UNIVERSITY OF TARTU

Institute of Computer Science
Computer Science Curriculum

Annett Saarik

Trajectory Reconstruction and Mobility Pattern Analysis Based on Call Detail Record Data

Master's Thesis (30 ECTS)

Supervisor: Amnir Hadachi, PhD

Tartu 2017

Trajectory Reconstruction and Mobility Pattern Analysis Based on Call Detail Record Data

Abstract: Up until now, GPS data has been greatly used for collecting highly precise locational data from moving objects including humans. In contrast, mobile phone data is becoming more and more popular in the last few years. The usage of mobile phone data, that is also known as CDR data, has many benefits over the widely used GPS. This means that the methods used for example in GPS trajectory reconstruction, need to have modifications made be compatible with CDR data.

The fact that telecommunication companies have started to cooperate more and share the CDR data with the public is also a boost to the usage of CDR data. The processed and analyzed CDR data can be used to get an overview of crowd movement in different scales, for example traveling inside a city as opposed to between countries. Extracting trajectories from CDR data has numerous complications. This is due to the fact that the data might not be continuous and discovering of the starting point of the object in motion is complicated.

The goal of this thesis is to use CDR data in the reconstruction of trajectories made by an anonymous user and to validate the results with GPS data generated in parallel to the CDR data. Reconstructed trajectories can be used for movement analysis and population displacement and would help city planning by optimizing the infrastructures.

Outcomes of this thesis are the reconstructed trajectories based on CDR data and the precisions of final paths. Also, the frequency of CDR events is analyzed in addition to distance distribution. After that the areas that the user visits most frequently are extracted, such as home and work locations.

Keywords: Call detail record, trajectory reconstruction, location data, crowd movement, mobility patterns

CERCS: P170

Trajektooride taastamine ja inimeste liikumise mustrite analüüs mobiiltelefoni andmete põhjal

Lühikokkuvõte: Tehnoloogiad, mis kasutavad geograafilisi andmeid, on muutunud meie igapäevaelu tähtsaks osaks. Tänu sellele on kasvanud asukoha andmete massiliine salvestamine ja kaevandamine. Seni on GPS tehnoloogiad olnud põhiliseks geograafiliste andmete kogumismeetodiks. Sellega paralleelselt on populaarsust kogunud mobiiliandmete kasutamine positsiooni tuvastamiseks ja liikumismustrite analüüsimiseks. Mobiiliandmete (CDR) põhjal trajektooride taastamiseks on vajalik meetodite kohendamine selleks, et tulemused oleksid korrektsed.

Tänu sellele, et telekommunikatsiooni ettevõtted on alustanud suuremat koostööd ja hakanud CDR-andmeid järjest rohkem avalikustama, on mobiiliandmete kasutamine mitmetel aladel suurenenud. Töödeldud mobiiliandmed aitavad anda ülevaadet rahvastiku liikumisest erinevates ulatustes. Samal ajal on trajektooride taastamine CDR-andmetest kohati raskendatud võrreldes GPS-andmetega. Suurimaks probleemiks on algus- ja lõpp-positsioonide asukoha määramine, mis on veelgi enam raskendatud juhul kui objekt liigub.

Selle lõputöö eesmärgiks on trajektooride taastamine anonüümsete kasutajate poolt genereeritud CDR-andmete põhjal. Tulemuste valideerimine GPS-andmetega, mis on loodud paralleelselt mobiiliandmetega ning on vajalik selleks, et määrata saadud trajektooride täpsust. Loodud trajektoore saab kasutada objektide, sealhulgas ka inimeste, liikumismustrite analüüsimiseks ja rahvastiku paiknemise tuvastamiseks, mis aitab linnade planeerimisel ja infrastruktuuride optimeerimisel. Lõputöö väljunditeks on trajektooride taastamine ja täpsuse analüüsimine, lisaks sellele inimese liikumismudelite tuvastamine ja tihedamini külastatavate asukohtade identifitseerimine nagu näiteks kodu, töökoht ja poed.

Võtmesõnad: Mobiiltelefoni andmed, trajektoori konstrueerimine, asukoha andmed, rahvastiku liikumine, inimeste liikumise mustrid

CERCS: P170

Acknowledgements

First, I would like to extend my gratitude to the best supervisor in the world, Dr. Amnir Hadachi for endless guidance, patience and the ability to keep a sense of humour, when I had lost mine. For the last year he has been a solid source of knowledge and wisdom.

I would like to thank my parents, Anu Saarik and Silver Saarik, who always believed in my capabilities even when I did not and supported me in my decision to pursue studies in Tartu University. I am grateful for the work they have done, to give me the life I have now. Without their encouragement and faith, I could not have done it.

Last but not least, I would like to thank Karl Rankla, Mari Krusten, Reimo Rebane and Jenny Holm for helping me by reading and giving valuable advice to improve my work. To the computer science ladies, Olha Shepelenko and Asmar Hasanova, big thanks for the support and for keeping the morale high during the thesis writing.

Nolite te Bastardes Carborundorum

Contents

List of Abbreviations

API Application Programming Interface

BTS Base transceiver station

CDR Call Detail Record

CI Cell ID

CS Computer Science

EVP Exceptionally Visited POIs

GPS Global Positioning System

GSM Global System for Mobile Communication

GU Geographical Unit

ITS Intelligent Transportation Systems

JSON JavaScript Object Notation

KNN K-Nearest Neighbor

LAC Location Area Code

LCSS Longest Common Subsequence

MCC Mobile Country Code

MMPP Markov modulated Poisson process

MNC Mobile Network Code

MVP Mostly Visited POIs

OSM OpenStreetMap

OVP Occasionally Visited POIs

POI Point of Interest

SMDR Station Messaging Detail Record

XML Extensible Markup Language

List of Figures

List of Tables

1 Introduction

The introduction chapter gives an insight about positioning and how this thesis work is contributing into the field of Intelligent Transportation Systems (ITS). The research that has been conducted is focused on investigating the use of mobile phone data also known as Call Detail Records (CDR) in reconstructing trajectories and extracting human mobility patterns.

1.1 General view

The usage of geospatial data in people's everyday lives has become more frequent over the years. Geospatial data is a dataset that includes geographic data. Spatial data makes mundane tasks easier by helping to make travel plans from short trips to the market, to longer more elaborate plans, to travel across country borders. This data can be presented with location attributes such as latitude and longitude, address, postal code, street name etc. In some cases the information is represented by more complex data types and structures.[PT14]

One option to collect enormous amounts of geospatial data is from Global Positioning System (GPS) devices, such as mobile phones, which are used frequently by people in their daily lives. The growth over the last 15 years in the number of mobile cellular services subscribers is remarkable. In the year 2015 the number of unique cellular (GSM) subscribers has increased to around 4.7 billion.[mob16]

There are some disadvantages with collecting GPS data, such as the user must enable the settings for GPS tracking and this drains the battery of the device faster. This is not the case in collecting CDR data. CDR is a record of every transaction a mobile subscriber makes. These transactions include calls, messaging or even just waking the mobile phone from sleep mode. This means that every time, when a mobile phone is connected to an antenna, CDRs are made continuously and stored by the telecommunication companies. Companies save the records for billing the subscriber for their use of the telecommunication services.[ict16]

In this thesis CDRs are acquired from a mobile application developed by the distributed systems group in University of Tartu, Computer Science Institute. This data is used to calculate the possible location data (latitude and longitude) of the mobile user at the moment of the CDR creation. From these sets of latitude and longitude it is possible to construct trajectories of users from a certain starting point A to destination point B. Using these trajectories helps in understanding complex movement patterns that mobile phone users make in their daily lives. These patterns can give an overview of the complicated migration of crowds in small and large scale. Small scale meaning commuting to workplace or traveling around the user's residential area. Large scale is more sophisticated movement from non-urban to urban areas and vice versa also migration to foreign countries.

1.2 Research questions and objectives

The goals of this thesis is to compute user trajectories from CDR data, validate them with the GPS data that was collected in parallel, to assess the legitimacy and precision of the CDR trajectory reconstruction results. An additional goal is to examine individual user's mobility patterns and areas of significance.

- Is it possible to to reconstruct human generated trajectories based on CDR data and what is the accuracy compared to GPS data?

- What are the individual user's mobility patterns and places of significance based on CDR data?

1.3 Scope

In this thesis, CDRs that are used in the trajectory reconstruction and further pattern analysis are generated by mobile users, who have the mobile application, developed by distributed group, *mobCollector* installed and have agreed to share their data for academic purposes. Data is collected inside Estonia country borders. In total 3721 CDR and 682 GPS records were used.

1.4 Contributions

Methodologies used in this thesis are mainly built on data analysis: understanding, preprocessing, cleaning, modeling and individualizing the data. The main objectives that are targets in this thesis are:

- acquiring cell ID location data from OpenCellID;
- calculating the cell coverage area polygon centroids;
- processing road data;
- discovering the nearest nodes from the road data;
- reconstruction of the trajectories based on CDR data;
- CDR event distribution analysis;
- CDR distance dispersion analysis;
- detecting frequently visited locations.

1.5 Road map

The road map gives a brief introduction to the structure and chapters of this thesis.

Chapter 2. Gives an overview of the current research that has been done in the related fields of this thesis. Subjects such as description of CDR data, trajectory reconstruction and human mobility patterns are covered.

Chapter 3. Includes a more detailed sight into the geospatial data used in this thesis and a thorough description of the steps and calculations made to generate the trajectories based on CDRs.

Chapter 4.Is about analyzing one anonymous user's event and distance distribution in addition to identifying most frequently visited places.

Chapter 5. Conclusion of the thesis and description the possible future work is in the last chapter.

2 State-of-the-art

This chapter gives an overview of research that has been conducted in the related CS fields. CDR data, trajectory reconstruction and crowd movement have been part of many academic research papers and this chapter will examine the possible overlapping of these subjects.

2.1 Call Detail Records

CDRs are records that are generated by the mobile subscriber of a telecommunication company. These records are generated, when the mobile phone is connected to an antenna and the user has interacted with their mobile phone. Different interactions are called events. The records can consist of various data and usually are not identical among the providers. There is no universal format implemented for this data and the providers can choose the content of the records themselves. Given the sensitivity of information in CDRs, it is a good practice to anonymize the identifying fields in the records. This means the names and/or mobile numbers are removed from the data and commonly replaced with unique integer numbers for specific subscribers.

2.1.1 Relevant issues with CDRs

Because CDR data is very different from GPS data it presents multiple challenges in processing it. Some of them are:

- Temporal sparseness - The CDRs are generated when the user interacts with their mobile phone. A large number of mobile users make infrequent calls and messages or the records made are periodically irregular. This is not the case with GPS generated data.

- Spatial sparseness - The location recorded with an event is the location of the cell tower and this brings the spatial sparseness into the CDR data.

- Non-routine events - Regular events like going to work or home are easier to detect. Non-regular events like football or some other social events are not part of the usual routine trajectories and therefor are more unpredictable. [DPG+15]

The privacy concerns with using and processing CDR data are an additional challenge. Even when names and mobile phone numbers are anonymized and all identifying data linked to the user is removed, there is an increasing awareness of the re-identification possibility. For example, identifying the specific field of work or a profession of a mobile subscriber from a seemingly random CDR dataset. By clustering one specific user's two most visited areas it is possible to discover significant locations by checking, if it is a residential area for home area.[Pul13]

2.1.2 Applications with CDR data

There are many possible development opportunities for mobile phone network data. Benefits in smart city planning and transportation are a given and presented in one of biggest projects with mobile phone data in fixing bus routes in the city of Abidjan. The telecommunication company released 2.5 billion CDRs to research the possibilities of improving the bus routes and scheduling in the city. The research included extracting frequent sequential patterns from the stops made and locating users' home and work areas, resulting in 65 improvement suggestions and two new added routes. This optimized the system enough for a 10 percent decrease in travel time for citizens.[BCDL+13]

In addition to transportation system planning, CDRs can also be used in disaster response. This was researched with data collected after the Haiti earthquake in 2010. It is a natural response to any disaster to flee from the affected areas, therefore finding out exactly how people react and move after catastrophes can help in organizing and managing first responders. Disasters also include infectious disease outbreaks and man-made hazards, for example terrorist attacks or industrial accidents.[BLT+11]

There have been many projects made in health research and disease prevention with mobile phone data. One of most significant studies was with quantifying malaria outbreaks in Kenya. Around 15 million mobile subscribers' data was acquired for a time period of one year, to map their regional travels. Together with the malaria transmission model, which shows the rate of infection, specific ares were located, where the probability of malaria spreading was higher.[WET$^+$12]

Social science research can also benefit from applying mobile phone data. In 2012 census data and CDR were used to research and find recognizable patterns in various social groups of the subscribers. In this project census data consisted of socioeconomic information from the same area as the CDRs are collected from. Around 10 million subscribers' data was acquired from 12 cities[FMV12]. Results showed a strong correlation between the socioeconomic level of a specific subscriber and the expenses, physical distance with the contacts and geographical areas where people travel. Another research in Republic of Côte d'Ivoire was conducted to use CDRs to determine and map poverty lines in that area. Greater amount of mobile communication between subscribers and larger range of calls are an indicator for larger prosperity. As a result poverty lines of eleven regions of Côte d'Ivoire were estimated.[SMC13]

2.2 Trajectory data mining

Spatial trajectories are location sequences generated by moving objects, such as humans, vehicles or even animals. As a result of rapidly advancing tracking technologies and mobile computing, processing and generating trajectories from that data is becoming more prominent. Research in trajectory reconstruction and data mining is extensive. Most of the studies inspect different techniques for trajectory computing with location data like GPS or CDR. In the research paper "Trajectory Data Mining: An Overview" by Dr. Yu Zheng, there is detailed description of trajectory construction methods. The five major subjects trajectory preprocessing, indexing, uncertainty, pattern mining and classification are introduced in the next four chapters.[Zhe15]

Figure 1 gives a simple overview of the process flow in trajectory data mining and subjects covered in this thesis. Spatial data in the beginning of the chart can represent both GPS and CDR data. In this thesis it represents CDR data collected from volunteer users.

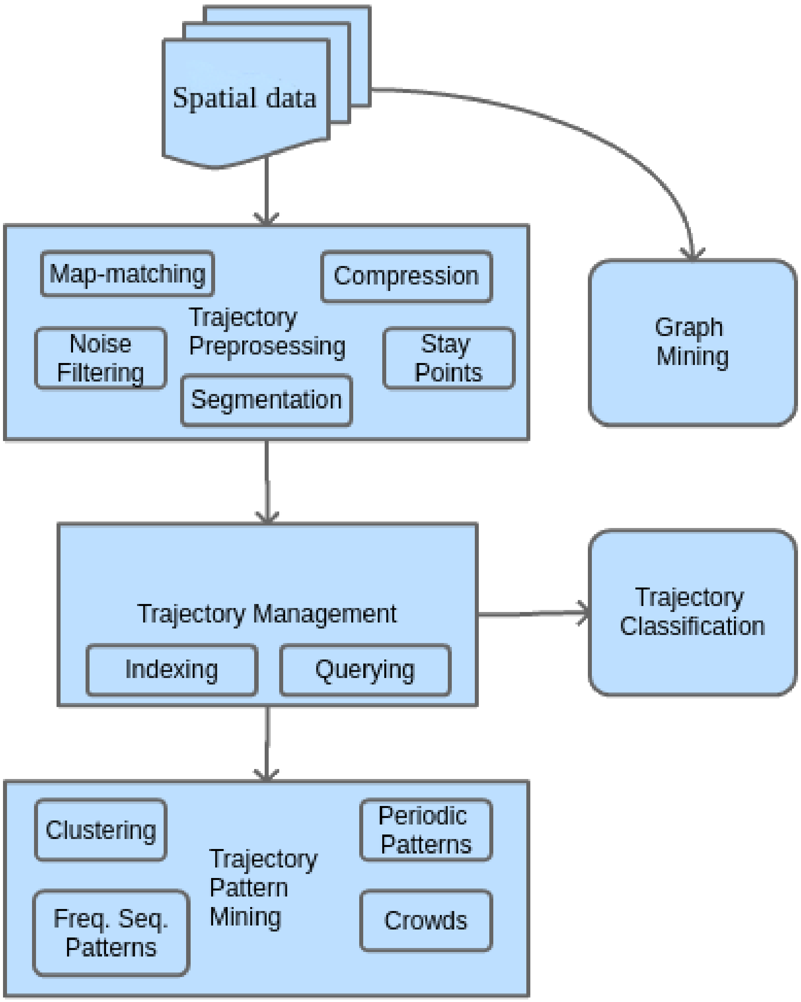


Figure 1: *Overview of trajectory mining methods.*

2.2.1 Trajectory preprocessing

Before using the trajectory data there are numerous problems that need addressing in order to start working with the data. More substantial issues concerning trajectory preprocessing can be covered with five suggested solution methods explained in more detail in the list below.

- *Noise filtering.* When working with trajectory data some location measurements may be incorrect by several (hundred) meters. These deviations in the data depend on the technology used and the physical objects that interfere with the signal near the true location. To remove the inaccurate points from the trajectory a median filter can be used. The filter algorithm calculates medians between a point n and its $n-1$ predecessors in a period of time. If the next median differs more than the agreed upon allowed error, it is a noise point. The median filter is not efficient for sparse trajectory data. This means the filter cannot be used with CDRs. For scattered data *Kalman* filter or particle filter can be used. *Kalman* filter takes a motion model into account and estimates states like speed with assuming linear models. The particle filter algorithm relaxes these assumptions and therefore is a less efficient algorithm.[LK11]

- *Stay points.* Some points in a trajectory are more significant than other, for example opposed to noise points there are also stay points. These points represent locations where an object has stayed for a longer period. When calculating data generated by humans, the stay points can be shops, malls, restaurants etc.[Zhe15]

- *Compression.* To minimize memory storage and computing time there are two compression methods. The offline method reduces the trajectory after it has been completely generated and the online method, which compresses trajectories instantly as an object moves.[LK11]

- *Segmentation.* For a very detailed analysis on trajectory segmentation it is split into smaller parts. Segmentation can be based on time, the shape of the trajectory or semantic meaning (walking, driving) of the parts.[Zhe15]

- *Map-matching.* Last preprocessing method is map-matching, which converts a sequence of latitude and longitude data to a road segment where the object generated corresponding points.[Kru11]

2.2.2 Trajectory data management

Searching and querying over enormous sets of trajectory data is time consuming. Indexing increases the efficiency of these queries and makes trajectory data storage management easier. Queries are divided into two types K-Nearest Neighbor (KNN) and range. KNN query recovers the top-K trajectories that are positioned inside the minimum accumulated distance range. In order to use KNN queries a distance (or similarity) function, that is based on minimum bounding rectangles between two trajectories, needs to be determined. Two concepts that are acquired from string matching can be implemented in function Longest Common Subsequence (LCSS) and Edit Distance. Range queries are also referred to as distance metric, meaning the evaluation of distances between two trajectories. Range queries fetch segments of the trajectory that are within a previously defined spatial range. The result of the query can be used to verify a feature within the segment such as object speed.[DXZZ11]

2.2.3 Trajectory uncertainty

As a result of spatial data sparseness uncertainty occurs in trajectories. As objects move constantly, but the number of recorded location points are limited, the locations of objects between two documented points are ambiguous. This problem arises with CDR data more frequently than in GPS.[Tra11]

2.2.4 Trajectory pattern mining

In trajectory pattern mining there are four distinguishable categories that observe and inspect patterns from one or more trajectory. These categories include moving together patterns, trajectory clustering, sequential patterns and periodic patterns.

- *Moving together patterns.* As the name suggests, this category discovers objects that move together in a certain period of time. These patterns are used in species migration, military surveillance and traffic event detection. It can be used to detect possible bottlenecks in city road systems during rush hours. The groups of objects traveling together are called flocks and swarms. Flock is a group that is moving simultaneously in at least k period of time. Swarm is a composed version of a flock leaving out the time requirement.[JYJ11]

- *Trajectory clustering.* Combining together objects, that have same paths at some point in their movements, is called trajectory clustering. The paths are split into segments and objects with similar ones can be identified by calculating the distances between two complete trajectories. Micro-and-Macro Clustering approach is also used to first find clusters in very small subsets and grouping small micros together to generate bigger macro clusters.[JYJ11]

- *Sequential patterns.* When objects travel in homogeneous paths through points and during similar times, sequential patterns can be identified. The sequences share same locations and comparable travel times, although the sequence does not have to be consecutive. As an example view the trajectories A and B in *Formula 1.*

$$A : l_1 \xrightarrow{\text{1.5h}} l_2 \xrightarrow{\text{1h}} l_7 \xrightarrow{\text{1.2h}} l_4. \quad B : l_1 \xrightarrow{\text{1.2h}} l_2 \xrightarrow{\text{2h}} l_4, \tag{1}$$

  where l is a location. In the example A and B share the same sequence $l_1 \to l_2 \to l_4$ although they are not consecutive. Discovering these patterns can enhance the accuracy of next location calculation, estimating similarities, trajectory compression and travel recommendation.[JYJ11]

- *Periodic patterns.* Searching and identifying recurring events results in periodic patterns. These patterns are made by objects that generate similar trajectories when certain time period has passed. For instance people going to the market on weekends or gift-shopping before holidays. For detecting these patterns two stage detection method is used. Density algorithm is implemented to find popular locations among objects. Considering the result, trajectories are reconstructed into time series with values *in* and *out* as a status of the moving object at a certain popular location. The final stage is to generate summaries from partial movement sequences using hierarchical clustering algorithm.[JYJ11]

2.2.5 Trajectory classification

Separating trajectories (or parts of it) by difference in status is called trajectory classification. Various states include movement, such as walking, biking, driving or using public transport. Adding these semantic descriptions to trajectories adds value in context aware computing. Trajectory classification process is broken down to three major stages:

1. Using segmentation methods to divide trajectories into sections.
2. Extracting characteristics from all sections.
3. Generating a model to identify every section.[Zhe15]

A research project based on GPS data categorizes object's trajectory by transportation mode [ZLWX08]. The main classes were driving, biking, walking and taking public transport and the reason for them is that a person can take multiple transportation method during one trajectory. The segments are identified with a class through Decision Tree Classifier. The principal is that movement information about heading change, stop and velocity adjustment rate are extracted and read into the Decision Tree after that results are classified the a model is implemented in the Decision Tree.[ZCL+10]

2.2.6 Applications of trajectory data mining

Numerous fields, such as transportation, urban planning, environment, energy, social, business and public safety, use trajectory mining applications. In this chapter urban planning and transportation applications are explained in detail [MT16]. For a regular smartphone user more familiar applications might be path discovery in transportation such as Google Maps[1] or any alternative, for instance WikiMapia Map[2], MapQuest Map[3] or Waze Map[4]. The main applications for trajectory data mining are explained in detail in the list below.

- *Path discovery.* Is the most common trajectory mining application, it is also very popular among users in their daily lives. Finding the most reasonable route for travels has been the focus point of many research projects [DYGD15]. As users' preferences for route attributes vary the path discovering algorithm also differ. As some people prefer shorter distance lengths

[1] www.google.ee/maps

[2] www.wikimapia.org/

[3] www.mapquest.com

[4] www.waze.com/livemap

for smaller gas consumption others only care about the time spent on the travels. More sophisticated path detections techniques also take into account the traffic and even weather. In ITS various research examines how to update the paths simultaneously with real time data from various sensing technology.[Zel98]

- *Destination prediction.* Is linked to path discovery and it is found that human mobility and movement is profoundly regular and therefore predictable in high precision. In a research paper about restraints of human mobility the prediction accuracy rate was 93% [SQBB10]. Large number of location based applications use destination prediction to send advertisements or special offers to potential clients. Recording past trajectories to databases improves destination predictions. When a user is traveling through a regular path destination corresponds with the final location in the past trajectories.[CLC10]

- *Movement behavior analysis.* Trajectory calculations gives many ways of analyzing object's movement and finding occurring patterns. One substantial research is conducted in determining patterns between sociodemographic groups based on age, wealth, gender, educational level and wealth [RBdM$^+$13]. Another research paper identifies groups, such as animals, humans and vehicles, traveling together in time intervals. The movement behavior for this type of event is called *gatherings* and are found by large dataset indexing, searching and updating issues. *Gatherings* can be celebrations, parades, protests, traffic congestions and other public assemblage. The five main characteristics of *gatherings* are:

  1. number of participants is high;
  2. participants arrange a compressed group;
  3. event should occur in a certain time interval;
  4. geometric attributes of groups stay the same;
  5. there is a number of participants, who stay in the group at any time for a certain interval.[ZZYS13]

  In the research a sizable dataset, consisting of location data, is collected from taxicabs in Beijing [ZZYS13]. Research in discovering object's, in this case a human's, rationality to enter a certain point of interest (POI). POI can be any potential stopping point from shopping malls, restaurants and to bus stops, hospitals.[LQW15]

- *Group behavior analysis.* Examines clusters of objects, that are likely to generate between groups, during motion. These clusters develop due to their social behavior and to discover them techniques from *Chapter 2.2.4* are used. Research has been conducted in trajectory modeling to describe the movement patterns in shifting groups. These patterns include events like parades, protests and traffic bottlenecks.[ZZYS13]

- *Urban computing.* Obtaining data from sources such as sensors, devices, vehicles, buildings, humans and analyzing it to find better solutions for problems in the city. Using data to solve these issues, like air pollution, energy consumption and traffic bottlenecks, in cities is called *urban computing* [FZ16]. The usage of trajectory data in urban development has many benefits, for example processed trajectories can be used to optimize public transportation schedules and routes, also in planning and building new roads. The identification of regions, such as residential, business and education, in cities helps urban planners to understand the complexity of cities.[ZCWY14]

- *Understanding trajectories.* Making sense of trajectory data without semantic descriptions can be problematic and to simplify this attributes are added to segments of trajectories by modeling data with specific features. More frequently used semantic attributes are divided by the mean of transportation used, such as *walk*, *drive*, *bus*[PSR$^+$13]. Some location applications require a semantic attribution of locations for instance *work* or *home* [LCC12]. Another method to make the trajectory data understandable is to use visual analytics methods such as map-matching, graphs, images.[AA13]

2.3 Human mobility patterns

Human mobility has been mentioned several times in *Chapters 2.1* and *2.2.4*. This chapter gives a more detailed analysis about human mobility pattern discovery and presents prominent research in the field. This chapter is divided into four subchapters: transportation, urban planning, event detection, semantic analysis. All of these subchapters cover a specific field, which uses mobility pattern applications.

2.3.1 Transportation

Detecting and analyzing human mobility patterns from location data has numerous benefits in transportation, such as rescheduling public transport when needed and discovering traffic anomalies. Planning and building streets and infrastructure is also moderately connected to transportation but in this paper, will be covered in *Chapter 2.3.2.* This chapter gives an overview of techniques for processing CDR data to improve the transportation infrastructure in a certain region.

Understanding the human mobility in a city and detecting POIs for the purpose of improving transportation is crucial. There is regularity in the trajectories and time in human mobility, that lead to the following characterizing aspects:

- movements of individuals are summarized in as a set of points, where they stay the longest time;
- places visited only once are time consuming;
- humans travel between points based on temporal distance;
- most frequently visited POIs are *home* and *work*.[PJZ+16]

Traffic anomalies such as congestion, accidents, bottlenecks can greatly affect normal traffic flow in cities. To identify unusual events from mobile data metrics, such as trip rates, travel distance and travel time need to be calculated [CMS+16]. In the research project about determining these values trajectories of individuals were mostly used.[GHB08]

Observing the traffic flow in cities, at a crowd level can give additional insight to problems. Big datasets of trajectories are an effective base for understanding mobility patterns at society-wide range. For example a study in Milan identified roads most used by the morning and evening commuters in addition to the exact times when traffic bottlenecks developed in the city.[GNP$^+$11]

2.3.2 Urban planning

Using spatial data in urban planning can increase the understanding of urban dynamics and human movement flows. Analyzing location data allows urban planners to monitor the fast changing urban dynamics. Also to detect upcoming trends in movement of the citizens[DL99], which can be very time consuming and difficult using traditional surveys, such as questionnaires.[RFPW06]

Approximately 15 million CDRs were collected around city of Morristown to get an overview of residents' daily travels. The geographical areas surrounding the city, where workers live are given a semantic name *laborshed*. In contrast, areas where people would frequently visit the community location such as bars and restaurants are called *partyshed*. By grouping residents by their preferred activities around the city it is possible to model typical flow of the residents between different parts of the city.[BCH$^+$11]

Human mobility is complex, but almost never random. The movement of people is affected by their needs, commitments and social obligations. As an outcome of these factors human mobility patterns show regularity in daily (weekly, monthly, etc) movements. These regularities can be characterized by defining the *Relevance Ratio* and POI [NSL$^+$12] of the user u under observation, such as in *Formula 2.*

$$RR(POI, u) = \frac{d_{visit}(POI, u)}{d_{total}(u)} \tag{2}$$

where $d_{visit}(POI, u)$ is the sum of days that the location has been visited and $d_{total}(u)$ is the sum of all days in the user data. These locations are categorized into three classes Mostly Visited POIs (MVP), Occasionally Visited POIs (OVP) and Exceptionally Visited POIs (EVP), by frequency of visits.[JZGR16]

2.3.3 Event detection

Event is considered a substantial activity not regular in daily human patterns. To discover these type of *events*, object's history and regular movements are taken into account and any sort of deviations from regular paths are detected. After that Poisson and exponential distribution models are used. With these steps it is possible to characterize regular behavior and recognize anomalous events. If the thresholds of frequencies and spans are smaller than for any *event* that occurred, it can be categorized as an anomaly.[ZD12]

Alternative method for detecting unusual events, from more than one object, is more suitable for CDR data, because of the issues described in *Chapter 2.1.1*. The process is divided into multiple steps, firstly CDRs are received and split into clusters, after that crowds are detected from sequences of clusters. Afterwards constraints are verified for each crowd detected and a tag (*unusual*) is added to it. One or more crowds construct an unusual event.[DPG^{+}15]

Definition 2.1. (Crowd) A crowd C is $\left\{CC_{tm}, CC_{tm+1}, ..., CC_{tn}\right\}$ that represents consecutive clusters with three constraints. Movement, because number of points visited needs to be over one. Durability, number of consecutive clusters is bigger than threshold. Commitment, number of participants in any given moment is greater than the threshold.

Unusual events were detected from CDRs in Abidjan, Côte d'Ivoire. Markov modulated Poisson process (MMPP) was modified to detect hourly and daily behavioral anomalies from spatial data. As result unusual events found from mobile data correlated with events such as protests, holidays and major sport events that actually occurred in the area. Additionally, analyzing mobile data as a time series gives better output in tracking masses during movement.[GISL16]

2.3.4 Semantic analysis

From object spatial data such as CDRs most visited places, also known as POI, are attainable as mentioned in previous chapter. Giving these POI a semantic meaning adds more personal information about the object under observation. Two places with the highest frequency of visits often are *home* and *work*. Finding other semantic locations from trajectory data is more complex.

The two issues with combining the location data to a semantic meaning, are obtaining the spatial data and after that labeling the locations with semantic meaning. To obtain most occurring points from location data, clustering (partitioning, density-based, time-based) algorithms are used. The disadvantage with these techniques is that the result is a geographic point with a radius, but does not include a semantic meaning. Another approach uses hierarchical algorithm compiled with both time-based and density-based clustering.[LCC12]

Definition 2.2. (Location point) Is P from the trajectory as a pair $p = (lat, lon)$. Where lat is the latitude and lon is the longitude.

Definition 2.3. (Trajectory) Is a sequence of location points with added timestamps, represented as $traj = \{(p_0, t_0), ..., (p_n, t_n)\}$, where p is a location and t is the timestamp.

Definition 2.4. (Visit point) Is a location with two timestamps, defined as $vp = (p, t_{in}, t_{out})$, where p is the location and t_{in} is arrival time and t_{out} is departure time.

Definition 2.5. (Physical place) Is a cluster of location points, represented as $pp = (vp_1, ..., vp_n)$, where vp_1 and vp_n are nearby.

The time-based algorithm checks, if vp is located in the cluster of points already generated, and compares the time intervals and distances. If the time period is greater and distance is less than the tolerated threshold, it is in fact a visit point. [LCC12]

In another research same types of clustering algorithms were used to find various user groups from CDR data. The main attributes for clusters, were time (day, week) and distance (Euclidean distance). All of the unique users' CDR data was aggregated into 1-hour blocks by day of the week. Results show that it is possible to identify student mobile users by their daily and hourly usage pattern. The other group identified is commuters, who use their mobile phones more during the morning and evening rush hours.[BCH+]

In comparison, location data can be nowadays collected from social sites, such as Twitter[5]. This was tried in a project for mining users mobility patterns within

[5] www.twitter.com

urban context. As a result it was possible to identify most frequent route patterns between famous London landmarks. The outcome can be used to generate personalized travel recommendations for tourists.[CFT16]

Because mobile phones are ubiquitous nowadays and present in high and low income households, CDR data can be used to recognize and analyze needs and habits of various groups. In a project conducted in Latin America mobile phone data was combined with socio-economical census data, collected by the National Statistical Institute, during a period of five years and divided into Geographical Units (GUs). CDR data was grouped into polygons using Voronoi diagrams and then merged with GUs. Results display correlations among various socio-economic levels. Larger distances of the call maker and receiver means a higher economic background of the users. Accuracy of these calculations is $R^2 \approx 0.82$.[FMV12]

2.4 Conclusion

To summarize, there is extensive work done in the fields of trajectory data mining and human patterns, but many problems are still unsolved. Thanks to open data movements gaining support, many telecommunication companies have started to take steps towards sharing their CDR data. This gives more opportunities to investigate human mobility more extensively and with less financial losses. At the moment CDR trajectory reconstruction methods are not highly accurate. This is due to the fact that origin and destination point recovery is difficult and usually not very precise.[LWB+13]

3 Calculating the trajectories

This chapter introduces the methods and technologies used in the processing of CDR data that has been collected by a mobile application developed by the distributed systems group at the University of Tartu, Computer Science Institute. In parallel to CDRs, GPS data was also gathered in the same time period. This enables the verification of data legitimacy in trajectory reconstruction covered in *Chapter 3.5*. Additionally to CDRs multiple resources were used to collect and filter data. In *Chapter 3.1* there is a detailed description of CDR data, after that in *Chapters 3.2.1* and *3.2.2* two additional services are introduced that were used in the trajectory reconstruction process.

3.1 Call Detail Records

GSM is an international mobile phone standard that provides connection services for subscribers. The GSM system consists of base transceiver stations or BTSs. Each system covers a geographical area that is called a cell coverage area (polygon). BTS enables the gathering of information about every GSM device that is in connection. This type of data is known as *call detail records* or CDRs (also *station messaging detail record* or SMDR). Information collected in CDRs can include:

1. metadata;
2. phone number of the mobile subscriber or when anonymized, user id (SID);
3. timestamp;
4. cell location data such as MCC, MNC, LAC, CI (descriptions in *Chapter 3.2.1*);
5. event type (i.e. handover, pickup).

In addition to information in the list above, telecommunication operators can include or remove fields as they choose. CDRs are described as passive location data compared to GPS data, which is considered active. CDRs are generally formatted in XML (eXtensible Markup Language) based tagging schema that is defined by the operators. CDR data is used to generate invoices for subscribers of telecommunication companies and to analyze network traffic.[cdr13]

Table 1: *Example of CDR data used in this thesis.*

ID	SID	tascii	CGI
1	100562421962333	2016-10-16 09:00:39.853	248-2-1002-54412
2	100562421962333	2016-10-16 09:00:53.227	248-2-1002-54415
3	100562421962333	2016-10-16 09:00:58.546	248-2-1002-56593
4	100562421962333	2016-10-16 09:01:10.524	248-2-1002-54412

Example of a CDR used in this thesis, is shown in *Table 1*. Irrelevant fields, to this thesis, were left out of the table. User generated CDRs are represented as a row and contains SID, timestamp (tascii) and CGI that includes location data divided by a dash.

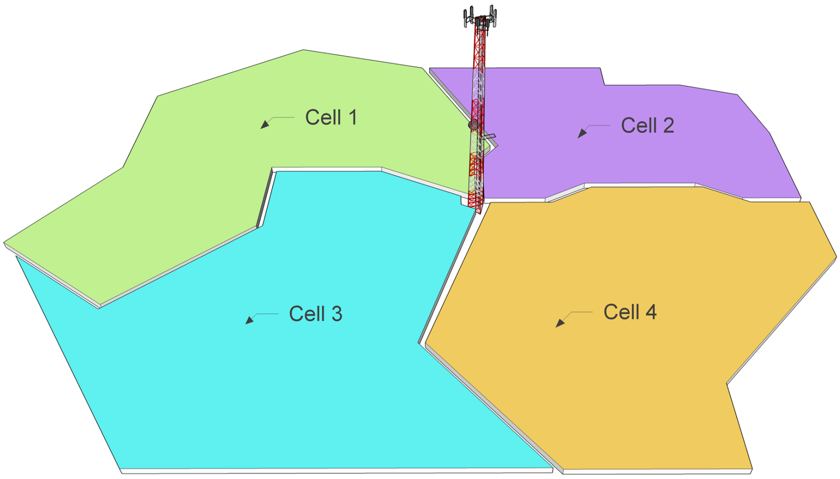


Figure 2: *Antenna and four cell coverage areas (polygons) surrounding it.*

Figure 2 is an illustration of one antenna that can four coverage areas (polygons). In real life the antenna can have multiple cell towers mounted on it. The polygons are irregular and can overlap each other. In more complex cases, one coverage area can wrap another entirely.

3.2 Technologies applied

Various libraries and software was used in this thesis. Reading in and processing CDR data was made by using Python programming language[6] library named Pandas[7]. Pandas is an open source library offering high-performance, easy-to-use data structures and data analysis tools. Jupyter Notebook[8] was used as a programming environment because it contains live code, equations, visualizations and supports Python and all the libraries used in this thesis. Other technologies used in this thesis:

- OpenCellID[9] - explained in *Chapter 3.2.1*;
- OpenStreetMap[10] - explained in *Chapter 3.2.2*;
- NetworkX[11] - explained in *Chapter 3.2.3*;
- QGIS[12] - free open source software for creating, editing, analyzing geospatial information. In this thesis QGIS was used to visualize CDR paths;
- Numpy[13] - package for scientific computing with Python was used to conduct multiple calculations with location data and distances;
- Matplotlib[14] - Python library for generating high quality graphs and plots;
- Requests[15] - the HTTP library for Python was used to make queries to various services;
- Geopy[16] - Python client for numerous geocoding web services was used in trajectory reconstruction;

[6] www.python.org
[7] www.pandas.pydata.org
[8] www.jupyter.org
[9] www.opencellid.org
[10] www.openstreetmap.org
[11] www.networkx.github.io
[12] www.qgis.org
[13] www.python.org
[14] www.matplotlib.org
[15] www.python.org
[16] www.geopy.readthedocs.io

- Scipy[17] - Python-based system of open-source software for mathematics, science and engineering was used in various computations.

3.2.1 OpenCellID

OpenCellID is an open-source community collaboration project that aims to collect and document GPS locations of cell towers and share this data. Volunteer users can upload and download spatial data through a Location API (Application Programming Interface). OpenCellID data together with CDR data can be used to replace GPS as a tracking method with cell IDs, which helps to save device battery power and to track a device in a building, where GPS is not available.

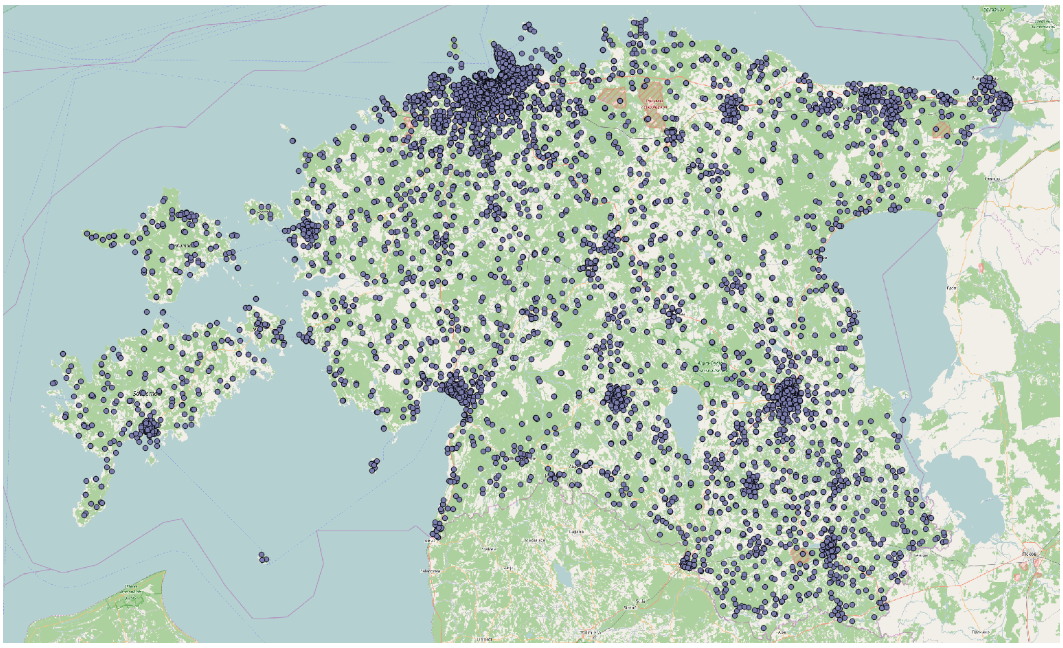


Figure 3: *Cell tower locations in Estonia from OpenCellID API.*

In this thesis four fields of data was used to identify the cell ID. An example of location data is in *Table 1* where cell tower data is in column CGI. When splitting the column by dash, the fields are:

- MCC - Mobile Country Code represented as a *integer* number (Estonia is 248);

[17] www.python.org

- MNC - Mobile Network Code represented as a *integer* number;
- LAC - Location Area Code of the operator network;
- CI - Cell ID.

Accuracy of the data may vary, because OpenCellID is open-source and community based service. This means that some API queries got inaccurate results back or no result at all and these errors were removed. In *Figure 3* all the cell tower positions are shown inside Estonia. In urban areas there are clearly more towers to serve bigger number of users as opposed to rural areas.

3.2.2 OpenStreetMap

OpenStreetMap (OSM) is an online service that collects world's geographic data and distributes it for free. The service has about 20 000 active users, who volunteer to assemble and upload geographic data. All maps generated with collected data are adjustable by users and can be used as they see fit. The data can be downloaded in the format of XML. The XML files include tags:

- *node* is a geographical point on earth with latitude and longitude as attributes;
- *way* is an ordered sequence (list) of nodes that all together make up a portion (polyline) of a road or a street;
- *relation* is a connection to model logical (usually local) or geographic relationships between objects.

OSM data includes additional data about road segments, for example max speed, bus stops etc. In addition to the data, OSM has multiple functions to query and manipulate spatial data. OSM has an API for querying and saving data. [HW08] For this thesis the road data is downloaded by boundary box query to the API. For example the downloaded road data for Tartu had 24655 nodes and 27933 ways.

3.2.3 NetworkX

For the purpose of searching and processing OSM road data faster, nodes, ways and relations were read into a network structure with NetworkX. NetworkX is a Python language software package for creating, manipulating and studying structures and functions of the network, in addition it is possible to read in and store data as a graph. While reading OSM data into structure, every node and way was identified with an ID number from the original OSM road data file. In *Figure 4* nodes in Tartu are shown in a graph structure using NetworkX. Nodes have id, latitude and longitude as data. As seen in the figure, nodes make up a network that resembles Tartu's road system.

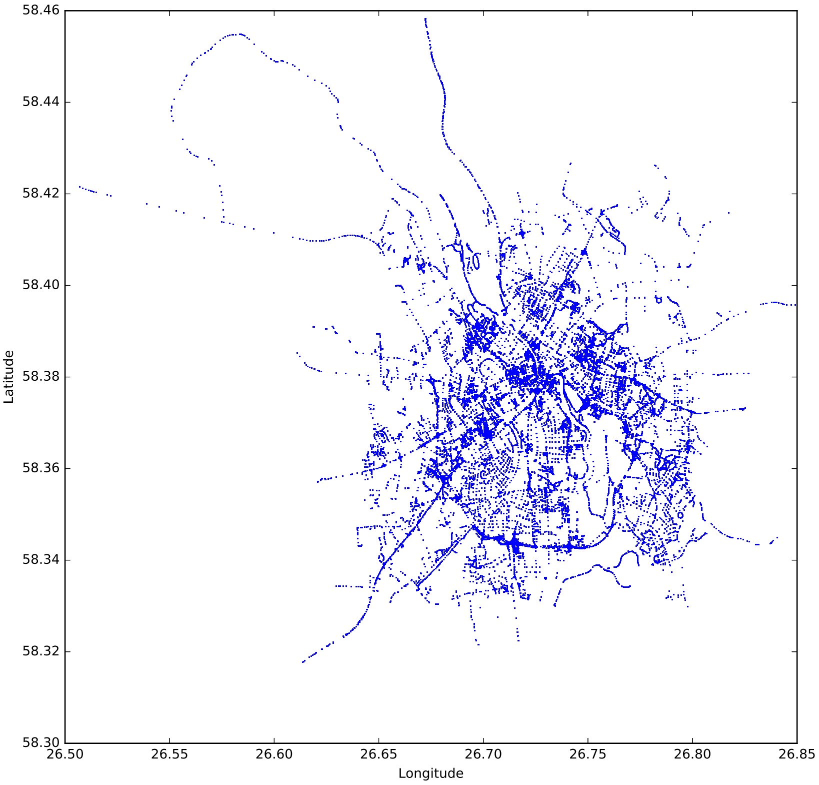


Figure 4: *OSM road data of Tartu shown in NetworkX structure.*

3.3 Trajectory preprocessing

As mentioned in the previous chapters, multiple technologies were used to reconstruct the trajectories from original CDR data. In this section all the steps in data processing are described in detail and intermediate results are visualized.

3.3.1 Cell ID query

OpenCellID API is used to search and download cell IDs in Estonia. CDR data does not include latitude and longitude parameters, because of this querying the OpenCellID API is necessary. When an API query is successful JSON format file is downloaded as a response. An example of a JSON file for one query is below.

```
{
  "lon": 24.774672192307694,
  "lat": 59.377463307692295,
  "mcc": 248,
  "mnc": 2,
  "lac": 1,
  "cellid": 55145,
  "averageSignalStrength":  10,
  "range": 14157,
  "samples": 26,
  "changeable": true,
  "radio": "GSM"
}
```

Listing 1: *OpenCellID API query response in JSON.*

As can be seen from the example above, desired cell tower latitude and longitude attributes are included in the JSON file. Due to the fact that OpenCellID data is gathered by volunteer users, the data may have errors or be missing in some cases. Around one third of the results were empty.

3.3.2 Geometric center

Because the cell ID latitude and longitude are not the cell center, the geometric centroid from a polygon is calculated and the total number of the vertices is unknown in the polygons. Formulas (3),(4) and (5) were used to find the latitude

$C_{latitude}$ and longitude $C_{longitude}$ of the center point of the polygon.

$$C_{latitude} = \frac{1}{6A}\sum_{i=0}^{n-1}(x_i + x_{i+1})(x_i y_{i+1} - x_{i+1} y_i) \tag{3}$$

$$C_{longitude} = \frac{1}{6A}\sum_{i=0}^{n-1}(y_i + y_{i+1})(x_i y_{i+1} - x_{i+1} y_i) \tag{4}$$

$$A = \frac{1}{2}\sum_{i=0}^{n-1}(x_i y_{i+1} - x_{i+1} y_i) \tag{5}$$

In *Formula* (5) polygon's area A is calculated. In *Formulas* (3) and (4) latitude and longitude coordinates is calculated by using the intermediate area result and number of n vertices $(x_0, y_0), (x_1, y_1), ..., (x_{n-1}, y_{n-1})$.

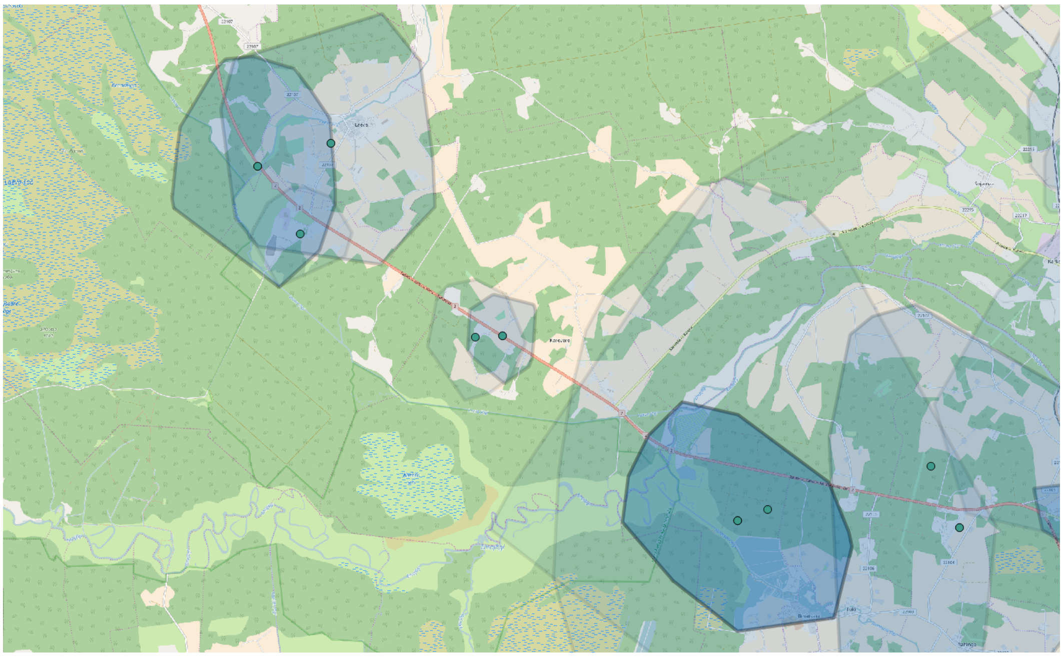


Figure 5: *Calculated centroids with coverage areas (polygons). The centroids are in green and polygons are in blue.*

Calculated centroids are shown in *Figure 5* with the cell coverage area polygons. As can be seen from the figure, some cell areas are overlapping others and some even hide multiple smaller areas underneath.

3.3.3 Importing Tartu road data into a network

Nearest road to the centroid is calculated using OSM service that was described in *Chapter 3.2.2*. Using the API, tiles of OSM with a boundary box are downloaded and processed. Boundary box query consists of attributes, such as minimum longitude, minimum latitude, maximum longitude and maximum latitude. Due to the fact that downloaded OSM files are very big (OSM file of Tartu has 24655 nodes and 27933 ways) a script was made to download road data according to latitude and longitude points. Two different methods were used to calculate bounding box attributes to download road data. Calculating bounding box from one point and two points. The last one was used to reconstruct path between two CDR points. The XML files of road data were read into a complex network with NetworkX. Road data in a network structure enables nearest node and path search.

3.3.4 Nearest road

To find the nearest road OSM data is used to find the closest node to the centroid. Vincenty's inverse formula [Vin75] was used to calculate distances, because of the high accuracy of the calculations with geographic points. The high precision is achieved by calculating geodesic distances on ellipsoids, this gives results within a one millimeter error radius. The high accuracy is important when reconstruction the trajectories and also in mobility pattern analysis. Formula of the Vincenty's inverse equation is shown below and also the notations are explained in *Table 2* .

Table 2: *Notation to the Vincenty's inverse formula.*

a	radius at equator, 6378137.0 meters in WGS-84
f	flattening of the ellipsoid, 1/298.257223563 in WGS-8
$b = (1 - f)a$	radius at the poles, 6356752.314245 meters in WGS-84
$U_1 = arctan[(1 - f)tan\Phi_1]$	reduced latitude
$U_2 = arctan[(1 - f)tan\Phi_2]$	reduced latitude
$L = L_2 - L_1$	difference in longitude of two points
λ_1, λ_2	longitude of the points on auxiliary sphere
α_1, α_2	forward azimuths at the points;
α	azimuth at the equator
s	ellipsoidal distance between two points
σ	arc length in the middle of positions on auxiliary sphere

$$\sin\sigma = \sqrt{(\cos U_2 \sin\lambda)^2 + (\cos U_1 \sin U_2 - \sin U_1 \cos U_2 \cos\lambda)^2} \tag{6}$$

$$\cos\sigma = \sin U_1 \sin U_2 + \cos U_1 \cos U_2 \cos\lambda \tag{7}$$

$$\sigma = \arctan\frac{\sin\sigma}{\cos\sigma} \tag{8}$$

$$\sin\alpha = \frac{\cos U_1 \cos U_2 \sin\lambda}{\sin\sigma} \tag{9}$$

$$\cos^2\alpha = 1 - \sin^2\alpha \tag{10}$$

$$\cos(2\sigma_m) = \cos\sigma - \frac{2\sin U_1 \sin U_2}{\cos^2\alpha} \tag{11}$$

$$C = \frac{f}{16}\cos^2\alpha\big[4 + f(4 - 3\cos^2\alpha)\big] \tag{12}$$

$$\lambda = L+(1-C)f\sin\alpha\left\{\sigma + C\sin\sigma\left[\cos(2\sigma_m) + C\cos\sigma(-1 + 2\cos^2(2\sigma_m))\right]\right\} \tag{13}$$

When λ has been assembled using *Formulas 12* and *13* to the wanted degree of precision (10^{12} corresponds to 0.06mm). Assembling λ is done by iterating over *Formulas(6* to *11.* After all that proceed to evaluate the following formulas.

$$A = 1 + \frac{u^2}{16384}\left\{4096 + u^2\left[-768 + u^2(320 - 175u^2)\right]\right\} \tag{14}$$

$$B = \frac{u^2}{1024}\left\{256 + u^2\left[-128 + u^2(74 - 47u^2)\right]\right\} \tag{15}$$

$$\begin{aligned}\Delta\sigma = B\sin\sigma\Big\{&\cos(2\sigma_m) + \tfrac{1}{4}B\big[\cos\sigma\big(-1 + 2\cos^2(2\sigma_m)\big)- \\ &\tfrac{1}{6}B\cos(2\sigma_m)(-3 + 4\sin^2\sigma)\big(-3 + 4\cos^2(2\sigma_m)\big)\big]\Big\}\end{aligned} \tag{16}$$

$$s = bA(\sigma - \Delta\sigma) \tag{17}$$

, where $\Delta\sigma$ is a result of *Formulas 14, 15* and *16.*

$$u^2 = \cos^2\alpha\frac{a^2 - b^2}{b^2} \tag{18}$$

After all that the result can be calculated by *Formulas 19* and *20*.

$$\alpha_1 = \arctan\left(\frac{\cos U_2 \sin\lambda}{\cos U_1 \sin U_2 - \sin U_1 \cos U_2 \cos\lambda}\right) \tag{19}$$

$$\alpha_2 = \arctan\left(\frac{\cos U_1 \sin\lambda}{-\sin U_1 \cos U_2 + \cos U_1 \sin U_2 \cos\lambda}\right) \tag{20}$$

Explanations to variables in Vincenty's inverse formulas are in *Table 2*. In this thesis WGS84 is used in distance calculations and visualizations in QGIS. WGS84 is an Earth-centered reference system and geodetic datum. WGS84 projection is based on a set of constants and parameters that describes the Earth's size, shape.

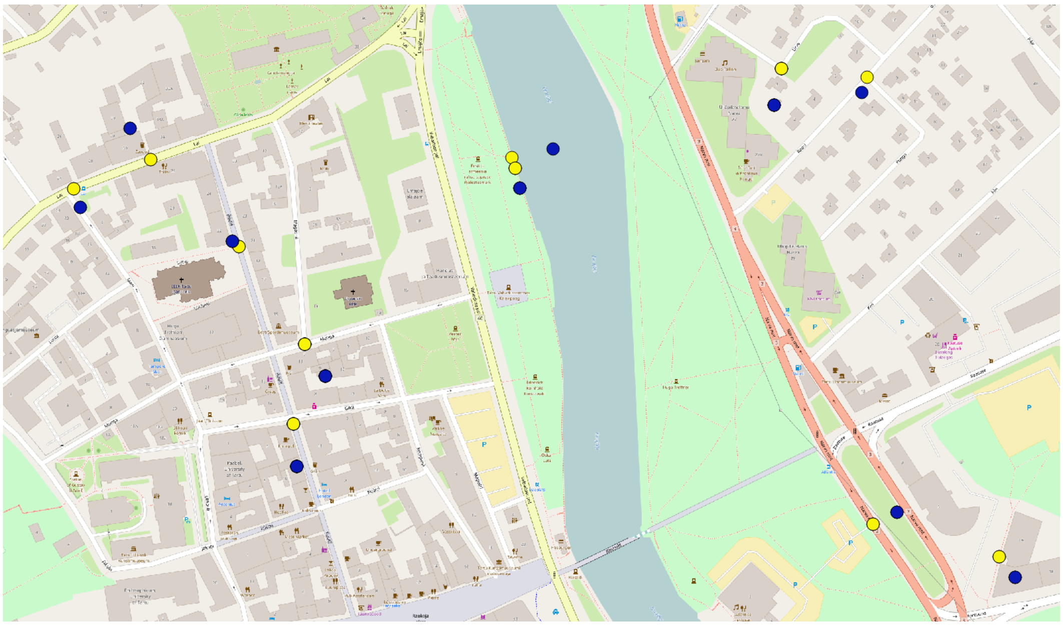


Figure 6: *Cell coverage area centroid and nearest node in OSM. The centroids are in blue and nearest nodes are in yellow.*

Results of the nearest node search from OSM road data are shown in *Figure 6*. Blue points are the calculated centroids from polygons and yellow are the nearest nodes from the road data.

3.4 Reconstructing the trajectories

After completing all of the steps in preprocessing, path reconstruction is possible. Firstly the shortest path between two nearest node points, that are in chronological order, are calculated. With the length of shortest path a cutoff is calculated and used to find paths between two nodes with depth-first search. This is covered in mored detail in *Chapter 3.4.2*. To choose the final path from intermediate results, time intervals are calculated in *Chapter 3.4.1* and compared to timestamps of two CDRs from the original data and the closest time interval is the resulting trajectory.

3.4.1 Calculating the time intervals

Time intervals are used to find closest time to the CDR timestamps. Two different times were calculated to find the most accurate path. The average car driving time and the average walking time for previously generated paths using depth-first search. The speed limits were extracted from road data using the OSM service. When data about the legal speed limits was missing, it was set as 50 km/h (city limit) and 90 km/h (highway limit) for driving and 5 km/h for walking. After calculating the result of estimation times for paths, they were compared to the starting point time and destination point time. The path with the closest estimation time was chosen as the trajectory.[LV11]

3.4.2 Reconstructing paths

Paths are generated by first finding the shortest path between two nodes. After that using length of the shortest path to run depth-first search with a cutoff. Because of the fact that a smartphone can ping-pong between cell towers and connect multiple times to one cell tower even when not in movement, there might be misleading data in CDRs. These type of double nodes were filtered out from CDRs during the path reconstruction process, to reduce inaccuracies of path generation.

```
procedure DFS iterative(G, start, end, cutoff):
    let S be a stack
    S.push(start)
    while S is not empty
        v = S.pop()
        if S is smaller or equal to the cutoff:
            if start is not labeled as discovered:
                label start as discovered
                for all edges from start to end in G.adjacentEdges(start) do
                    S.push(end)
```

Listing 2: *Pseudocode of the depth-first search with a cutoff.*

In *Listing 2*, the cutoff if-statement is used before going through the graph nodes and edges. Intermediate results from depth-first search were saved and time intervals were attached to them. After this, time periods were compared and the closest one was resulting path.

3.4.3 Cell ping-pong handover problem

While processing the CDRs for the trajectory reconstruction an anomaly in the spatial data was discovered. For some sequences of CDRs, multiple cell towers reoccurred numerous times. This problem is called the cell tower ping-pong handover problem, which means that when the user is located between two cell towers the connection can be passed from one tower to another, due to network traffic fluctuations. In trajectory reconstruction, these double cell towers were removed to get more accurate results. This was done by going through the sequence of cell towers and comparing patterns of three last cell towers to the next ones. When the patterns matched, double occurrences were removed.

3.5 Validating with GPS data

As mentioned previously, GPS is considered an active tracking method and CDR a passive one. In this thesis GPS location data was used to verify the results of the trajectory reconstruction. CDR characteristics, such as total number of CDRs, CDR per reconstructed trajectory, total number of OSM nodes per trajectory and the percentage of accuracy of the reconstructed path are covered in *Table 3.* The accuracy percentage is calculated by splitting the CDR trajectories into road segments and dividing the accurate number of segments with the total number of segments in the trajectory.

$$A_{percentage} = \left(\frac{S_{accurate}}{S_{total}}\right) \cdot 100 \tag{21}$$

In *Formula 21* $A_{percentage}$ is the correctly reconstructed trajectory the percentage, $S_{accurate}$ is the number of road segments that are equal to the GPS trajectory and S_{total} is the total number of road segments in the path.

Table 3: *Table of the processed CDR characteristics.*

Characteristics of CDR trajectories				
Date	Total number of CDRs	CDRs per trajectory	Number of trajectory nodes	Percentage of accuracy
1 May	230	15	87	$\sim$ 66%
16 May	240	18	93	$\sim$ 75%
19 May	340	12	46	$\sim$ 69%
20 May	204	16	34	$\sim$ 57%
23 May	61	6	37	$\sim$ 70%

In *Table 3* five datasets of CDRs are shown, separated by date, that were used to create the overview of CDR trajectory reconstruction accuracy. First dataset might have a lower accuracy because the number of CDRs was somewhat lower, but competed to the fourth dataset, it was located near the city border of Tartu and because lower cell towers the accuracy is lower. Second dataset has a much higher percentage than others. This could be because the area where the trajectory was did not have many roads. This mean only one or two paths could be generated between two simultaneous CDRs and the correct one was chosen from the results more often. Fourth dataset gave the most inaccurate result. This might be because the CDR in this dataset were more sparse than in others and this made the trajectory error bigger. Accuracy in the fifth dataset is one of the highest, but not as high as in the second one. This might be due to the fact that the number of CDRs is much lower.

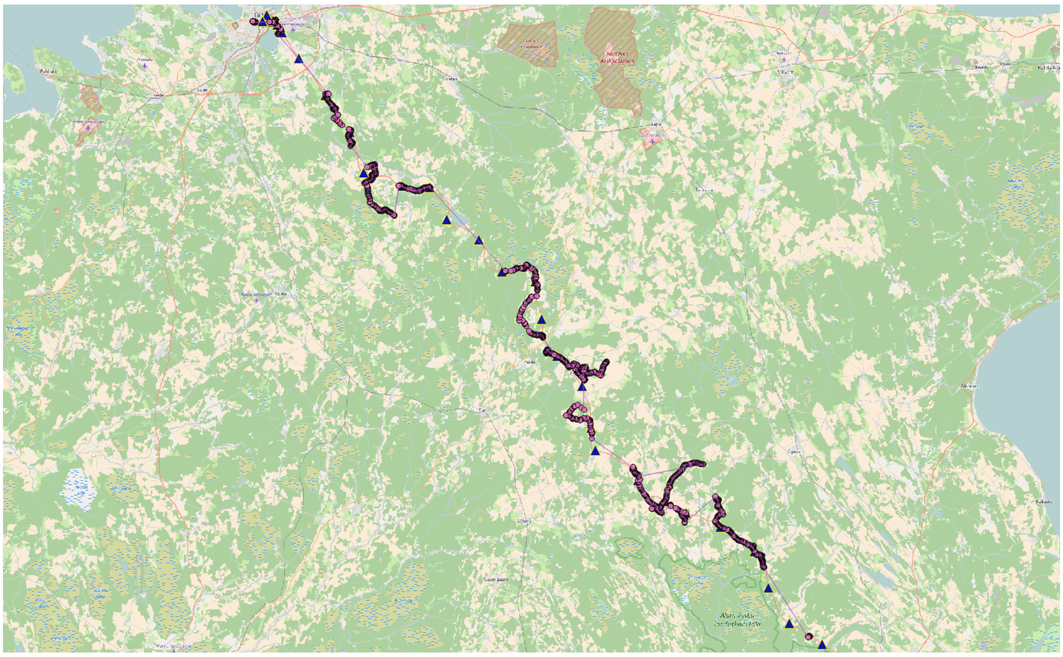


Figure 7: *Trajectory reconstruction based on CDR data. Trajectory points are in pink and GPS locations are shown as blue rectangles.*

Trajectory reconstruction is shown in *Figure 7*, for CDR dataset of 19th of May, outside the city of Tartu. Reconstructed path nodes are pink and GPS positions are represented with blue triangles. The path is not continuous and breaks multiple times. Also the trajectory seems to go along numerous side roads and not go along the Tallinn-Tartu highway in a straight line. This result is expected, because in the reconstruction phase object's real starting position can be different from the cell coverage area centroid. In cases where the object starts movement somewhere near the cell polygon edge, the difference in the calculated and real starting point can be up to hundreds of meters. In *Figure 9* that situation is displayed in cell number 3. This difference in real positions and the centroid affects trajectory reconstruction less in urban areas, because the number of antennas and cell areas is increased. This means that the coverage areas are smaller and close together. This decreases the amount of side roads in the final trajectories and the paths are more accurate. The overall accuracy of the trajectory reconstruction in this thesis is $\sim 67.4\%$.

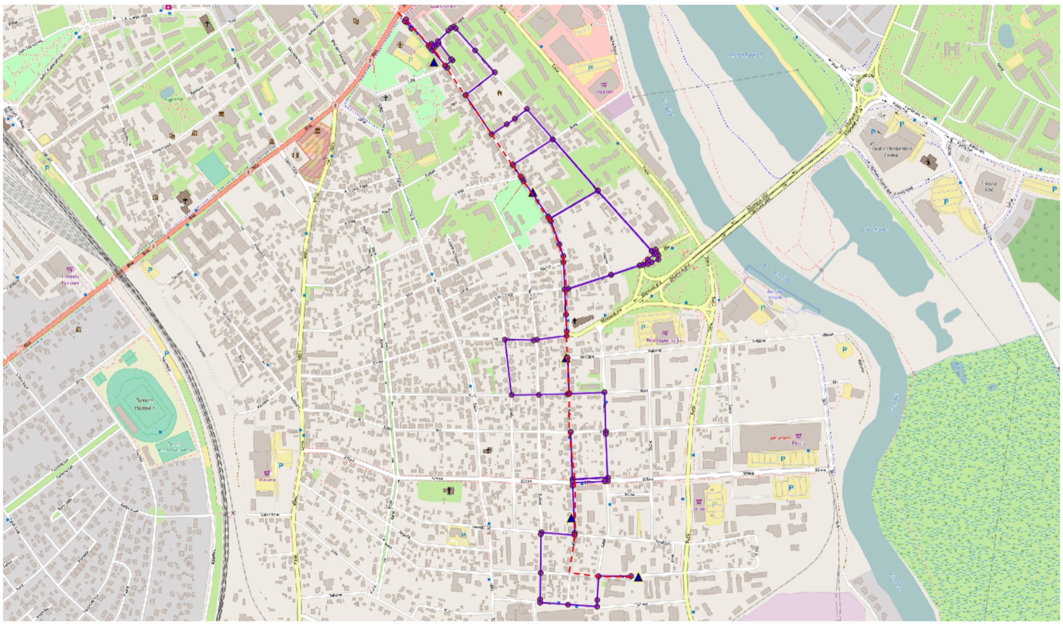


Figure 8: *Trajectory reconstruction based on CDR data. GPS trajectory is in red and CDR trajectory is in blue.*

In *Figure 8* reconstructed CDR trajectory is illustrated. The path seems to take many side streets and this is because of the fact that object's real position and the polygon centroid are not int the same location. There is method for trying to fix this problem using *Kalman* filters and identifying if the cell is a stay, bypassing or a jump cell and by using these identifiers the real position of the user is calculated with a higher accuracy in stay cases, while in bypassing it is lower.[BHLV15]

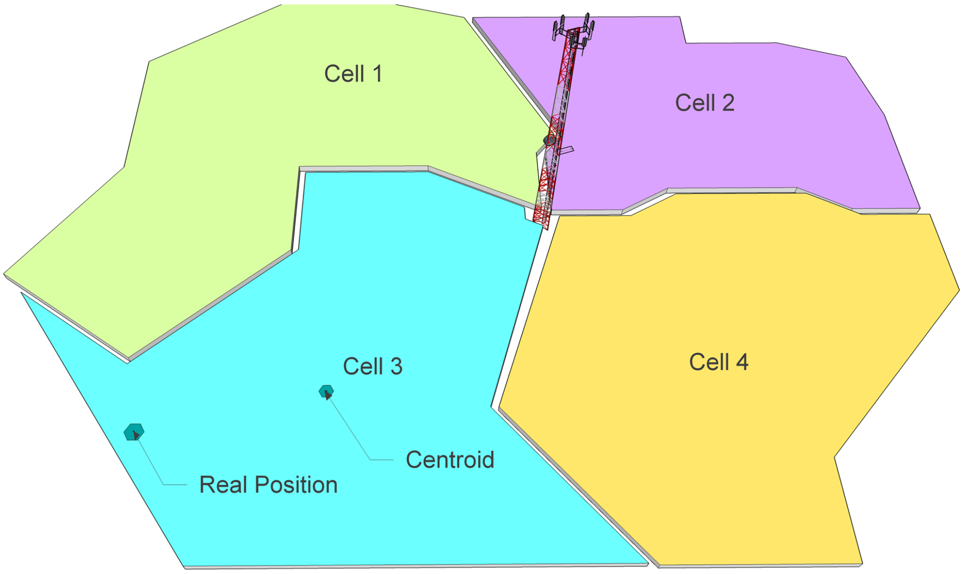


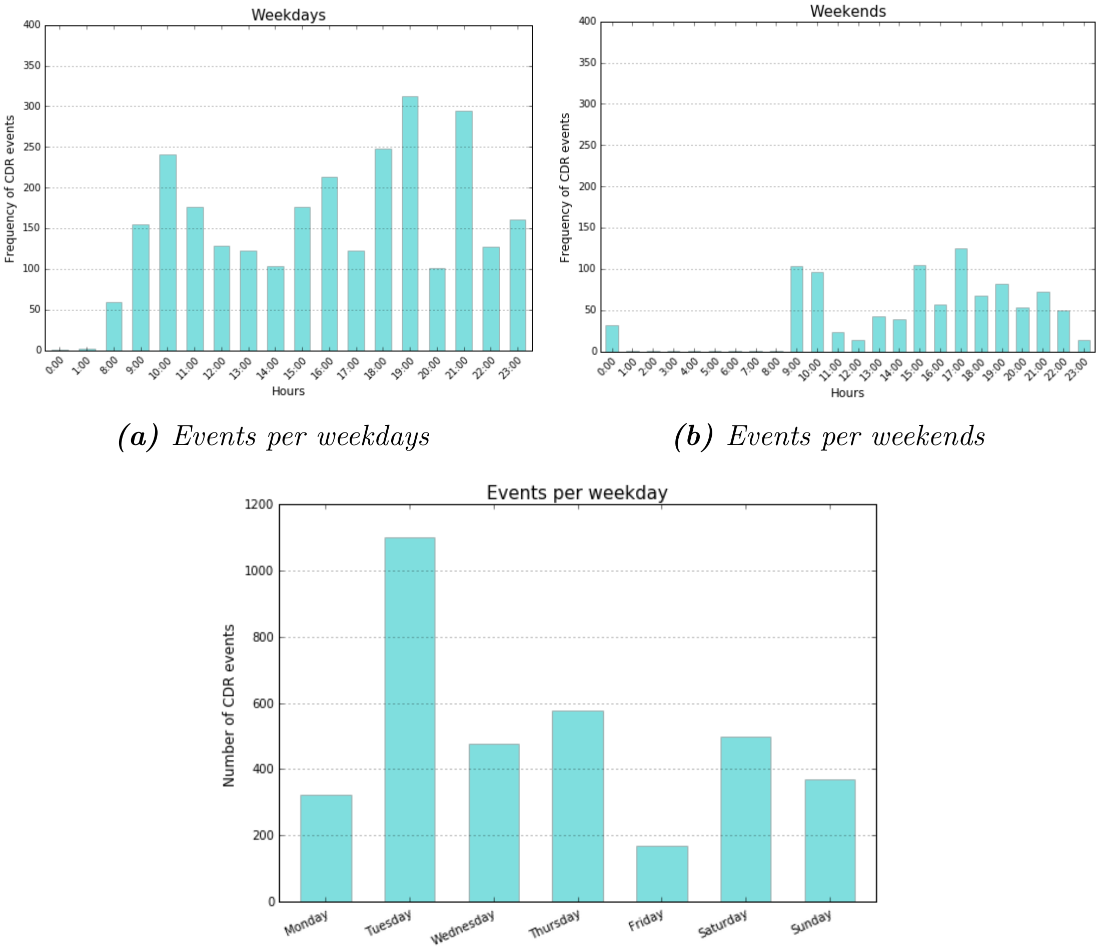
Figure 9: *Cell 3 centroid and the real starting position of the object.*

4 Human mobility patterns

In this thesis individual user mobility patterns are investigated. In *Chapers 4.1* and *4.2* the distribution of user generated CDR events and distances traveled are visualized with time intervals in histograms. To find these patterns in the frequency of the CDR events and the distribution of distances traveled many methods from *Chapter 3*, such as using road data from OSM and calculating distances. After that the user's most frequently visited POIs are identified. In the last chapter CDR cell tower connecting issue is described. The inspected CDRs start from the date 16th of October, 2016 and end in 24th of November, 2016. In total there were 3721 CDRs.

4.1 Frequency of CDR events

Visualization of the CDR event frequencies gives an overview of time periods when the user is most active during the day or week. In *Figures 10a* and *10b* the frequency of the CDR events is shown during workdays and weekends. The number of users events over business days is a lot higher than in weekends. This might be due to the fact that the user works during the week and weekends are leisure time. The first peak in *Figure 10a* starts around 9 a.m. and ends around 11 a.m, these are the usual commuting to work hours. There is another smaller peak in the lunch time. The biggest peak is during the after office hours from 6 to 7 in the evening, when the user might leave work. There is a smaller peak around 9 p.m., which might indicate some leisure activity for the user. The weekends in *Figure 10b* have far less activity and the day starts from 9 a.m. to 11 a.m. with the first peak. Overall the events in weekends are evenly distributed and low. In *Figure 10c* the frequency of events are shown by weekdays. The busiest days of the week are Tuesdays and Saturdays. Weekend days have a lot less activity, with Sundays having the smallest frequency of events. In *Appendix 5.1* the daily event frequencies of the weekend days are shown. Mostly, the daily histograms conform the overall patterns.



***(a)** Events per weekdays*

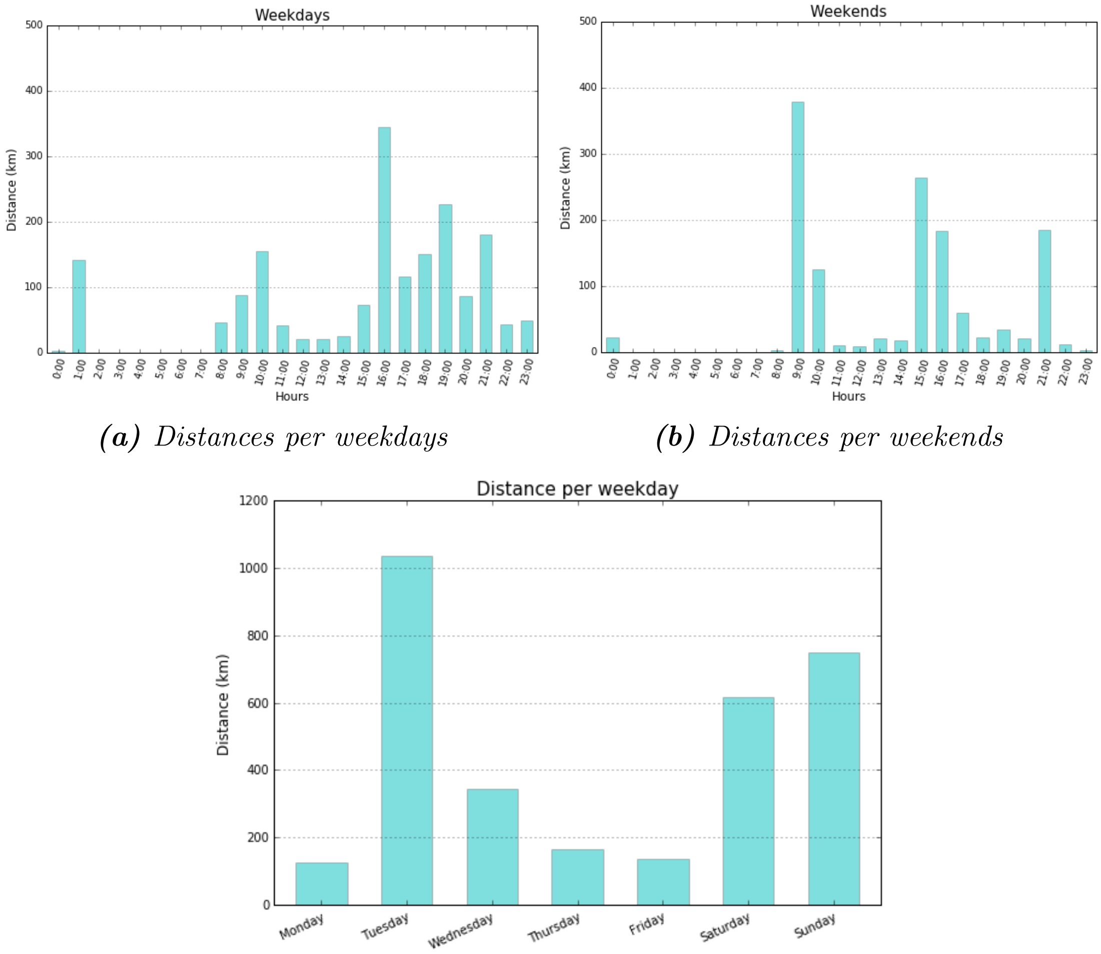
***(b)** Events per weekends*

***(c)** Events per weekday*

***Figure 10:** Histograms of event frequencies for the entire dataset.*

4.2 CDR dispersion over distance

Plotting the distances of the CDR events in chronological order gives an overview of the times when the user is travels the most distances during the day or week. In *Figures 11a* and *11b* the distance dispersion is shown during workdays and weekends. There are many similarities between the event and the distance distribution histograms. The user moves a lot more during weekdays than weekends. In weekdays there are three noticeable peaks. The smallest is from 8 a.m. to 9 a.am., which can be the time interval when the user travels to work. Second one is 4 p.m. and the third one 7 p.m. These hours might be when office workers usually goes home from work and complete other daily obligations (e.g. shopping, visiting friends). In the evening at 9 p.m. there is a peak, which can indicate to recreational activities. The peak at 1 a.m. is very unusual. *Figure 10c* confirms that the user travels more during the weekdays and less during weekends. The daily distance distribution is shown in *Appendix 5.1*. The daily frequencies in the *Figure 19* are similar to the overall distance distribution during the weekends, which shows that the user does not travel a lot in the days off from work.



(a) *Distances per weekdays*

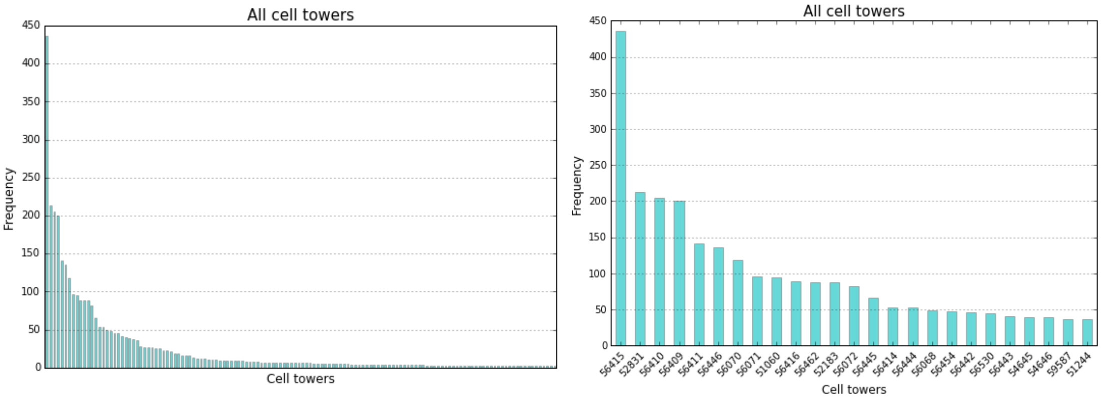
(b) *Distances per weekends*

(c) *Distances per weekday*

Figure 11: *Histograms of distance distribution for the entire dataset.*

4.3 User POI

To find the user's POI the CDRs were analyzed. In the data there were 883 unique cell IDs that represent sell towers the user had been connected to at some point of the time. In *Figure 12a* all of the connections are shown and in *Figure 12b* the 18 most frequently connected towers are shown. Around 763 towers had connections less than 10 times, which means these areas were most likely visited only once. This means that these cell towers were used, when the user was in motion.



(a) *All the cell tower connections.*

(b) *Most popular cell towers.*

Figure 12: *Cell tower connections.*

4.3.1 Home and work

The most frequently connected cell towers were analyzed to find the home and work areas of the user. From *Figure 12b* the ten first cell locations were investigated, to find the area where the user most likely lives and works. Most frequently visited cell location polygons (cell coverage areas) were extracted and visualized. The cell coverage areas that were very close together or even overlapped were considered one area of interest. Area with the most frequently visited cell ID is considered to be the area where the user lives. Because there was insufficient location data for four of the most frequently visited cells additional tasks had to be made. To confirm the home and work place hypothesis the timestamps were extracted and compared. The cell tower that was consistently connected to first in the mornings and last in the evenings was considered the home location and the second area was the place of work. The resulting home and work areas are shown in *Figure 13*, where the user's home seems to be in the south part of Tartu, in an area called Karlova. Because the original home area was very large, topographical attributes were considered, to crop the area to be even more precise. For example the Tartu river and big shopping center were left out, because there are not any residential areas in this region. In addition, the workplace seems to be in Tartu city center. Due to the fact that there are many polygons overlapping around the work area, there can be other POI in that area. This hypothesis is further investigated in *Chapter 4.3.2*.

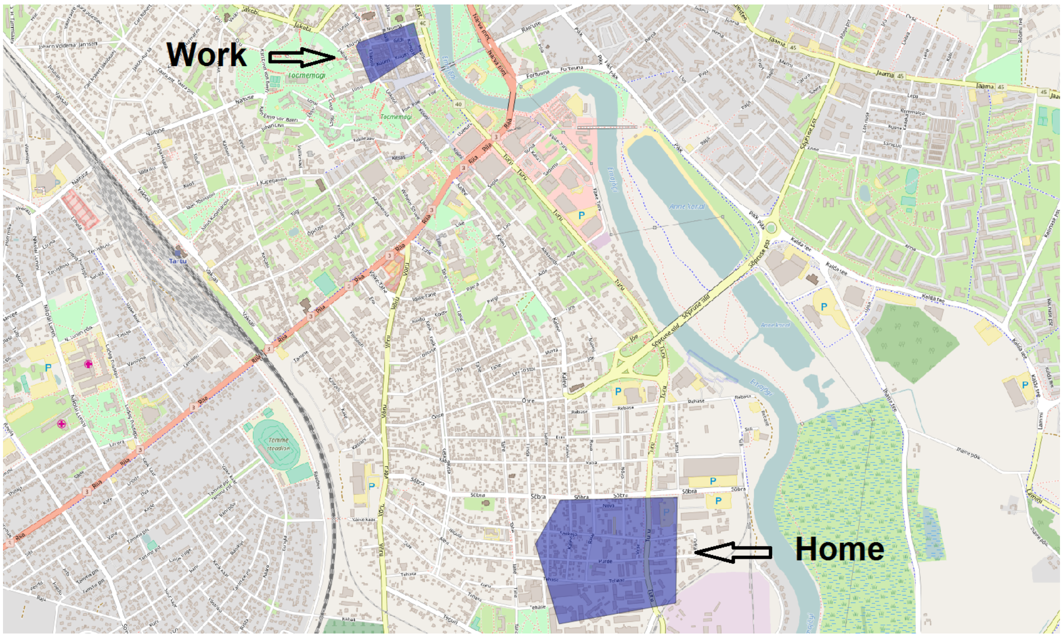
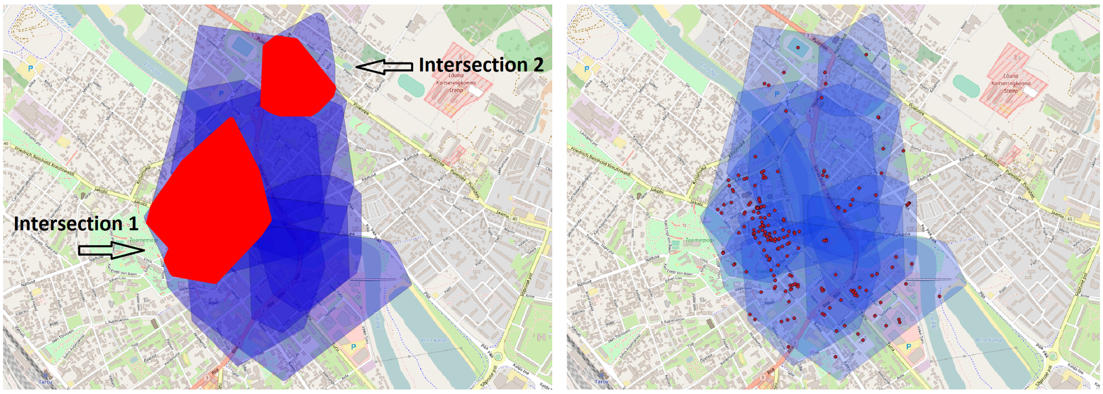


Figure 13: *Work and home areas shown in blue polygons.*

4.3.2 Additional POIs

To detect user's other POI the frequently visited areas were furthermore examined and intersections were generated. The overlapping areas represent potential locations, where the user has another POI. The two intersections found are presented in *Figure 14a* with the color red and the frequently visited areas are blue. The bigger overlapping area seems to be due to the fact that it is the area, where the user works. Second intersection on the other hand seems to be a POI. To confirm this the possible locations of stay, such as shops, gyms and restaurants were extracted from OSM service. The result is shown in *Figure 14b*, where the frequently visited areas are in blue and locations of potential POI are in red.



(a) *Intersections of frequently visited areas. Intersections are shown in red and frequently visited areas are in blue.*

(b) *Most visited areas with POI from OSM. Most visited areas are shown in blue polygons and POI are in red.*

Figure 14: *Frequently visited areas.*

Determining the exact location, that the user visits in the second intersection, timestamps from the CDRs were investigated again, to find time intervals of the stays. Results showed that the user usually arrived to the area between 6 p.m. and 7 p.m. and left around 9 in the evening. This means that the user stayed in the area for over three hours per every visit. The OSM data about potential locations, where the user could visit, consisted of two retail establishments, three food shops or restaurants, two sports locations (gym, stadium) and an educational building. The retail and education locations were left out because these institutions are closed during the visitation time interval and the sport and food businesses remain. After examining opening hours of the remaining locations only three remained, two sporting establishments and a food market. Due to the fact that visits were regular and mostly made on Tuesdays and Wednesdays (also some Sundays), it is highly plausible that the user visited the sport center shown in *Figure 15*.

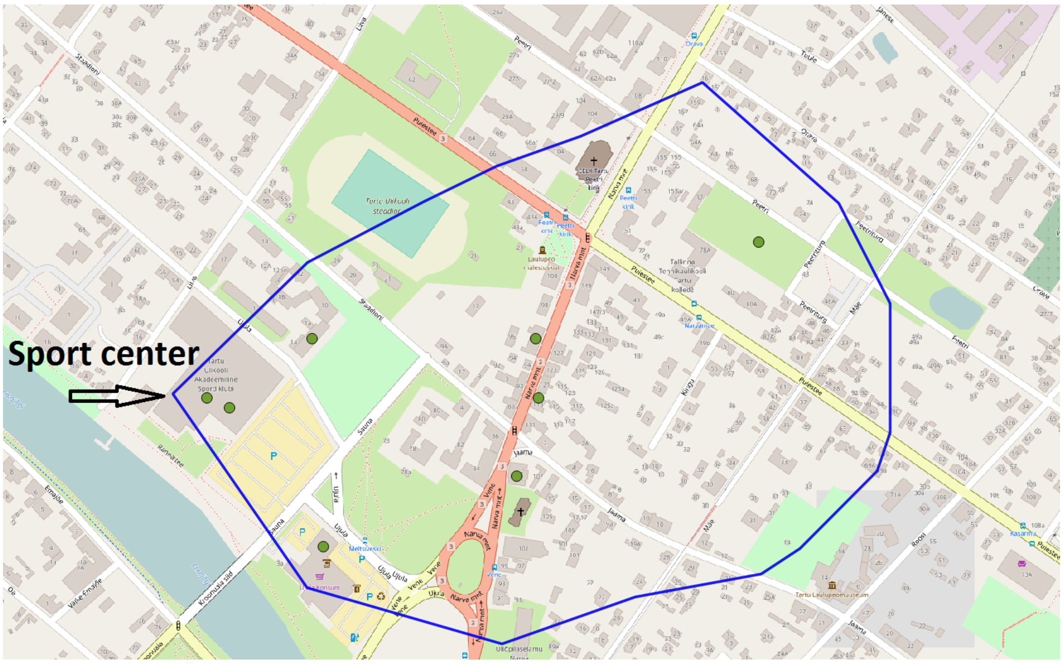


Figure 15: *POI inside the frequently visited area. Visited area is shown within a blue polygon line and POI are in green dots.*

5 Conclusion

Geospatial technology in the daily lives of people has become an essential commodity and many applications use different types of spatial data to make mundane tasks easier, such as making travel plans or finding a suitable path. The rise of these technologies enables massive geospatial data collection and analysis. The results can be used in optimizing transportation, urban planning, event detection and semantic studies. Numerous research shows the benefits in using spatial data, for example GPS or mobile phone (CDR) data, to improve the daily lives of citizens.

In this thesis, CDR data was used to reconstruct user paths. Due to the fact that CDR and GPS data have various differences many preprocessing phases were completed before beginning with trajectory reconstruction. For example CDR data does not include location data, but the information about the cell tower that it has been connected to. Location data was obtained by using an open source service and calculating the centroid of cell coverage areas. The center found as a result was appointed as user's location at a certain moment in time. The trajectory reconstruction results showed that the paths generated took on some cases many side streets or roads. This is because the real location of the user was in the edge of the coverage area and the result was less accurate. Comparing rural and urban areas showed that because in cities the number of cell towers is larger the paths generated were more accurate.

In addition to the trajectories, human mobility patterns were investigated, by analyzing one user's CDR data. The data consisted of CDRs starting from 16th of October, 2016 and ending with 24th of November, 2016. CDR distance and event distribution was visualized and patterns detected. Additionally user's significant places, such as living and work, were located from the intersections of frequently visited cell coverage areas. To find the user's recreational destinations times interval of the visits were compared to the points of interest in the coverage area. As a result the sport center, which the user visited regularly, was located.

5.1 Future work

There are many areas, where future work could improve the accuracy of trajectory reconstruction based on CDR data and there are various ways to do this. Some suggestions for the future work are in the list below.

- Starting and ending point detection could be improved, by giving attributes, such as *Stay*, *Jump* and *Move* to the cell coverage areas, where the object has been in. This could be done by using *Kalman* method with three ingrained movement models.

- The path choosing could be improved by taking into account public transport, in addition to the *walking* and *driving* times.

- Calculate more accurate average speeds by including times of traffic congestion in the city.

- Alternative path generating algorithm could be used to find more accurate trajectories.

Bibliography

[AA13] Natalia V. Andrienko and Gennady L. Andrienko. Visual analytics of movement: A rich palette of techniques to enable understanding. In Chiara Renso, Stefano Spaccapietra, and Esteban Zimnyi, editors, *Mobility Data*, pages 149–173. Cambridge University Press, 2013.

[BCDL$^+$13] Michele Berlingerio, Francesco Calabrese, Giusy Di Lorenzo, Rahul Nair, Fabio Pinelli, and Marco Luca Sbodio. *AllAboard: A System for Exploring Urban Mobility and Optimizing Public Transport Using Cellphone Data*, pages 663–666. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[BCH$^+$] Richard A. Becker, Ramn Cceres, Karrie Hanson, Ji Meng Loh, Simon Urbanek, Alexander Varshavsky, and Chris Volinsky. Clustering anonymized mobile call detail records to find usage groups. *AT&T Labs - Research.*

[BCH$^+$11] Richard A. Becker, Ramón Cáceres, Karrie Hanson, Ji Meng Loh, Simon Urbanek, Alexander Varshavsky, and Chris Volinsky. A tale of one city: Using cellular network data for urban planning. *IEEE Pervasive Computing*, 10(4):18–26, 2011.

[BHLV15] Oleg Batrashev, Amnir Hadachi, Artjom Lind, and Eero Vainikko. Mobility episode detection from cdr's data using switching kalman filter. In *Proceedings of the Fourth ACM SIGSPATIAL International Workshop on Mobile Geographic Information Systems*, MobiGIS '15, pages 63–69, New York, NY, USA, 2015. ACM.

[BLT$^+$11] Linus Bengtsson, Xin Lu, Anna Thorson, Richard Garfield, and Johan von Schreeb. Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: A post-earthquake geospatial study in haiti. *PLOS Medicine*, 8(8):1–9, 08 2011.

[cdr13] Call detail records. the use of mobile phone data to track and predict population displacement in disasters. *ACAPS*, 2013.

[CFT16] Carmela Comito, Deborah Falcone, and Domenico Talia. Mining human mobility patterns from social geo-tagged data. *Pervasive and Mobile Computing*, 33:91 – 107, 2016.

[CLC10] Ling Chen, Mingqi Lv, and Gencai Chen. A system for destination and future route prediction based on trajectory mining. *Pervasive and Mobile Computing*, 6(6):657 – 676, 2010. Special Issue PerCom 2010.

[CMS+16] Cynthia Chen, Jingtao Ma, Yusak Susilo, Yu Liu, and Menglin Wang. The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transportation Research Part C: Emerging Technologies*, 68:285 – 299, 2016.

[DL99] Guozhu Dong and Jinyan Li. Efficient mining of emerging patterns: Discovering trends and differences. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '99, pages 43–52, New York, NY, USA, 1999. ACM.

[DPG+15] Yuxiao Dong, Fabio Pinelli, Yiannis Gkoufas, Zubair Nabi, Francesco Calabrese, and Nitesh V. Chawla. *Inferring Unusual Crowd Events From Mobile Phone Call Detail Records*, pages 474–492. Springer International Publishing, Cham, 2015.

[DXZZ11] Ke Deng, Kexin Xie, Kevin Zheng, and Xiaofang Zhou. *Trajectory Indexing and Retrieval*, pages 35–60. Springer New York, New York, NY, 2011.

[DYGD15] J. Dai, B. Yang, C. Guo, and Z. Ding. Personalized route recommendation using big trajectory data. In *2015 IEEE 31st International Conference on Data Engineering*, pages 543–554, April 2015.

[FMV12] Vanessa Frias-Martinez and Jesus Virsesa. On the relationship between socio-economic factors and cell phone usage. *ICTD2012 Special Issue*, 9:35–50, 2012.

[FZ16] Zhenni Feng and Yanmin Zhu. A survey on trajectory data mining: Techniques and applications. *IEEE Access*, 4:2056–2067, 2016.

[GHB08] Marta C. Gonzalez, Cesar A. Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, June 2008.

[GISL16] Didem Gundogdu, Ozlem D. Incel, Albert A. Salah, and Bruno Lepri. Countrywide arrhythmia: emergency event detection using mobile phone data. *EPJ Data Science*, 5(1):25, 2016.

[GNP+11] Fosca Giannotti, Mirco Nanni, Dino Pedreschi, Fabio Pinelli, Chiara Renso, Salvatore Rinzivillo, and Roberto Trasarti. Unveiling the complexity of human mobility by querying and mining massive trajectory data. *The VLDB Journal*, 20(5):695, 2011.

[HW08] M. Haklay and P. Weber. Openstreetmap: User-generated street maps. *IEEE Pervasive Computing*, 7(4):12–18, Oct 2008.

[ict16] *Measuring the Information Society Report*, chapter Chapter 5. Measuring mobile uptake, pages 157–192. International Telecommunication Union, 2016.

[JYJ11] Hoyoung Jeung, Man Lung Yiu, and Christian S. Jensen. *Trajectory Pattern Mining*, pages 143–177. Springer New York, New York, NY, 2011.

[JZGR16] Karim Keramat Jahromi, Matteo Zignani, Sabrina Gaito, and Gian Paolo Rossi. Simulating human mobility patterns in urban areas. *Simulation Modelling Practice and Theory*, 62:137 – 156, 2016.

[Kru11] John Krumm. *Trajectory Analysis for Driving*, pages 213–241. Springer New York, New York, NY, 2011.

[LCC12] Mingqi Lv, Ling Chen, and Gencai Chen. Discovering personally semantic places from gps trajectories. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 1552–1556, New York, NY, USA, 2012. ACM.

[LK11] Wang-Chien Lee and John Krumm. *Trajectory Preprocessing*, pages 3–33. Springer New York, New York, NY, 2011.

[LQW15] Siyuan Liu, Qiang Qu, and Shuhui Wang. Rationality analytics from trajectories. *ACM Trans. Knowl. Discov. Data*, 10(1):10:1–10:22, July 2015.

[LV11] Dennis Luxen and Christian Vetter. Real-time routing with openstreetmap data. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '11, pages 513–516, New York, NY, USA, 2011. ACM.

[LWB$^+$13] Xin Lu, Erik Wetter, Nita Bharti, Andrew J. Tatem, and Linus Bengtsson. Approaching the limit of predictability in human mobility. *Scientific Reports*, 3, 2013.

[mob16] The mobile economy, 2016.

[MT16] Jean Damascène Mazimpaka and Sabine Timpf. Trajectory data mining: A review of methods and applications. *Journal of Spatial Information Science*, (13):61–99, 2016.

[NSL$^+$12] Anastasios Noulas, Salvatore Scellato, Renaud Lambiotte, Massimiliano Pontil, and Cecilia Mascolo. A tale of many cities: Universal patterns in human urban mobility. *PLOS ONE*, 7(5):1–10, 05 2012.

[PJZ$^+$16] Michela Papandrea, Karim Keramat Jahromi, Matteo Zignani, Sabrina Gaito, Silvia Giordano, and Gian Paolo Rossi. On the properties of human mobility. *Computer Communications*, 87:19 – 36, 2016.

[PSR$^+$13] Christine Parent, Stefano Spaccapietra, Chiara Renso, Gennady Andrienko, Natalia Andrienko, Vania Bogorny, Maria Luisa Damiani, Aris Gkoulalas-Divanis, Jose Macedo, Nikos Pelekis, Yannis Theodoridis, and Zhixian Yan. Semantic trajectories modeling and analysis. *ACM Comput. Surv.*, 45(4):42:1–42:32, August 2013.

[PT14] Nikos Pelekis and Yannis Theodoridis. *The Case of Big Mobility Data*, pages 211–231. Springer New York, New York, NY, 2014.

[Pul13] United Nations Global Pulse. Mobile phone network data for development. October 2013.

[RBdM$^+$13] Chiara Renso, Miriam Baglioni, Jose António F. de Macedo, Roberto Trasarti, and Monica Wachowicz. How you move reveals who you are: understanding human behavior by analyzing trajectory data. *Knowledge and Information Systems*, 37(2):331–362, 2013.

[RFPW06] Carlo Ratti, Dennis Frenchman, Riccardo Maria Pulselli, and Sarah Williams. Mobile landscapes: Using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design*, 33(5):727–748, 2006.

[SMC13] Christopher Smith, Afra Mashhadi, and Licia Capra. Ubiquitous sensing for mapping poverty in developing countries, 2013.

[SQBB10] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[Tra11] Goce Trajcevski. *Uncertainty in Spatial Trajectories*, pages 63–107. Springer New York, New York, NY, 2011.

[Vin75] T. Vincenty. *Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations*, chapter Survey review, pages 88–93. 1975.

[WET^{+}12] Amy Wesolowski, Nathan Eagle, Andrew J. Tatem, David L. Smith, Abdisalan M. Noor, Robert W. Snow, and Caroline O. Buckee. Quantifying the impact of human mobility on malaria. *Science*, 338:267–270, October 2012.

[ZCL^{+}10] Yu Zheng, Yukun Chen, Quannan Li, Xing Xie, and Wei-Ying Ma. Understanding transportation modes based on gps data for web applications. *ACM Trans. Web*, 4(1):1:1–1:36, January 2010.

[ZCWY14] Yu Zheng, Licia Capra, Ouri Wolfson, and Hai Yang. Urban computing: Concepts, methodologies, and applications. *ACM Transaction on Intelligent Systems and Technology*, October 2014.

[ZD12] Huiqi Zhang and Ram Dantu. *Event Detection Based on Call Detail Records*, pages 305–316. Springer London, London, 2012.

[Zel98] J. S. Zelek. Complete real-time path planning during sensor-based discovery. In *Proceedings. 1998 IEEE/RSJ International Conference on Intelligent Robots and Systems. Innovations in Theory, Practice and Applications (Cat. No.98CH36190)*, volume 3, pages 1399–1404 vol.3, Oct 1998.

[Zhe15] Yu Zheng. Trajectory data mining: An overview. *ACM Transactions on Intelligent Systems and Technology*, 6(3):29:1–29:41, May 2015.

[ZLWX08] Yu Zheng, Like Liu, Longhao Wang, and Xing Xie. Learning transportation mode from raw gps data for geographic applications on the web. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 247–256, New York, NY, USA, 2008. ACM.

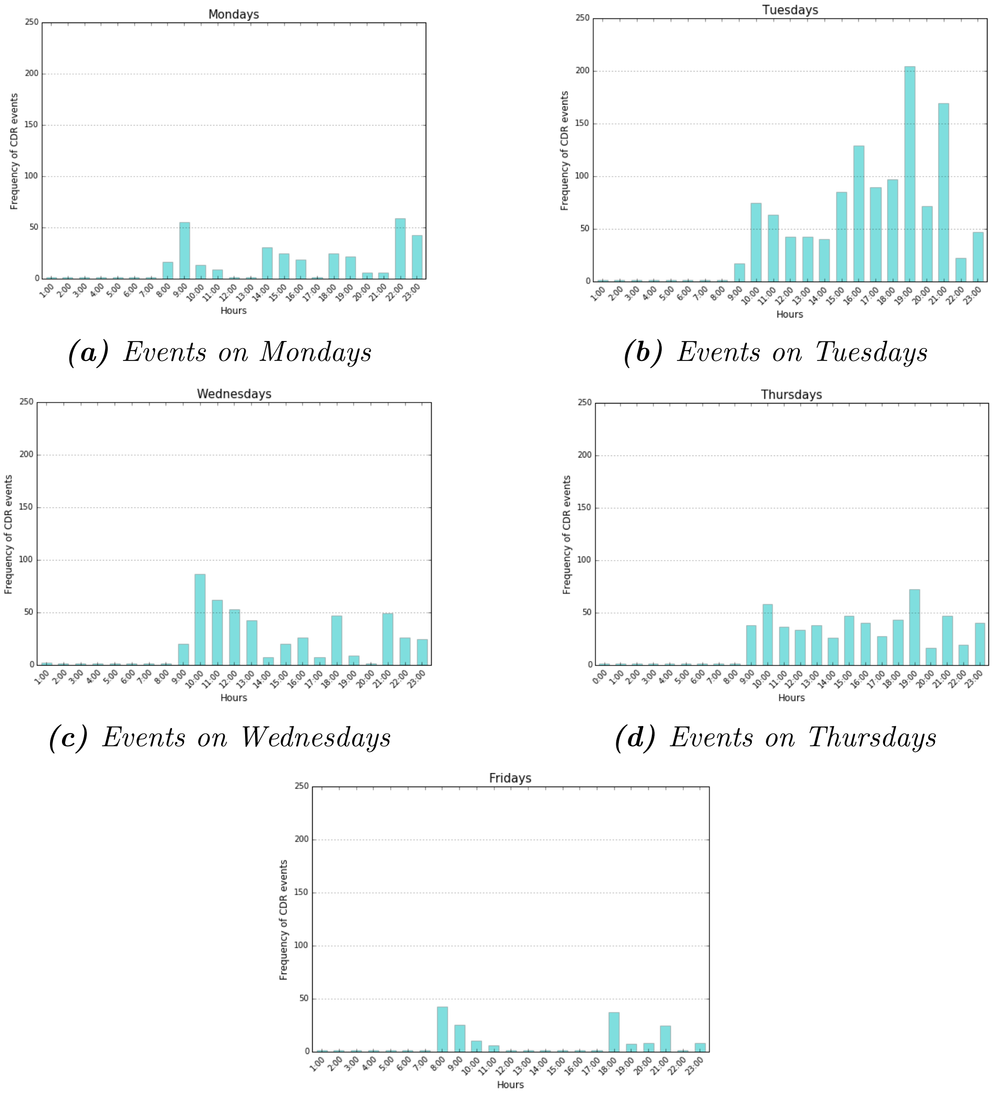
[ZZYS13] Kai Zheng, Yu Zheng, Nicholas Jing Yuan, and Shuo Shang. On discovery of gathering patterns from trajectories. ICDE 2013, April 2013.

Example of an OSM road data file

```
<?xml version="1.0" encoding="UTF 8"?>
<osm version="0.6" generator="CGImap 0.0.2">
 <bounds minlat="58.3774406" minlon="26.721533" maxlat="58.3790047" maxlon="26.7315473"/>
 <node id="377093405" lat="58.3791890" lon="26.7181400" user="Ivo" uid="46882" visible="true" version="1" changeset="676636" timestamp="2008 09 21T21:37:45Z"/>
 <node id="418065240" visible="true" version="4" changeset="48083103" timestamp="2017 04 24T08:17:25Z" user="CharlieHotelRomeo" uid="5718944" lat="58.3771593" lon="26.7271446">
  <tag k="crossing" v="traffic_signals"/>
  <tag k="highway" v="traffic_signals"/>
  <tag k="traffic_signals" v="crossing"/>
 </node>
 <way id="29980910" visible="true" version="6" changeset="14894434" timestamp="2013 02 03T11:27:20Z" user="k__" uid="156900">
  <nd ref="330042903"/>
  <nd ref="2139961606"/>
  <tag k="highway" v="residential"/>
  <tag k="name" v="Vallikraavi"/>
  <tag k="source" v="Tartu City Government"/>
 </way>
<relation id="2959823" visible="true" version="2" changeset="16333718" timestamp="2013 05 29T07:00:54Z" user="MHohmann" uid="129688">
  <member type="relation" ref="2959820" role=""/>
  <member type="relation" ref="2959819" role=""/>
  <tag k="network" v="Tartu linn"/>
  <tag k="operator" v="Sebe"/>
  <tag k="ref" v="2"/>
  <tag k="route_master" v="bus"/>
  <tag k="type" v="route_master"/>
 </relation>
</osm>
```

Listing 3: *Example of an OSM road data*

Event frequency daily



(a) *Events on Mondays*

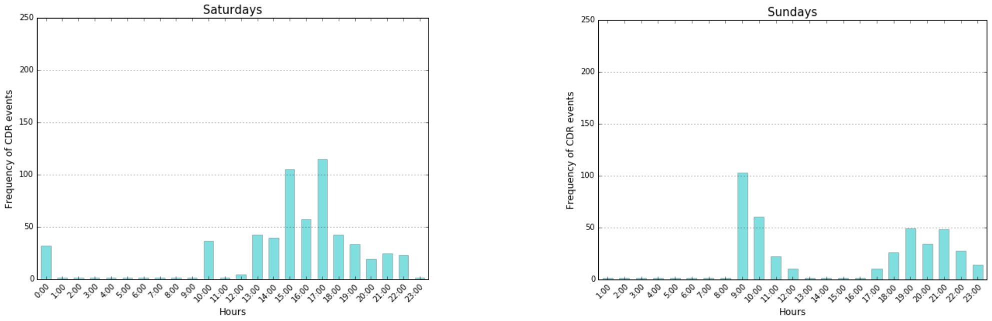
(b) *Events on Tuesdays*

(c) *Events on Wednesdays*

(d) *Events on Thursdays*

(e) *Events on Fridays*

Figure 16: *Event frequency during weekdays.*

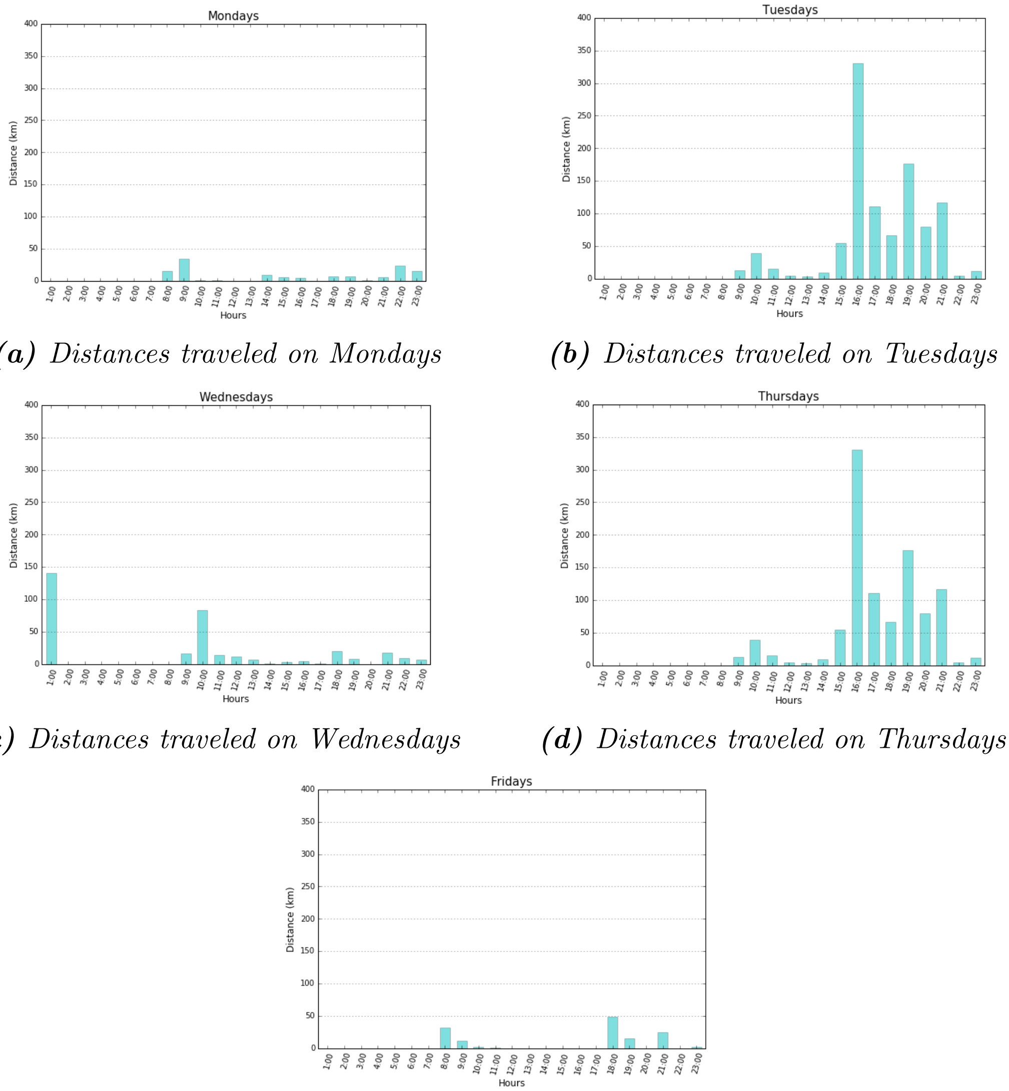


(a) *Events on Saturdays*

(b) *Events on Sundays*

Figure 17: *Event frequency during weekends.*

Distance distribution daily



(a) *Distances traveled on Mondays*

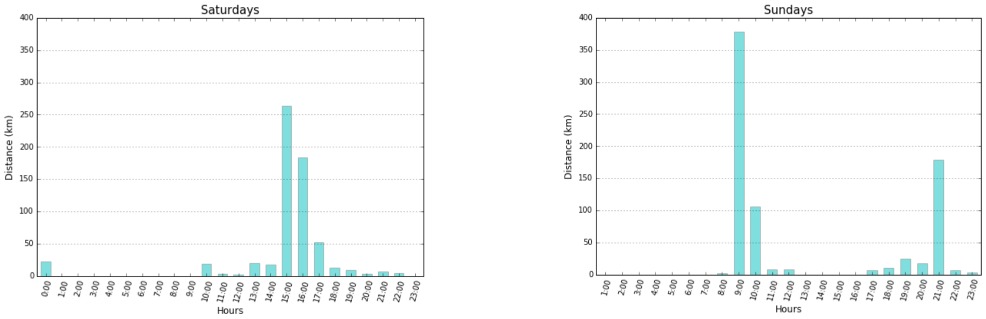
(b) *Distances traveled on Tuesdays*

(c) *Distances traveled on Wednesdays*

(d) *Distances traveled on Thursdays*

(e) *Distances traveled on Fridays*

Figure 18: *Distance distribution during weekdays.*



(a) *Distances traveled on Saturdays*

(b) *Distances traveled on Sundays*

Figure 19: *Distance distribution during weekends.*

Non-exclusive license to reproduce thesis and make thesis public

I, Annett Saarik,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to:

1.1 reproduce, for the purpose of preservation and making available to the public, including for addition to the DSpace digital archives until expiry of the term of validity of the copyright, and

1.2 make available to the public via the web environment of the University of Tartu, including via the DSpace digital archives until expiry of the term of validity of the copyright,

Trajectory Reconstruction and Mobility Pattern Analysis Based on Call Detail Record Data

supervised by Amnir Hadachi, PhD

2. I am aware of the fact that the author retains these rights.

3. I certify that granting the non-exclusive license does not infringe the intellectual property rights or rights arising from the Personal Data Protection Act.

Tartu, 18.05.2017