weights is large we are updating weights that do not actually contribute to our posterior distribution. One solution to the degeneracy problem is to introduce a re-sampling step.

The idea behind the resampling step is to eliminate the particles with low weight and replace them with duplicates of particles that have higher importance. We generate a new set of particles $\{z_t^{s*}\}_{s=1}^S$ by sampling with replacement S times from the weighted distribution in 5. The probability of choosing particle j for replication is w_t^j. The new samples are unweighted, so we need to set their weights to $w_t^s = \frac{1}{S}$. This is step number 3, called in Figure 2 as resample.

3.2 Particle filtering for mobile user localization

The problem of the user localization has been solved until now using linear Bayesian methods like Switching Kalman Filter. Considering that the users do not move linearly we adapted a nonlinear method like particle filtering to solve the same problem. This section will give an overview of our adaptation of the particle filter. The main steps of the workflow are represented in the flowchart in the Figure 3.

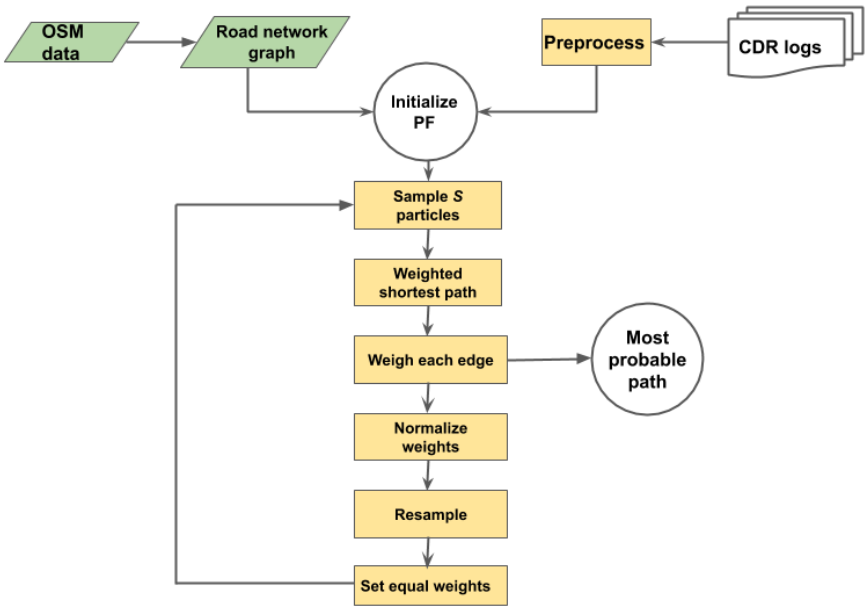


Figure 3. Mobile positioning flow chart

Preparing the data

The algorithm starts by receiving mobile data events as input. Every record in the trace should have at least the timestamp and the Cell ID. The Cell ID is coupled with the information about the coverage area of the specific cell. This information is expressed as a list of coordinates (latitude and longitude) that represent the corner points of the cell polygon. Any desired preprocessing on the trajectory should happen at this stage. For example, it is beneficial to detect ping-pong effects, to detect long gaps between consecutive events, etc. We will speak on more details about this step in Chapter 4.

As the second parameter for the input serves the geographic area where the CDR records are spread. We are using the OpenStreetMap (OSM) API to download the data in XML format. On a second step, the data is read from the file and converted to a graph structure

where each building is represented as a node. If two nodes are connected to each other in any way, then an edge $e(u, v)$ is added to the graph where u marks the starting point expressed in geo-coordinates and v the endpoint. OpenStreetMap is an open-source project and relies on the contribution of its members for the completeness of its data. The edge element in OSM output has a tag for maximum speed allowed but that is not the case for every edge. When this tag does exist, we add the maximum speed as a property to the edges of our newly created graph. The rest of the time, when the maximum speed limit is missing we add a constant value. This value is calculated based on the average traffic characteristics of the geographical area where the movements of our users are spread.

The third parameter that the algorithm receives is the necessary number of particles S, to model the distribution of the user location. After we have specified all these three inputs we are ready to start with the next phase which is initialization.

Initialization

At the initialization stage the algorithm randomly selects S edges in the complete graph from the uniform distribution, Each edge endpoint is assigned a weight of $\frac{1}{S}$ as a possible location of the user. At this step, we do not have any prior knowledge on the user location therefore there are no restrictions on the graph. The user is equally probable to have started his journey at any point. The time parameter at the initialization phase is set as well to 0. In Figure 4 we are going to visually demonstrate the stages of the algorithm with an example which uses three particles. As discussed previously the hexagons represent a cell and the grid with squares represents the road network. Until now, we have presented only Step 1: Initialization. The red segments show the randomly selected edges. Notice that the line width is proportional to the probability of every edge, $\frac{1}{S}$.

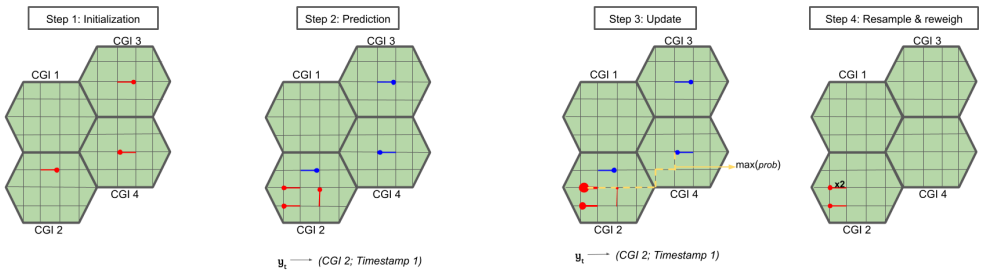


Figure 4. User positioning using mobile data

The initialization process is the first step of the algorithm and happens only once in the beginning. The next steps we will present below, from update to resampling, are iterative and happen for every record in the CDR trajectory.

Update

The algorithm selects all the edges which have both endpoints in the given cell. Assuming a uniform distribution of the user location within the cell, we select randomly S edges from this group. In Figure 4, this process is represented as Step 2: Prediction. The road segments marked with red are the predictions based on the belief and the road segments are the previous state. Until this point the weights are equal. From the selected S edges, the endpoints are taken into consideration. The weighted shortest paths between every edge endpoint from the previous step and the new proposed edges endpoints are calculated. The results have the form of a matrix with a size SxS. The complexity of this step increases the time requirements, therefore it should be computed in parallel. As the weight for the edges, while calculating the shortest path, serves the estimated travel time which is calculated as the ratio between the endpoints distance and the maximum speed limit. For every path, we need to estimate how likely the user has traversed them given the actual traversal time. The actual traversal time, we note it with TT, is calculated as the difference between the current CDR record timestamp and the previous record timestamp. The probabilities of traversal for every path are estimated using the formula below:

$$probs = \frac{abs(TT - TD)}{TT} \tag{6}$$

In this formula, TT stands for the actual time difference between the CDR records, and TD stands for the estimated travel time using the shortest path algorithm. The probabilities are then scaled to a range from 0 to 1. Given the SxS matrix with all the probabilities, we select the highest one. The associated path that produced this probability is assumed to be the trajectory of the user. The selected path with the highest probability is shown in Figure 4 Step 3: Update with yellow color. The line size and the size of the circles which represent the road segment end are proportional to the updated. The endpoint of this path is the estimated location of the user within the cell.

Now we need to update the weights related to the possible user locations. We consider for that purpose the edge endpoints. There are two components that decide the weights. Firstly, we take into account the transition probabilities. For every new proposed vertex, its maximum probability from the matrix of probabilities SxS is retrieved. As a result, we have a vector with S probabilities, each one assigned to one edge endpoint or vertex.

The second factor that defines the weights is what we call the evidence probabilities. In a GSM network, it is more probable that you will trigger an event if you are closer to the cell center due to signal strength. To take into account this effect we calculate the evidence probabilities which are the probabilities of triggering an event and depend on the distance of the proposed vertex towards the cell centroid. The evidence probabilities and the transition probabilities are then multiplied together. The weights from previous steps that are all $\frac{1}{S}$ are multiplied to these probabilities and the result is normalized.

Resampling

A resampling step happens right after where the vertices with the strongest weight are chosen to be duplicated and their weights are assigned to 1/S. In Figure 4 this step is represented by Step 4: Resample & reweigh.

You will find all the steps described above, with the details of the implantation, in the pseudo code of the Algorithm 1.

Algorithm 1: Particle Filtering for user localization

Input: CDR records with length N including the cell plan for each record,
$G(V, E)$ - graph of road segments and connecting nodes,
S - number of particles

Result: N estimated user positions and $N - 1$ paths

1. Initialization
t_{-1} = 0
Sample e^s_{-1} from uniform distribution
$w^s_{-1} = 1/S$ $\quad \forall s \in 1, 2...S$
$V_{start} = v_1, v_2...v_S$ $\quad \forall e^s_{-1} = (u_s, v_s)$

for $i = 0 \rightarrow N - 1$ **do**

2. Calculate path probabilities
t_i = CDR[i].Timestamp
V' = $v_1, ...v_K$ in CDR[i].Cellplan
Get E'= $e_1, ...e_K$ $\quad \forall e_k = (u, v)$ where $u \in V$ and $v \in V$
Sample e^s_i from E'
V_{end} =$v_1, v_2...v_S$ $\quad \forall e^s_{-1} = (u_s, v_s)$
TT = $t_i - t_{i-1}$;
TD[m,n] = shortest_weighted_path(V_{start}[m], V_{end}[n])
$probs = \frac{abs(TT-TD)}{TT}$
$probs = 1 - probs/probs.max()$
$trans\_probs = probs.max(axis = 0)$
$evid\_probs$ = evidence_probabilities (V_{end}, CDR[i].$Cellplan.centroid$)

3. Calculate weights
$probs = trans\_probs * evid\_probs$
$w^s_t = w^s_{t-1} * probs$
Normalize weights $w^s_t = \frac{w^s_t}{\sum_{s=1}^{s=S} w^s_t}$

4. Resample
Resample

4 Experiments

This chapter will describe all the experiments done within the scope of this thesis and report the results in the form of tables, figures, or graphs.

Acquiring real CDR data requires great effort and collaboration with MNO-s. Generally, it is possible for an individual to retrieve his personal record. However, in this case, we have not removed all constraints as the general requests do not give the individuals the right to access information related to the cell borders. Only the MNO-s can provide this kind of information. For this thesis, we received access to only a couple of hundreds of records together with the cell shapes from one operator in Estonia. In order to overcome the limitations of a small dataset we firstly decided to evaluate the method in synthetic data. Additionally, by creating an experimental set up we will have almost full control over the parameters of the system. We can evaluate the effect that each parameter has on the outcome of the experiments. At the same time, during this process, we could see if some changes are necessary to help improve performance. The synthetic CDR data are generated taking into consideration public GPS trajectories. The prediction error after running the algorithm is evaluated as in subsection "User positioning evaluation". There were three parameters that we explored the most. Firstly, we tried to estimate the effect of the particles on accuracy improvement. Later on, we analyzed the effect of the cell surface and time granularity.

The second set of experiments is performed over real data. Here we keep constant the number of particles, meanwhile, the cell surface and time granularity are determined by the dataset itself. Being that for this dataset we have the full GPS trace we add one extra step on the evaluation, path accuracy. You will find the description of the metric used in this case in the subsection "Path evaluation". For the real data, it is necessary that the trajectories are checked for any inconsistency that would affect the particle filter. We had such a case with long gaps between two records. The particle filtering bases its probability calculation in the traversal time TT and estimated time TD. In cases when the TT is considerably large compared to all TD-s, the algorithm will always define the longest path as the most probable. In order to avoid this bias, we take two measures. Firstly, if the time difference between two consecutive records is more than three hours we consider the trip to be finished. At this moment we interrupt the particle filtering algorithm and restart it from the beginning using the new trajectory that starts from the second record. The second step we perform is calculating the time that is needed to traverse the most distant points between two cells. If this time is lower than the difference between two timestamps it means the user has stayed for a long time in one of the cells without any activity. In this case, we calculate TT as the time to traverse the path between two centroids of the cells to avoid the bias towards the longest path.

4.1 Evaluation metrics

For our system, the GPS locations serve as the ground truth. Our main intention is to position the user as close to the GPS position as possible. As an output from our algorithm, for every CDR record, we get a possible trajectory comprised of the most probable road segment that the user has followed from the previous to the present timestamp. The end node of this path is our predicted user location. In order to evaluate the accuracy of our algorithm, we have to compare this location with the given GPS position. Additionally, as an extra step, we can evaluate if the predicted path is similar to the actual traversed path. Below we are going to show the evaluation metrics for both cases.

User positioning evaluation

To evaluate our model we calculate the haversine distance [25] between our predicted user location for the timestamp and the actual GPS location.

$$a = sin^2(\frac{\Delta\Phi}{2}) + cos(\Phi_1) \cdot cos(\Phi_2) \cdot sin^2(\frac{\Delta\Lambda}{2}) \tag{7}$$

$$d = R \cdot 2 \cdot atan2(\sqrt{a}, \sqrt{1-a}) \tag{8}$$

Where Φ is the latitude, Λ is longitude, R is earth's radius (mean radius = 6,371 km). d in our use case represents the error that the algorithm has made towards the ground truth.

Path evaluation

Given that we are provided with a full GPS trajectory in the time-span between the capture of two CDR records, we can evaluate the accuracy of the proposed path from our algorithm for the user movement. Every GPS point in the trajectory is mapped to the closest edge on the graph. The result is a list of edges that serve as the ground truth. We find the intersection between this list and the list proposed by the algorithm 1. Afterward, the accuracy is calculated as a ratio between the common elements of the list and total elements in the GPS trajectory, expressed as a percentage.

$$E_{gps} = e_1, e_2, ...e_k$$

$$E_{pf} = e_1, e_2, ...e_j$$

$$Acc = \frac{len(E_{gps} \cap E_{pf})}{len(E_{gps})} \cdot 100 \tag{9}$$

4.2 Model evaluation in synthetic data

The first set of experiments were performed in a controlled setting. This section will describe how we generated the dataset, different parameters taken under observation, and the outcomes for each.

4.2.1 Dataset description

The data used in this section belongs to the T-drive [26][27] dataset that contains a one-week GPS trajectory of 10,357 taxis in the city of Beijing. Due to time complexity restrictions of the Particle Filter algorithm we have selected the trajectories of only two taxis. In Figure 5 we can see the distribution of the taxis' locations on top of Beijing city center map. The total 1972 data points are represented with two different colors, which distinguish the two taxis. For the rest of the thesis we are referring to this data as T-Drive dataset, nevertheless, this is only a small sample from the original data.

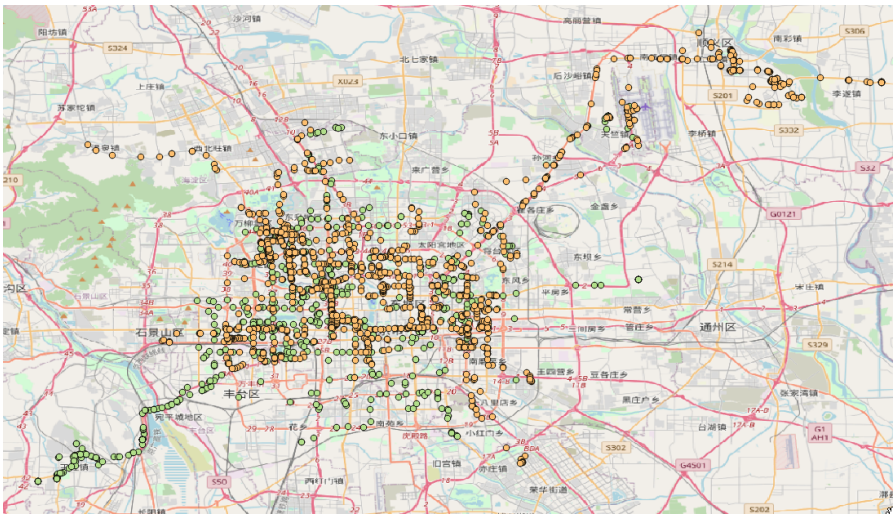


Figure 5. GPS data from T-drive dataset

In Figure 6 we can see the distribution of frequencies for the GPS records. Most of the records are retrieved within less than five minutes from each other. In those 5 minutes, most of the time the taxis have traveled between 0 and 2 km.

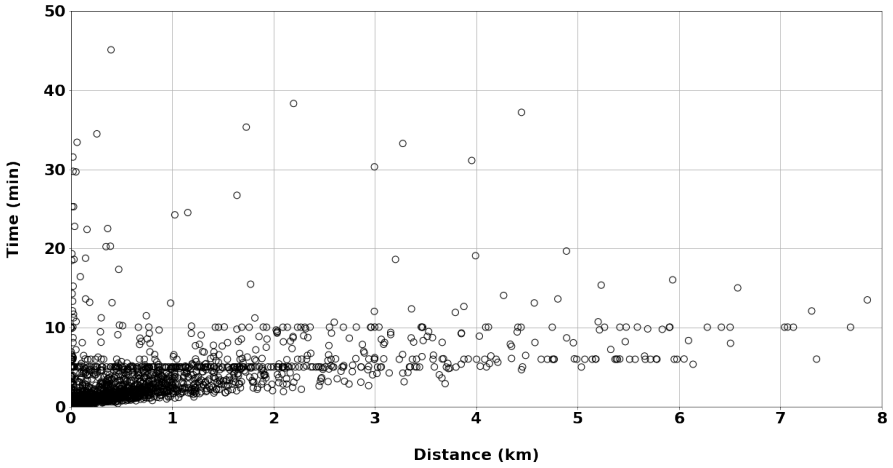


Figure 6. Time and distance relation in T-drive dataset

After the visualization of the dataset, we have to construct our own CDR data. In order to represent the cell coverage from MNOs we have split the geographical area into equal hexagons with diameter 1800 meters. The diameter is representative of most of the cells in urban areas like Beijing city. A unique ID is given to each hexagon. In a second step, every GPS record is assigned to one of these hexagons. The data now has the features *Timestamp*, *Longitude*, *Latitude*, *CGI*, *Cellplan* and it is ready to be retrieved by the algorithm 1. We are making the assumption that when a GPS event is triggered, a CDR event is triggered automatically at the same time. Figure 7 visualizes the spread of the cells in the map together with the GPS points assigned to them.

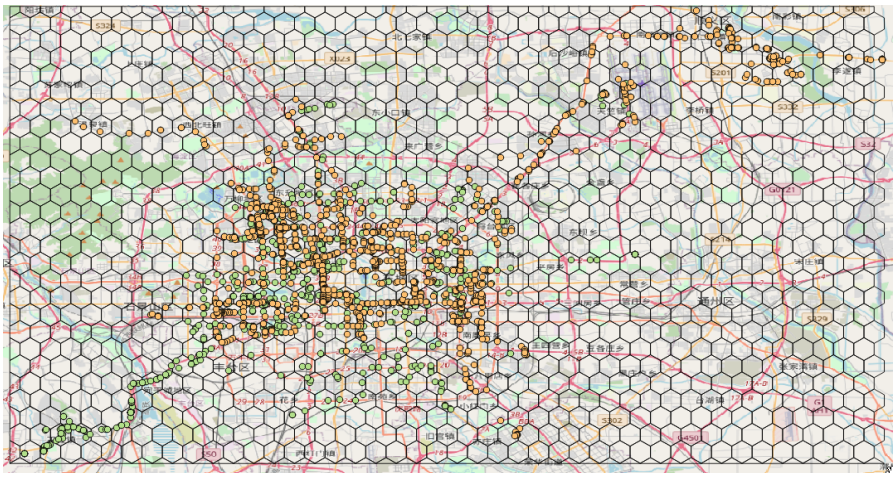


Figure 7. CDR generated events from T-drive dataset

4.2.2 Results

We have selected to initialize the Particle Filter algorithm with 50 particles as a reasonable number considering the small size of the cells. The dataset corresponding to the two taxis locations from T-drive project was fed to the algorithm together with the map extracted from Open Street Map of the Beijing city. The results received are displayed in the histogram in Figure 8. The distribution is a positively skewed distribution with a mean 678 m and a standard deviation of 360 m. There are 81 records or 4.2% of the cases where the error is less than 100 m. Out of this, the minimum error is only 4.8 m. However, we have a maximum error of 1735.7 m as well.

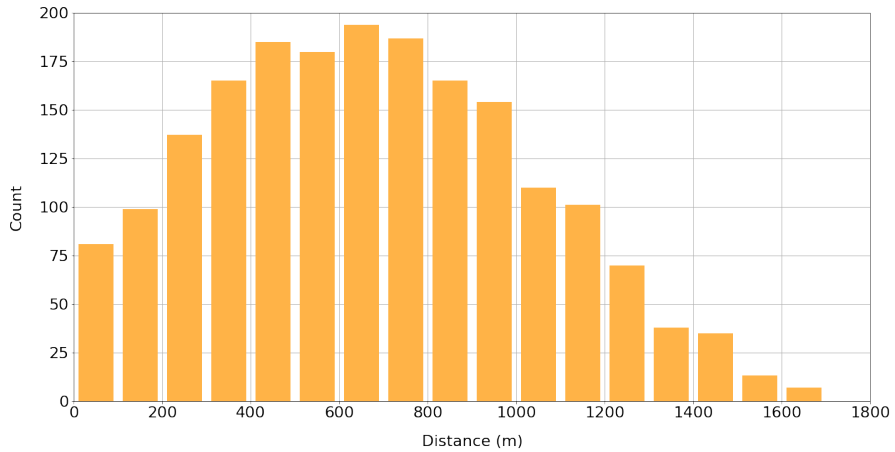
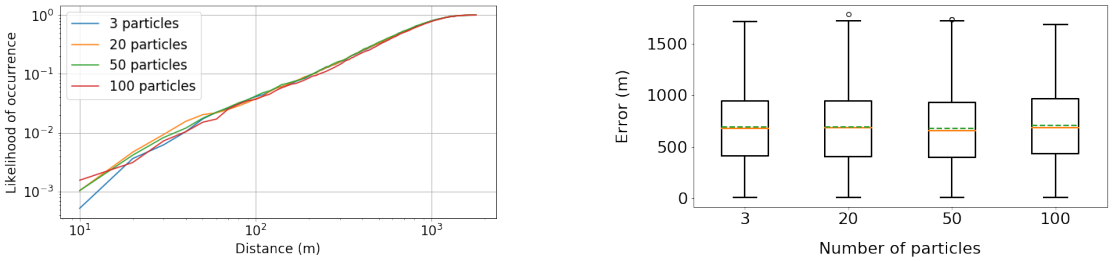


Figure 8. Error distribution using 50 particles in generated CDR events

Sample size
In many particle filter applications, the number of particles used is set in an arbitrary way and kept constant throughout the application. Theoretically, when the number of particles increases near the infinity we are able to model the desired phenomena in a perfect way. Hence, we wanted to study the effect that the number of particles will have in the error for the user positioning. Due to the high time requirements for the execution of the algorithm, we have considered these levels: *3 particles*, *20 particles*, *50 particles*, *100 particles*. The results are presented in Figure 9. When looking at the results we did not notice a considerable difference when the number of particles changes. In 9 (a) we have shown the cumulative distribution of the errors. Both axes are in logarithmic scale to help with the comparison in the visualization. In 9 (b) we are showing the boxplot representation of the errors.



(a) CDF (b) Boxplot

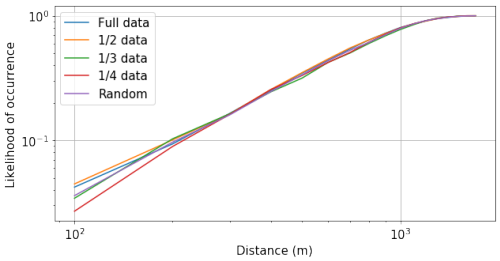
Figure 9. The effect of sample size

Time granularity

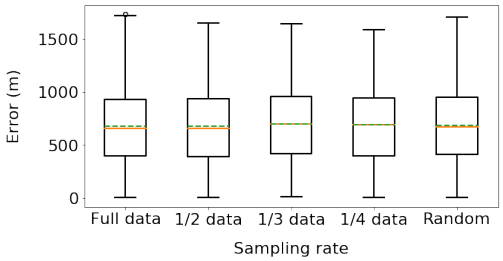
Another factor to consider when dealing with user positioning is the temporal uncertainties which are closely related to the time gaps between the generated CDR data. In this dataset, the time granularity of the CDR data, based on the applied assumption to generate the synthetic data, is the same as with the GPS traces. We already presented it visually in Figure 6. Most of the CDR events in this dataset are generated in less than 5 minutes away from each other. However, nowadays the CDRs are not that frequent yet. Based on the increased usage of mobile devices, maybe in the future, this will be the case. It was in our interest to understand what is the effect that the time granularity has over the resulting errors. To quantify this effect we have used different levels of data sampling. The first sample was generated keeping every second record from the original dataset, resulting in 963 records. As you can imagine most of the data now have a time distance of fewer than 10 minutes. The second sample was generated taking every third record and the later one by keeping every fourth record. In addition to these regulated samples, we generated a random one. A random number (keep probability) between 0 and 1 was drawn from the uniform distribution for every record in the CDR trajectory. If the number is higher than 0.5 we keep the record, otherwise we drop it. The results from our experiments using the newly created datasets with different sampling rates are displayed in Figure 10.

Cell sizes

The third factor we have to take into consideration is the spatial uncertainties. The error distribution for user positioning is limited within the range of the longest diameter of the hexagon. However, these limits vary when it comes to MNO cell towers coverage. We previously mentioned that the diameter ranges from some meters in the urban areas up to 30-40 km in rural areas. Hence, our next will be to study how the change in cell size affects the distribution of our errors. We have tried four different levels of the diameter sizes: *800 m, 1800 m, 2800 m and 3800 m* that are suitable for an urban area like Beijing center. The results of the experiments are shown in Figure 11.

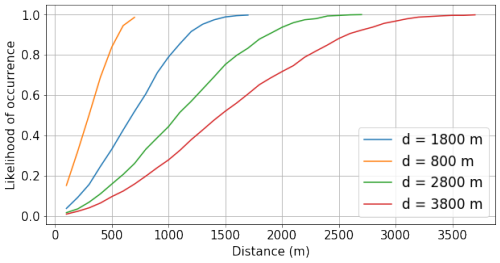


(a) CDF

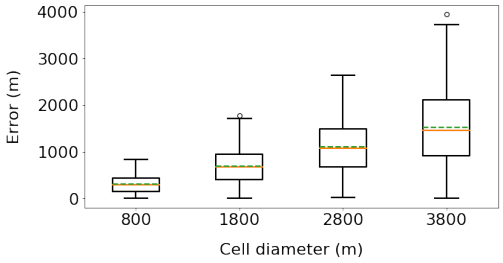


(b) Boxplot

Figure 10. The effect of time granularity



(a) CDF



(b) Boxplot

Figure 11. The effect of cell coverage surface

You will find the statistics about the mean and standard deviation of completed experiments related to time granularity in Table 1. Table 2 contains the statistics about the sample size experiments and finally, Table 3 has the results about the cell size.

Sampling	Mean	Std
Full data	678.4	360.4
1/2 data	676.8	362.9
1/3 data	696.6	361.0
1/4 data	689.9	355.7
Random	687.8	360.5

Table 1. Time granularity

Particles	Mean	Std
5	689.2	359.4
20	689.5	360.5
50	678.4	360.4
100	704.1	358.8

Table 2. Sample size

Diameter	Mean	Std
800 m	305.0	177.4
1800 m	689.5	360.5
2800 m	1105.9	556.5
3800 m	1519.3	772.6

Table 3. Cell size

4.3 Case study: Real CDR data

When it comes to similar applications we already mentioned that there is not enough related work. However, the authors in [19][18], have taken the same problem under the consideration and used the Switching Kalman Filter approach to solve it. We have received access to the dataset they used in their study and the results acquired.

4.3.1 Dataset description

The dataset consists of CDR data of five mobile owners in Estonia, accompanied by the file with their GPS data locations. The period of data extraction is between April and August 2015. This dataset corresponds to a more realistic scenario from the previous dataset, and we can clearly detect that from Figure 12.

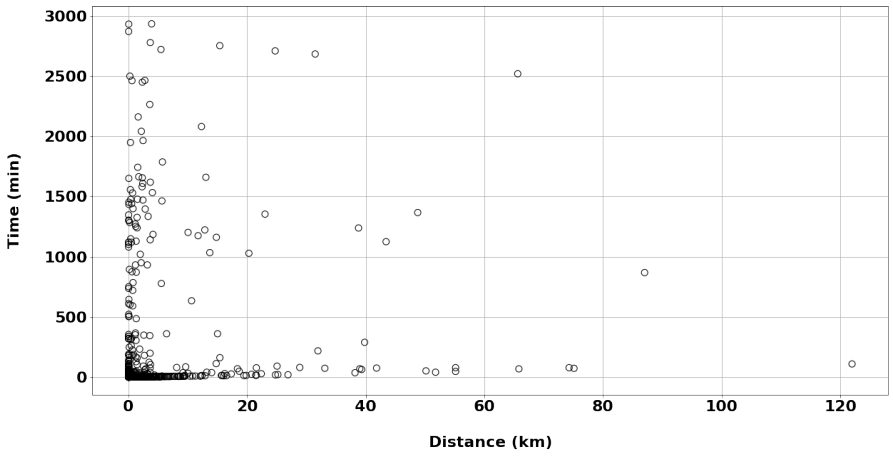


Figure 12. Time and distance relation in real CDR events

Most of the data are captured between every 0 and 500 minutes and the users have moved within that period from 0 to 20 km. It is easy to notice here the lack of information and the challenges with real data we introduced at the beginning of the experiment section. For example, a user who traveled 0 km between two events that have a gap of 2500 minutes can represent two cases. The first one would be that the user has not moved at all from his location i.e. the events are triggered in the evening and later in the morning. In this case, the user has been in his home location for the whole time. However, the second case is more complicated. The user might have traveled from the initial point A to point B and came back to point A. During this period the two events are generated only when starting from point A and after returning to point A. However, we do not have any trace

for the intermediate event. Another challenging point is the users that traveled around 20 km or more in almost 0 minutes. This happens in cases where there are problems with CDR data receiver and the timestamp is corrupted. The newly generated CDR record will be assigned to the last seen timestamp. This figure gives a good overview of the nature of real data. Moreover, some outliers were removed and not shown in the figure in order to have better visualization.

In contrary to our previous example the level of uncertainty related to user positioning in this dataset is significantly increased. In addition, the cell surfaces in real scenarios do not have a perfect hexagonal shape. Their shape resembles more of a distorted polygon or hexagon. One of the geographic information systems, called QGIS, was used to visualize the coverage areas. In many cases, we could notice parts of them overlap with each other. In Figure 13 we can see the distribution of the diameter of the cells in this dataset. Notice here, the diameter calculation is an approximation as the cells do not have regular hexagon or circular shape. We have calculated the surface of each cell and estimated the diameter of the largest circle we could create within this surface. The majority of the cells have a diameter of less than 1 000 m. This is expected as the majority of the movements of these users happen in the city. The rural cells vary from 5 000 to around 25 000 in diameter which makes the user positioning task really challenging in these cases.

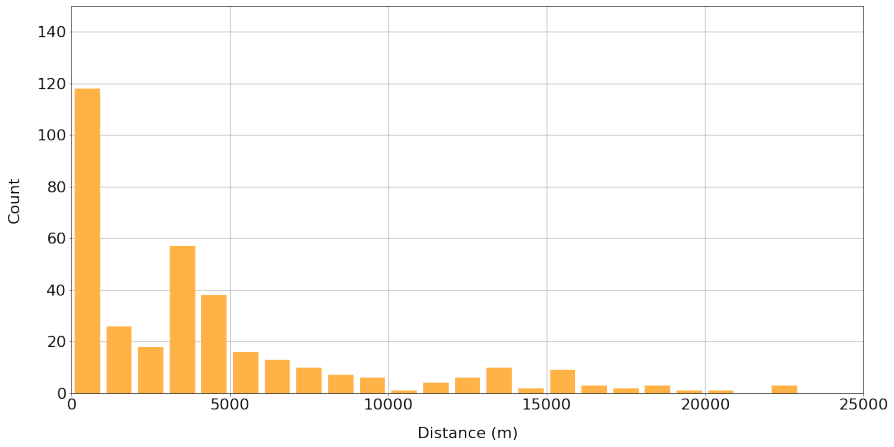


Figure 13. Cell diameter distribution in real CDR events

In total the dataset had 699 records. We have run our algorithm using 20 particles through the complete chain of provided CDR traces. However, we noticed that some GPS records did not fall into the cell area given for the same timestamp. This is normal in the cases where the GPS receiver was not always turned on or it was impossible to collect all the GPS trajectories for the users. In this case, the closest GPS location in time with

the triggered CDR record is selected. This gap can be from several seconds to several minutes and the user might have moved in a considerable distance. Due to the design of the particle filtering for mobile user localization a proper evaluation can be done only when the GPS information matches the CDR. Hence, we have decided to exclude the other cases from the report and show the result only for 359 records.

In addition, for two out of five user IDs we had been provided with complete GPS trajectories of user movements. For these users, we performed the step of path evaluation. You can see the results of these experiments in the next section.

4.3.2 Results

Comparison with Switching Kalman Filter

In Figure 14 we have shown a comparison between the distribution of the results achieved by our particle filter application on user positioning and the results achieved by the proposed Switching Kalman Filter in [19]. The histogram represents bins of size 500 meters. As we can see it seems the Switching Kalman Filter is performing better in the first bin. There are 220 elements grouped in contrary to 200 for Particle Filter. In the second bin, it seems there are more elements from Particle Filter and in the third, the situation is the same. In general, Switching Kalman has a mean of 988 m and a standard deviation of 1631 m. The Particle Filter has a mean of 1571 m and a standard deviation of 2586 m. We can see the particle filtering has more outliers with 5 present bins between 10000 m and 17000 m while Kalman has only one bin.

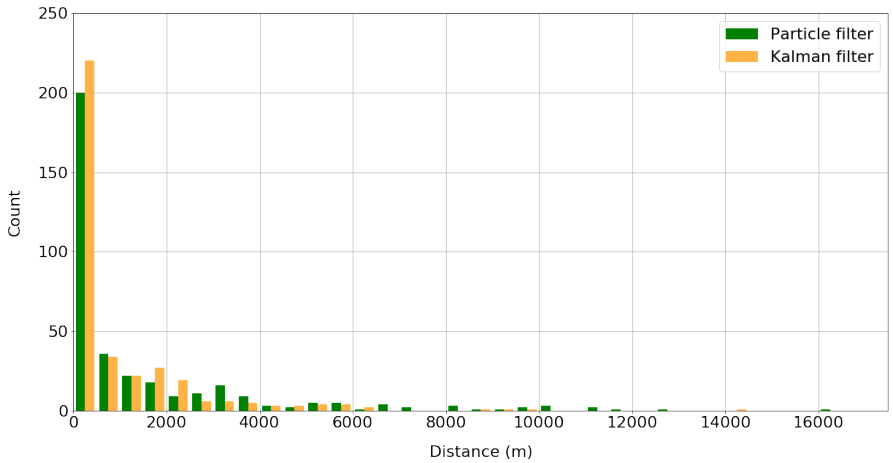


Figure 14. Distribution of errors for Particle Filter and Switching Kalman Filter in real CDR data

Cell size effect

Another point we are interested in, are the dynamics of our predictions related to the cell size. If the elements of the first bin in the histogram come only from cells with a small size, then it shows that our method is somehow biased. Remember that the maximum error we can have is at most equal to the largest diameter of the hexagon. In this case approximately as the estimated diameter of the circle. Figure 15 represents the relationship between the algorithm's accuracy and diameter size for both Particle Filtering and Switching Kalman Filter. It is clear that the errors are positively correlated with the diameters. The larger the diameter, the wider the spread of the errors. However, it is visible that even in some large cells there are errors less than 2500 m.

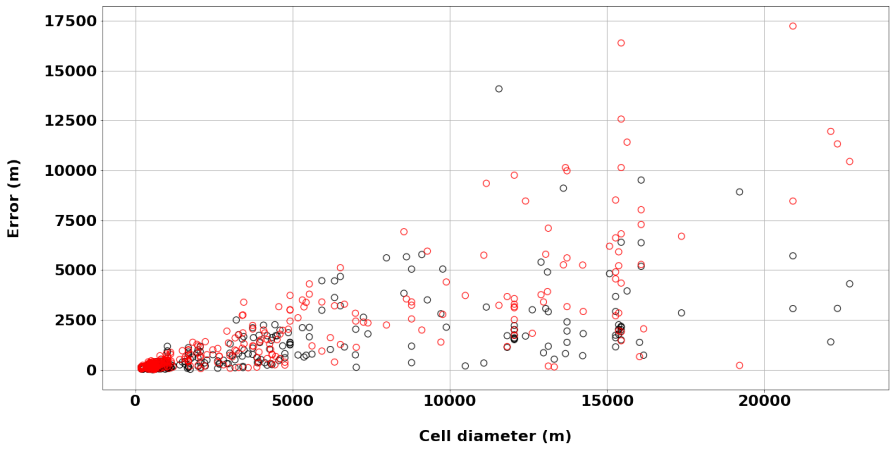


Figure 15. The relation between the accuracy and the diameter size in real data

For the Particle Filtering the results about mean and standard deviation for different cell sizes are given in Table 4.

Diameter	Mean error	Standard deviation
< 5000 m	508 m	631 m
5000 - 10000 m	2928 m	1443 m
10000 - 15000 m	4434 m	2816 m
15000 - 20000 m	5790 m	3923 m

Table 4. The location error for different cell sizes groups in real CDR data

Path evaluation

For almost half of the data, we were provided with full GPS trajectories. This allowed us to evaluate our predicted paths using the method presented in the section "Path Evaluation". The mapping of GPS location to the edges was really time-consuming because the GPS data were extracted every 1 second, and the road network graph of Estonia had a hundred thousand of edges. Hence, we decided to use a sampling strategy depending on the size of GPS trajectories. I.E every 10th record in cases where there are more than 500 GPS points in one trajectory. The loss of information is not a concern as the GPS data was already really dense and the path accuracy formula considers GPS as ground truth.

For Switching Kalman Filter the path predictions were not given. However, we estimated them to be the shortest path between every two consecutive user locations predicted by the model. The edges in the generated shortest path are then compared with the edges in the GPS trajectory. The results for both methods are shown in Figure 16. Most of the accuracy values fall within the first bin of 10%. The rest of them seems to be almost uniformly distributed. There are elements even in the last bin with 90 - 100% accuracy. The Switching Kalman filter has more elements in the first bin and less in the last compared to Particle Filtering. The average accuracy for our method is 17% and for the Switching Kalman Filter the average is 14%.

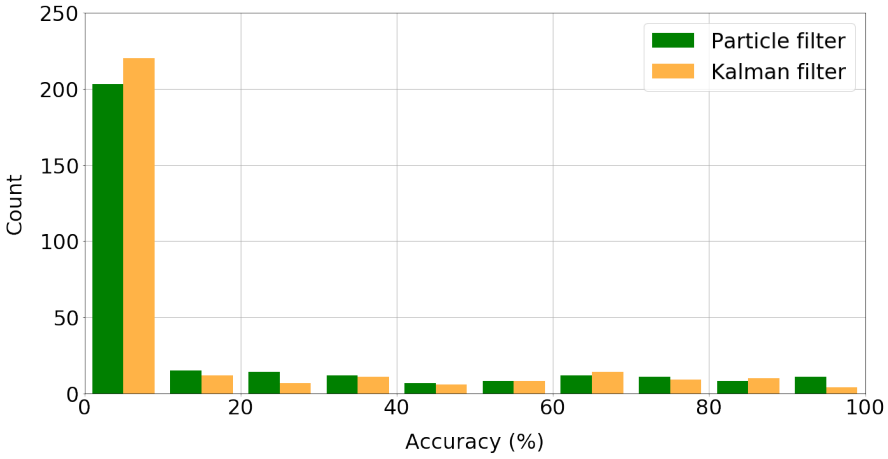
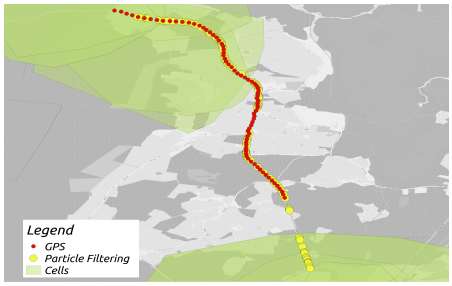


Figure 16. Path accuracy in real data

The two images in Figure 17 are extracted from the generated user trajectories by the particle filter. In Figure 17 (a) the path accuracy is only 19%. The user, in this case, has moved within the same cell, which has a very large size. You can notice how it overlaps with the smaller cells. Despite the wide coverage area, our algorithm has been able to locate the user in the south of the cell, really close to the actual position. In

the second example, Figure 17 (b), the path is matching almost completely. It can be noticed that the GPS records, marked with red in this image, do not start exactly within the cell. We mentioned previously that one of the reasons might be that there are not enough GPS data during the time frame between two CDR events. A second reason was introduced in section 2.1 Global System for Mobile communication. Due to mobile network functionalities, the mobile devices are not always connected to the closest base station. Therefore, the GPS location is not within the cell.

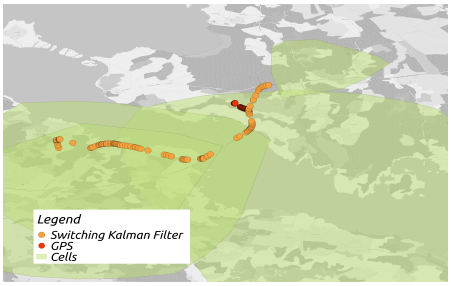
(a) Path with accuracy 19%



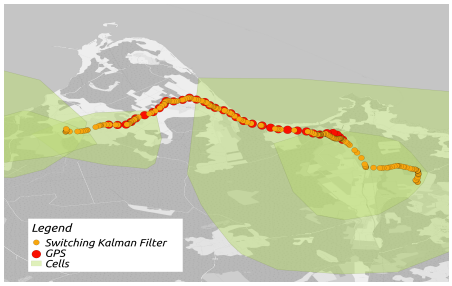
(b) Path with accuracy 87%

Figure 17. Examples of predicted trajectories from Particle Filtering compared to real GPS locations

In Figure 18 are displayed two random samples from path predictions of Switching Kalman Filter. In the first subfigure (a) the path has an accuracy of only 15% and the prediction is far away from the real position. In subfigure (b) the method has achieved to depict almost perfectly the trajectory followed by the user. Both methods seem to perform better in cases where there is some movement episode with non-overlapping cells.



(a) Path with accuracy 15%



(b) Path with accuracy 85%

Figure 18. Examples of predicted trajectories from Switching Kalman Filter compared to real GPS locations

5 Discussions

This chapter will discuss the challenges and opportunities of the non-linear Particle Filtering method. Furthermore, the insights that we derived from the experiments and the future work will be described.

It was Gustafsson et al. [20] who first said that the mobile data could potentially be used for positioning, in particular with particle filtering. The paper dates from 19 years ago and since then it seems that the researchers have been more interested in using aggregated mobile data for fields like urban planning, transport mode detection, public health, crisis management similar to COVID-19 scenarios, etc. Today, it exists only one study which considers individual user positioning with CDR data. However, instead of particle filtering, this study uses the Switching Kalman Filter. In our knowledge, this thesis is the first work that considers the implementation and evaluation of a non-linear Sequential Monte Carlo Method like Particle Filtering for the task of user positioning, using only minimal information. The main drive behind this work was to understand if it is possible to improve the positioning accuracy using a non-linear method compared to existing linear methods. Any improvements in findings will lead to improved data quality for passive human mobility analyses that use CDR information.

5.1 Insights from the implementation and the results of the experiments

During the implementation phase, it was clear that one important factor that influences the accuracy of this application is the correct prediction of the travel time between two nodes. Currently, we are calculating this metric using the road segment length and the maximum speed attribute attached to each edge in the data extracted from Open Street Map. Nevertheless, a high percentage of edges are missing this attribute. For these cases, we have to consider a single default maximum speed that is not differentiated for the city roads or highways. The selection of the default speed limit is arbitrary and tries to take into account the general knowledge about the transport legislation in the area of interest. This approach introduces a second approximation step in top of the particle filtering which in itself provides approximate solutions. This is an important factor that affects the path accuracy results that we received in Figure 16. It was our expectation that the majority of predicted paths would not match the paths given by the GPS trace. First due to the approximate nature of Particle Filtering and secondly due to travel time estimations.

Another important factor that should be taken into consideration is the nature of coverage areas. From the study case on CDR data collected in Estonia, we could see that they were far from perfect. The coverage areas overlap with each other in a large portion and the umbrella cells were present. Moreover, they were not always optimal. Many mobile phones were connected to some nearby cells that were not foreseen. This means that the coverage areas provided by MNOs are not totally precise. One way to deal with them is by applying coverage optimization based on data-driven models like authors in [18], [19]. This approach has shown on average an improvement of 312 meters on error.

We have evaluated the effect of particle size on the results. In contrary to what the general theory of particle filtering describes and our expectations, a larger number of particles is not producing considerable improvements. However, it seems that 50 particles work best and the improvement stops when we increase the number to 100 particles. Nevertheless, from the CDF plot, we noticed that 100 particles have a higher likelihood in the first bins compared to the other sampling levels. It means the higher number of particles, the higher the probability that more elements will be in the first bins. There is a considerable difference in execution time when using 100, 50, or 20 particles. Hence, a decision on the trade-off between the improvements and the execution time is necessary. One interesting confirmation that we received from the experiments was the fact that the time granularity of generated CDR records is related to the accuracy of the particle filtering. The worst performing scenario was when only every fourth record was used in prediction. The last insight we had from this set of experiments was related to the cell sizes. There was an increase of almost 400 meters in mean every time that we increased the cell diameter with 1 000 m. When the cell was minimal (d = 800 m) we could see that the error was considerably small.

The second set of experiments was performed in real data. From the comparison between our method and the non-linear Switching Kalman Filter we noticed that although they produce similar results, particle filtering in this dataset was not able to outperform the Switching Kalman Filter. One characteristic that might impact positively the accuracy of the Kalman Filter is that it can depict the episode of Movement or Stay. Particle Filtering has a lower chance to predict the same location twice, in case the user is not moving throughout a series of records. First, because the newly sampled particles are selected randomly and the probability that some of them are similar to the last location depends highly on cell sizes. And second, even if the previous stay location is duplicated in the new set of particles, the probability that is selected again is high when the time gap between two events is small. The probability lowers when the time gap increases, which might be the case if the user is staying at home or office and does not use the mobile for a long time. Although we evaluated both methods in only 359 data points, it might be necessary to use a larger data-set to compare both methods for the results to be more generalizing.

Particle Filtering, in general, it is known to be a complex algorithm with high time requirements. This is the main reason why the other methods are preferred in linear models with a normal noise, even though it would be totally possible to use particle filtering as well. In this particular application, when we combine particle filtering methods with the shortest path calculations in a large graph the time complexity increases. Even after parallel processing performed on the shortest paths calculations, Particle Filtering continues to be slow, especially compared to the Switching Kalman Filter.

5.2 Future work

Based on the discussion above we have identified the areas that can be improved in this methodology.

First, there is a necessity to be able to better model the travel time, not depending on the maximum speed limit only. The sophistication in the method we use for travel time estimation can take into consideration the traffic situation, most common paths, traffic rules, etc. Therefore, we can model the spatial and temporal variations in speed. However, these models require access to large scale data of vehicle locations.

Second, in this thesis, we do not consider the separation between the pedestrian movements and the movements via different transportation modes. This is mainly due to the fact the data was coming only from moving vehicles. In future applications, it would be beneficial to add one step that detects the mode of transportation and the output serves as an input for the travel time estimation.

Additionally, in order to reduce the temporal uncertainties before applying the particle filtering model we can perform the step of trajectory reconstruction using one of the methods introduced in the works of [15], [17] or [16]. For the spatial uncertainties encountered with not accurate cell shapes, it might be necessary to apply some preprocessing like coverage area optimization or overlap detection.

Lastly, it will be interesting to see the results of applying the algorithm to real data with higher quantity and quality.

6 Summary

Geo-localization is becoming an important feature of many applications and data analysis projects and in many cases, it provides competitive advantages. Today the data extracted from Global Positioning System (GPS) are highly reliable, have high time precision (almost every second), and the first choice for real-time applications. On the other hand domains like urban planning, tourism, public health etc. on have the need to have large scale mobility data, with low acquiring and processing costs. Unfortunately, this is not possible with GPS data. Not all users are willing to share their GPS location. Furthermore, the data extraction and processing requires high power consumption. That is the reason why these domains are looking towards mobile data from Mobile Network Operators. Mobile data are retrieved from MNOs for billing purposes every time that we use our mobiles for calls, SMS, or 4/5G internet service. If we look at the data from all MNOs operating in a country we can extract the mobility patterns of all the citizens within a given period. According to the studies and statistics the mobile users are increasing in number as well as the CDR data frequency.

The quality of CDR data received from MNOs is closely related to the functionality of the mobile network. In Chapter 2 we gave an overview of the main standard for mobile networks, called GSM, and described its main components: Base station subsystem (BSS), Network switching subsystem (NSS), and Support subsystem (OSS). We described how the CDR event is triggered and its contents. Later on, we discussed the context in which the CDR can be used for user positioning and the spatial and temporal challenges that come with it. In section 2.3 we saw that different research areas have embraced CDR differently. They are serving mostly as support data for other research topics like tourism, transportation, etc. Meanwhile, there were studies that were dealing with two main challenges of CDR: temporal and spatial uncertainties. The first group uses trajectory reconstruction techniques and the second group deals with user positioning. We did take a look at the adaption of the particle filtering in localization and tracking problems and noticed that nobody so far has used it for user positioning with CDR data. This thesis will be the first method to use a nonlinear modeling technique like particle filtering for user positioning.

Particle filtering concepts have their beginnings in a publication by Gordon et al.[23] in 1993. Later on, the technique was improved and adopted in many fields. It falls under Sequential Monte Carlo methods and is used to model nonlinear phenomena with not normally distributed noise. The technique involves a set of particles that are drawn from the prior belief distribution. The main steps of the algorithm can be described as Predict, Update, and Reweigh. Our application follows the same phases to predict the location and movements of a mobile user. The first step required CDR data and a map of edges

and nodes representing the road segment. The user position is predicted to be normally distributed on the set of edges within the cell edges. A metric that correlates the actual time between CDR records and the predicted traversal time of the path is used to decide on the weights of each particle. In the end, a resampling step is applied were the more important particles are duplicated and the others are dropped.

We first tested our algorithm in an experimental setting using GPS data from taxis driving in Beijing center. After applying the steps described in section 4.2 the CDR events were generated. There were three main parameters that we kept in control and evaluated their effect on the algorithm accuracy: particle size, cell surface size, and time granularity. The results showed that in experimental settings the particle size did not have a significant impact on the accuracy of the model. On the other hand, the time granularity and cell size had a noticeable effect. Additionally, we evaluated the algorithm in a real case with data from [19]. We compared our results with the results that authors of the same paper have received. The results showed that our algorithm was performing at quite a similar scale but was not able to overcome their accuracy. Additionally, we have evaluated the path accuracy in a portion of the data.

7 Conclusion

This thesis aimed to evaluate if it is possible to improve the accuracy of the user positioning with mobile CDR data by introducing the adaptation of a non-linear particle filter algorithm to the problem. Our proposed model was evaluated in synthetic and real-case data. Based on the analysis we could conclude our approach was achieving similar results but not better compared to the previous linear techniques like Switching Kalman Filter. Although, our method had a 3% higher accuracy in the path evaluation. The major factor affecting the accuracy of our method is the travel time estimation. In the future, it is necessary to sophisticate the travel time estimation approach by modeling spatial and temporal fluctuations in speed. In addition, it would be in the interest of this research to compare both methods in a larger data sample. Our proposed method is completely new in a field that remains important for many areas and could be useful in different scenarios, especially particular cases like public health. The main benefit is the fact that it leverages mobile data, that have low extraction cost and are available for every mobile user, with only basic features required.

References

[1] K. Boyini, "Global system for mobile communications," 2018. [Online]. Available: https://www.tutorialspoint.com/global-system-for-mobile-communications

[2] The International Engineering Consortium, "Global system for mobile communication ({GSM})." [Online]. Available: http://www.uky.edu/{~}jclark/mas355/GSM.PDF

[3] N. Deblauwe, *GSM-based Positioning: Techniques and Applications, year = 2008*. Asp / Vubpress / Upa.

[4] "Cell sizes." [Online]. Available: http://www.wirelesscommunication.nl/reference/chaptr04/cellplan/cellsize.htm

[5] A. Turner, "1 billion more phones than people in the world," 2018. [Online]. Available: https://www.bankmycell.com/blog/how-many-phones-are-in-the-world

[6] P. Fiadino, V. Ponce-Lopez, J. Antonio, M. Torrent-Moreno, and A. D'Alconzo, "Call detail records for human mobility studies: Taking stock of the situation in the "always connected era"," in *Proceedings of the Workshop on Big Data Analytics and Machine Learning for Data Communication Networks*, ser. Big-DAMA '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 43–48. [Online]. Available: https://doi.org/10.1145/3098593.3098601

[7] M. Lebdi, "What happens to your data when you make a call ?" 2019. [Online]. Available: https://medium.com/sfu-cspmp/what-happens-to-your-data-when-you-make-a-call-f05812da4f8d

[8] T. Vajakas and J. Rõõmusaare, "On optimal spatial probability density estimation of passive mobile positioning events," in *2016 15th Biennial Baltic Electronics Conference (BEC)*, 2016, pp. 127–130.

[9] R. Sikder, M. Uddin, and S. Halder, "An Efficient Approach of Identifying Tourist by Call Detail Record Analysis," 2016.

[10] E. Saluveer, J. Raun, M. Tiru, L. Altin, J. Kroon, T. Snitsarenko, A. Aasa, and S. Silm, "Methodological framework for producing national tourism statistics from mobile positioning data," *Annals of Tourism Research*, vol. 81, p. 102895, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0160738320300396

[11] S. Qin, J. Man, X. Wang, C. Li, H. Dong, and X. Ge, "Applying Big Data Analytics to Monitor Tourist Flow for the Scenic Area Operation Management," *Discrete Dynamics in Nature and Society*, vol. 2019, pp. 1–11, 2019.

[12] G. Khodabandelou, V. Gauthier, M. El-Yacoubi, and M. Fiore, "Population estimation from mobile network traffic metadata," in *2016 IEEE 17th International Symposium on A World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, June 2016, pp. 1–9.

[13] M. Friedrich, K. Immisch, P. Jehlicka, T. Otterstätter, and J. Schlaich, "Generating Origin-Destination Matrices from Mobile Phone Trajectories," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2196, pp. 93–101, 2010.

[14] M. Pourmoradnasseri, K. Khoshkhah, A. Lind, and A. Hadachi, "OD-Matrix Extraction based on Trajectory Reconstruction from Mobile Data," 2019.

[15] S. Hoteit, S. Secci, S. Sobolevsky, C. Ratti, and G. Pujolle, "Estimating human trajectories and hotspots through mobile phone data," *Computer Networks*, vol. 64, pp. 296–307, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1389128614000656

[16] G. Chen, A. Viana, M. Fiore, and C. Sarraute, "Complete Trajectory Reconstruction from Sparse Mobile Phone Data," *EPJ Data Science*, vol. 8, 2019.

[17] T. Vajakas, J. Vajakas, and R. Lillemets, "Trajectory reconstruction from mobile positioning data using cell-to-cell travel time information," *International Journal of Geographical Information Science*, vol. 29, no. 11, pp. 1941–1954, 2015. [Online]. Available: https://doi.org/10.1080/13658816.2015.1049540

[18] A. Lind, A. Hadachi, and O. Batrashev, "A new approach for mobile positioning using the cdr data of cellular networks," in *2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*. IEEE, 2017, pp. 315–320.

[19] A. Hadachi and A. Lind, "Exploring a New Model for Mobile Positioning Based on CDR Data of The Cellular Networks," 2019.

[20] F. Gustafsson, F. Gunnarsson, N. Bergman, U. Forssell, J. Jansson, P. . Nordlund, and R. Karlsson, "A framework for particle filtering in positioning, navigation and tracking problems," in *Proceedings of the 11th IEEE Signal Processing Workshop on Statistical Signal Processing (Cat. No.01TH8563)*, Aug 2001, pp. 34–37.

[21] L. Hu and D. Evans, "Localization for mobile sensor networks," in *Proceedings of the 10th Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '04. New York, NY, USA: Association for Computing Machinery, 2004, p. 45–57. [Online]. Available: https://doi.org/10.1145/1023720.1023726

[22] B. Dil, S. Dulman, and P. Havinga, "Range-Based Localization in Mobile Sensor Networks," in *European Conference on Wireless Sensor Networks*, 2006, pp. 164–179.

[23] A. F. M. Smith, "Novel approach to nonlinear/non-Gaussian Bayesian state estimation," *IEE Proceedings F (Radar and Signal Processing)*, vol. 140, no. 2, pp. 107–113(6), apr 1993. [Online]. Available: https://digital-library.theiet.org/content/journals/10.1049/ip-f-2.1993.0015

[24] K. P. Murphy, *Machine learning: a probabilistic perspective*, ser. Adaptive computation and machine learning series. Cambridge, MA: MIT Press, 2012.

[25] D. Neff, "Deriving the {Haversine} {Formula}," apr 1999. [Online]. Available: http://mathforum.org/library/drmath/view/51879.html

[26] Y. Zheng, "T-Drive trajectory data sample," aug 2011. [Online]. Available: https://www.microsoft.com/en-us/research/publication/t-drive-trajectory-data-sample/

[27] J. Yuan, Y. Zheng, X. Xie, and G. Sun, "Driving with knowledge from the physical world," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '11. New York, NY, USA: Association for Computing Machinery, 2011, p. 316–324. [Online]. Available: https://doi.org/10.1145/2020408.2020462

II. Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Salijona Dyrmishi**,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

   Creating a novel approach for mobile positioning based on CDR data,

   supervised by Amnir Hadachi, PhD.

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Salijona Dyrmishi
13/05/2020