UNIVERSITY OF TARTU

Institute of Computer Science
Computer Science Curriculum

Anna Bilmaijer

Urban Mobility Sensing Using CDRs

Bachelor's Thesis (9 ECTS)

Supervisor: Amnir Hadachi, PhD

Tartu 2017

Urban Mobility Sensing Using CDRs

Abstract:

The aim of this thesis is to show that CDR (Call Detail Record) data holds potential in analysing human mobility in urban areas. CDR data is continuously gathered by mobile service providers, and it can be used as relatively low-cost and endless source of human displacement. This work tries to find meaningful places like people's work and home locations based on their cellular activity. Finding clear patterns proved difficult because CDR data tends to be imprecise due to large coverage areas and handovers between cellular towers. Some common paths between people were recognised; however, finding clearer urban and suburban communities would need applying more methods. Analysing human mobility patterns gives an overview on how to better plan urban areas. It can benefit in transportation, improving cellular networks, and also give input for targeted advertisement.

Keywords: CDR, human mobility, Thiessen polygons, OpenLayers, cellular network, mobility data

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

Linnasisese liikuvuse uurimine mobiilside andmete põhjal

Lühikokkuvõte:

Selle töö eesmärk on näidata, et mobiilside andmetel on potentsiaali inimeste liikumise uurimiseks linnades. Mobiilside andmeid kogutakse pidevalt mobiilsideoperaatorite poolt ning seda on võimalik kasutada kui pidevat inimeste liikumise allikat. See töö püüab leida inimeste mobiilside aktiivsuse põhjal nende tähtsad kohad nagu kodu ja töö. Täpsete liikumismustrite leidmine osutus raskeks, kuna mobiilside andmed on ebatäpsed suurte sidemastide katvusalade tõttu ning kuna ühest asukohast tulenevaid kõnesid võib vastu võtta üks mitmest seda kohta teenindavatest mastidest. Leiti mõningased mustrid erinevate inimeste liikumises, kuid täpsemate kesklinna ning äärelinna kommuunide leidmiseks on vaja rakendada rohkem meetodeid. Inimeste liikumise uurimine annab ülevaate, kuidas saab paremini linnasid planeerida. Sellest võib kasu olla transpordivõrgustiku arendamises, mobiilsidevõrkude parandamises, ning lisaks annab võimaluse kindlatele sihtrühmadele reklaami levitamiseks.

Võtmesõnad: CDR, inimeste liikumine, Thiesseni hulknurgad, OpenLayers, mobiilsidevõrk, liikuvuse andmed

CERCS: P170 Arvutiteadus, arvutusmeetodid, süsteemid, juhtimine

Acknowledgements

I am very grateful to my supervisor, Amnir Hadachi, for all his support and guidance. He was there for me from day one and I really appreciate all he has done to help with this thesis.

A big thank you to Priit Salumaa for his help in the search for the topic.

I am thankful to Tele2 Eesti for providing the data that was used in this work.

Contents

List of Abbreviations

CDR:	Call Detail Record
GPS:	Global Positioning System
IMSI:	International Mobile Subscriber Identity
MSISDN:	Mobile Station International Subscriber Directory Number
IMEI:	International Mobile Equipment Identity

Chapter 1

Introduction

In Estonia, almost every grownup has a cell phone. In 2012, 84.2% of the population, aged 16-74 used mobile device [stab]. Each call from a mobile phone user generates a CDR, a timestamped data with information about the cell tower, that processed the request, and also contains information about the user, user's device, event type and more. This information varies from service provider to service provider. Every second a large amount of data about human mobility is generated without interference from third parties. This means, there is big data just waiting to be used.

Many companies try to use this kind of data to accomplish targeted advertisement. In 2013, a UK mobile service provider tried to locate Manchester United fans to send them game vouchers [itl16].

Another use for cellular network providers would be to better plan their network. When analysing human mobility, it is possible to extract movement patterns and also how people are distributed during a big event such as sports games and festivals. This would help in finding locations where cellular towers are or will be overloaded and fix the problem to offer better service for clients.

Also, traffic planning can be done using CDR data. Knowing how people move, it could be possible to avoid traffic congestion during big events or know how most effectively evacuate part of or the whole city by reducing the number of vehicles on the busiest roads.

The problem with this kind of data is that it is imprecise. Meaning, that when a person makes a call, and it gets registered by a cell tower, then makes another call and it gets processed by another tower in another location, it does not mean that the person has moved. That person could have been just standing in an area where multiple cell tower coverage areas overlap. If one cell tower is processing too many events, the call could be picked up by another one to reduce the load on the first tower. This kind of phenomenon is known as handover.

Another challenge is that the user could be positioned anywhere in the tower coverage area. The person could be standing close to the actual tower position, or he or she could be at the edge of the coverage area, which for some towers could be 2 kilometres or more from the real tower position. Some coverage areas are smaller, giving a more precise caller location, and some bigger, for which the user's location can not be determined so easy.

Also, some people do not use their mobile devices frequently. This means that after positioning a person, he or she might not make a call for a long time and there is no way of saying where that person is based purely on CDR.

Although there are some limitations, the CDR data is valuable compared to other methods used. GPS is more precise considering spatial data, but that data is not so easily accessible as people need to give permission to use it and also not a lot of people use GPS devices. For better results, GPS data can be complementary to CDR for validating the results. Another popular way to gather information about population throughout times has been different questionnaires. This method is most precise as people themselves tell about their life habits, their usual locations and so on. However, planning and conducting such activity is very time consuming and expensive. Moreover, it does not track when people change their patterns and is not available in real time. CDR data, despite its imprecision, is very attractive as it can be applied in various fields and no extra effort is needed. All the data is already being collected by cellular service providers.

This thesis work gives a glimpse of what can be done with CDR data. Moreover, we will summarise how the CDR data can be used in analysing human mobility through the literature. Next, we will introduce our contribution in using the CDR data for discovering meaningful location such as home and work locations. Finally, we will discuss and analyse the results obtained and also conclude with a convulsion and future prospectives.

Chapter 2

State of the Art

2.1 Introduction

This chapter gives an overview of some articles that have been published on a similar topic. Human mobility has been a popular challenge in the past years as understanding communities and predicting their movements could help solve some urban problems. It could benefit in transportation, and also let service providers better plan and distribute cellular network coverage.

2.2 Similar Industrial Projects

Many researchers have been developing different methods to find people's home and work locations, some more successfully, some less. This chapter will cover two articles, that tried to solve a similar problem as this thesis.

In paper [ASJ+10] the authors presented a method on how to extract meaningful places for mobile phone users. The dataset was offered by EMT, Estonia's large telecom company. They used 12 months of data with more than 0.5 million anonymous users as their test base.

They processed data in eight steps. First, they differentiated regular calls from random calls. Random calls were distinguished when the network cell used for that call has been used only once per month, and those calls were extracted from further analysis. Next, they removed people with too few or too many calls. For every person, they found the cells which were used most frequently for the calls.

For some people, the most frequent cell was used less than seven days in a month. Those people were removed from the analysis as the did not give enough data to make valid assumptions about their meaningful points. Also if there were too frequent calls made by a person, that person was also excluded as it may have been a service centre or a technical device that used data. In the third step, they analysed two most frequent cells and tried to separate one cell as a home anchor and one as work anchor. The separation was done based on the time of day when the cell under observation was usually used for calls. The fourth stage tried to find out if the two cells separated in the previous step were two home or two work anchors. If these points were not in neighbouring cells, then it was concluded that they represented different anchors and they were moved to the next step. If they were neighbouring, the model took next frequent cell and re-analysed it in step three. There was a possibility that the model concluded that there were two home or two work anchors. If after n iterations of analysing the next most frequent cell the model still could not find the missing anchor point, the model moved on to stage 5 with two home or two work anchors. Next stage determined if a person had multiple anchor points or the cell acted as both home and work anchor. For this, it was observed whether or not the most frequent cell covers more than 75% of overall cells used. If it did, the second frequent cell was dismissed, and the model went on with one home-work multifunctional anchor. By the end of the fifth phase, the model had either found one home and one work anchor point, one multifunctional anchor point, two home anchor points or two work anchor points. First two moved on to eight stage, last two went ahead with the sixth stage. The sixth stage consisted of trying to find missing point again from next most frequent cell. Again, there was a possibility of only one anchor point or two same anchors. Next step tried to classify the missing point again. Regardless whether the last two steps were successful or not, the model went on to the last phase. That phase classified the points as one home and one work anchor, one multifunctional anchor, two home, and one work anchor, or one home and two work anchors. The rest of the cells were classified as secondary anchor points.

In [ASJ+10], the initial dataset consisted of 592 553 unique user IDs. After cleaning up unsuitable cells using previously mentioned steps, the number of IDs got reduced to 449 793. 282 572 people had both home and work anchor points, 178 458 had multifunctional anchor points, 10 777 had two home and one work anchor point, and 13 065 had one home and two work anchor points. After comparing the retrieved home-work anchor points with the Population Register, the authors pointed out that in most of the cities there were fewer home anchor points in the derived data than registered homes. For example, Population Register showed there was 39.4% of Estonian population living in Tallinn in 2007. The percent of home anchors was 38.4. The differences were about 1-3% for all cities and counties compared to the results from the article. Overall, the percentage weights by county from the register matched the home-anchor point percent.

The authors in [KGSR14] also tried to map people's home and work locations using CDR. Their data cleanup consisted of a bit different methods than in article mentioned before. The authors sampled data at 10-minute intervals, and if the time between events was more than 10 minutes, it was assumed that the user stayed in the same cell for the whole period because overlapping cells can lead to cell handover without the user changing position. They also used GPS data as one source of information. To make GPS data comparable to CDR, the observable city was divided into 0.5 x 0.5-kilometre grid system, each square representing a cell. To analyse the cells, the writers calculated the duration of how long each individual stayed in a cell. The cell where a user stayed the most would be ranked the highest, and so on. In the case of GPS dataset, the highest ranked grid was the one user visited the most. They then extracted the calls made on weekends as people normally do not work on these days. Too infrequent calls that were made less than 16 hours apart were also eliminated from further analysis as they did not give any meaningful information. They set 8 a.m. as the start of the daytime and 8 p.m. as the end of the daytime and start of the night. This means calls made between 8 a.m. and 8 p.m. were counted as daytime calls and nighttime calls otherwise. The home location was assigned to a cell that the person spent most of the time during night time, and work location would be the most frequent daytime cell. To count them as such cells, the user had to spend at least 50% of the day- or nighttime in that location. This validation gave a bigger chance of finding the actual home and work locations. If there were no such places for the user, the user was not considered any further. Only about 7-11% of users from the initial dataset were chosen for final analysis due to the strict filtering. GPS data was similarly filtered to find home and work location estimates. To compute the distance between found locations they used the crow-fly distance as the CDR data is not that precise and other more complicated methods might not give better results. After that, they tried to find the times when people moved between home and work. For that, they found the last call from a home cell and first call from work cell before noon, and also the last call from work and first call from home after noon. These two differences gave them morning and evening commuting times respectively. Also, they filtered out callers that produced cell events less that one call per hour. This left them with only 5% of the initial dataset. But considering a large number of callers in the original dataset, they still had many thousands of users left after all the strict filtering.

In the end, the results were analysed in many ways. As a start, they drew a graph for each city for how much time people spend in each cell, starting from the highest ranked cell. It was observed that during daytime the distribution follows Zipf's law [Hos09] while during night time there is a significant fall at about rank 10, giving distribution a sigmoidal shape. From this, the authors concluded that during daytime people visit various places quite often, while during night time they stick to most common ones. The next observation they made based on commuting distance is that most people live less than 10 kilometres from work. For the third graph, they mapped commute distances to their respected user and showed how these distance frequencies depended on the time of day. It showed that for most people the time they moved from home to work was between 5 a.m. and 10 a.m., and from work, to home, it was not as clear, presumably because people tend to stay out after work.

2.3 Human mobility patters

Observing population's home-work movements throughout the day is only one subject of interest for the companies that search for potential in CDR data. Next articles give an overview of some other observations of human mobility patterns made using mobile phone data.

In [FGNG12] the authors gathered human mobility data with rich context information. They ran an experiment to collect data about contacts between humans. Each participant filled out a questionnaire to discover their profile information. This gave an overview of each person based on location properties like their institute, office and lunch place locations, and social properties like their job position, what language they speak, do they smoke, and more. During experiment, they captured each contact between the participants in three weeks duration. The authors defined a new metric called social distance and predicted a number of contacts between people. The conclusion was that the people with same social properties tend to meet more. For example, if only people who speak the same language are considered, the mean of interactions grows by 18%. This was only an experiment with a small group of people, but the results were promising as they already showed the correlation between social ties and mobility.

A paper [CPDL^{+}10] was written on observing the human movements during big social events. The hypothesis was that different event types attract people from different parts of the city. This hypothesis was based on an assumption that people interest stay almost the same over time and keep visiting the same kind of events. They concluded that people living close to the event area are more likely to be interested in the event. Also, people from close origins are more attracted to the same type of events. This kind of study could help predict movements of big groups of people during events resulting in better traffic organisation and also planning emergency situations.

Chapter 3

Methodology

3.1 Data Description

The dataset was provided by Tele2, one of the largest telecommunication providers in Estonia, and was received as messages over specified period. When a person makes a call, one of the nearby cell towers picks up the request for processing. This tower is not necessarily the closest one, but the caller has to be in that cell's coverage area. These events processed by the towers are collected into messages. Each message starts with a timestamp followed by usually 3-6 events that occurred at that time. Each event consists of the following information: hashed IMSI, hashed MSISDN, IMEI, event type and the cellular tower ID, which processed the event. Event type can be calling, sending and receiving a text message, using the Internet, switching towers and other. All these events are produced by Tele2 clients and are anonymous. In this work to distinguish different people the hashed MSISDN number is used. Although multiple MSISDNs can belong to the same person, in this work, the term person or user is used as a reference to this number holder. In total there are about 300 000 unique MSISDNs in the dataset. Moreover, the study in this thesis will focus on urban areas. Hence, geographical area was limited to the city of Tartu.

3.2 Processing the Data

As with every large dataset, it is not known from the beginning what that data holds. CDR data can be imprecise as there are overlapping cellular coverages and some cell coverages are very large meaning that during an event in some cells the person could have been anywhere in a 2 or more kilometre radius. In this thesis, a term *cell* is used to refer to a cellular tower's coverage area. For visualisation, an OpenLayers [ope] map is used.

To better understand how cell towers are positioned, *Figure 3.1* gives an overview of all cellular coverage area centres in Estonia. There are bigger cell densities around larger cities as the population is higher in those areas *Figure 3.2* [staa] and so the cellular load is increased. Looking at *Figure 3.3*, we see that on the city level, in Tartu the cells are more frequent in the city centre, due to the density of population (*Figure 3.4*) in more populated areas that need to be all the time connected to the network. In total there are 337 cell towers in the city of Tartu [tar]. The cellular network usually consists of macro-cellular networks and small-cellular networks, like micro- and pico-cell nets [JKA11]. Cells in a macro-cell network are large compared to smaller ones, usually ranging from 1 to 20 kilometres. Micro and picocells are ranging from a few metres to around a kilometre. These small cells are used to cover the urban areas on a street or a building level, where cellular activity is higher, and the network needs to be always online and processing the events. Macro cells are usually found in rural areas where the traffic is low, or in the city, covering the areas where small cells do not reach. *Figure 3.5* shows most of the real coverage polygons in Tartu. This detailed cellular tower data is a bit imprecise, and there are some polygons missing; however, it gives us an idea of how the polygons look like, and also it helps to discover that for some cellular towers the coverage area can consist of multiple polygons. Using macro-cells to position people would give a too broad estimation of person's location, so the cells with large coverage areas were extracted from Tartu's cell list. After removing cellular towers, that have large coverage area, from the data, we get *Figure 3.6* showing a bit less overlapping and smaller coverage areas. The number of cell tower got down to 262. Comparing *Figure 3.5* and *Figure 3.6* gives an example, that macro-cells cover mostly the suburban areas and gaps between micro-cells. Micro-cells and pico-cells are located more in the city centre.

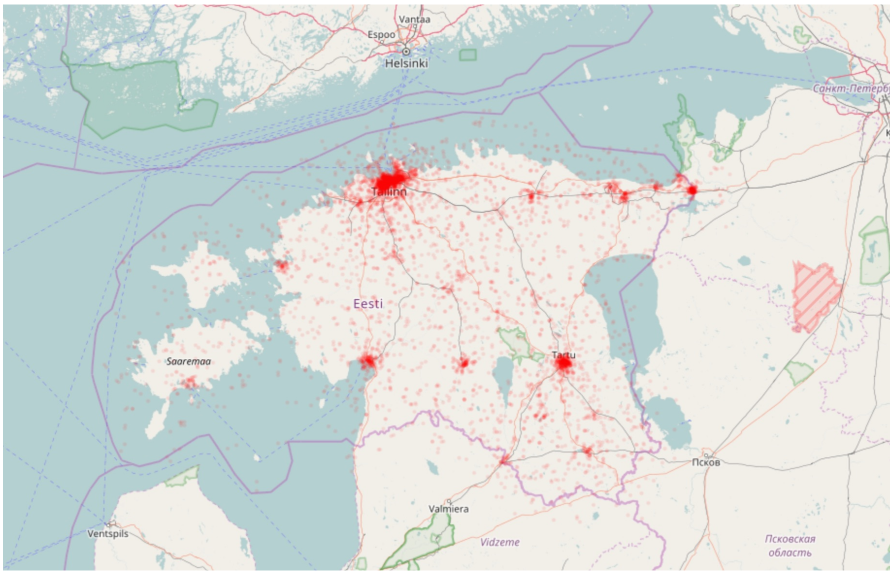


Figure 3.1: Location of cellular coverage area centroids in Estonia.

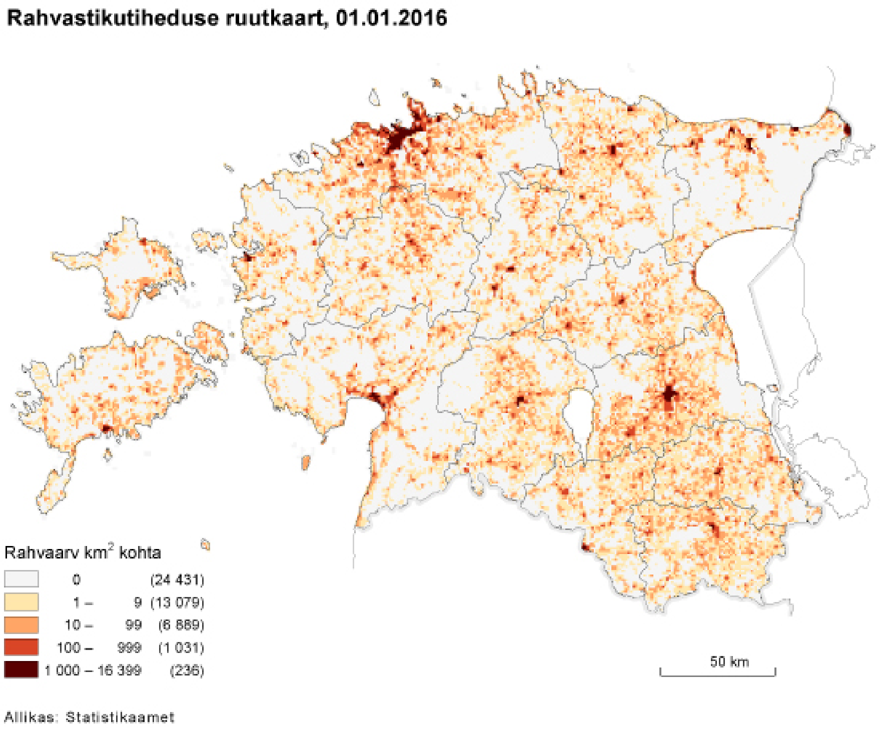


Figure 3.2: Estonian population density in 2016.

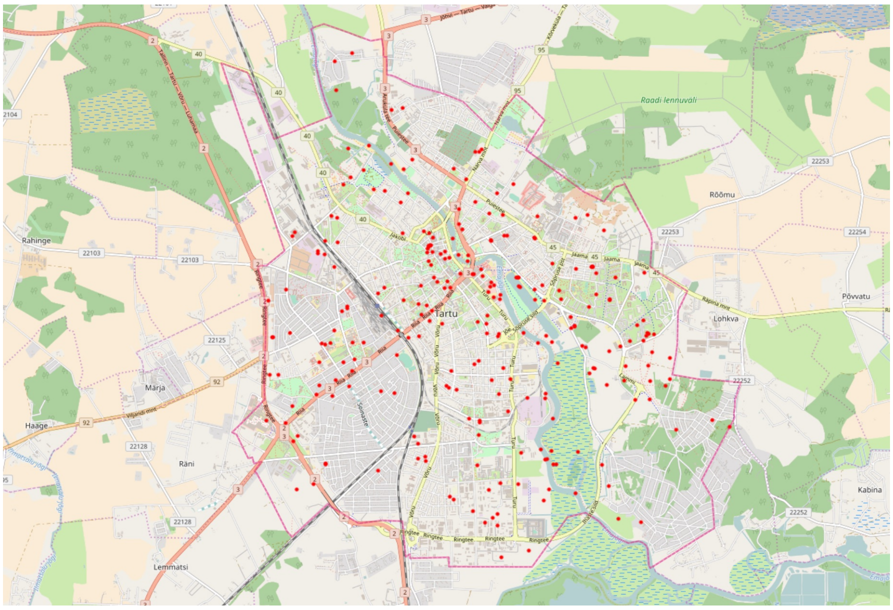
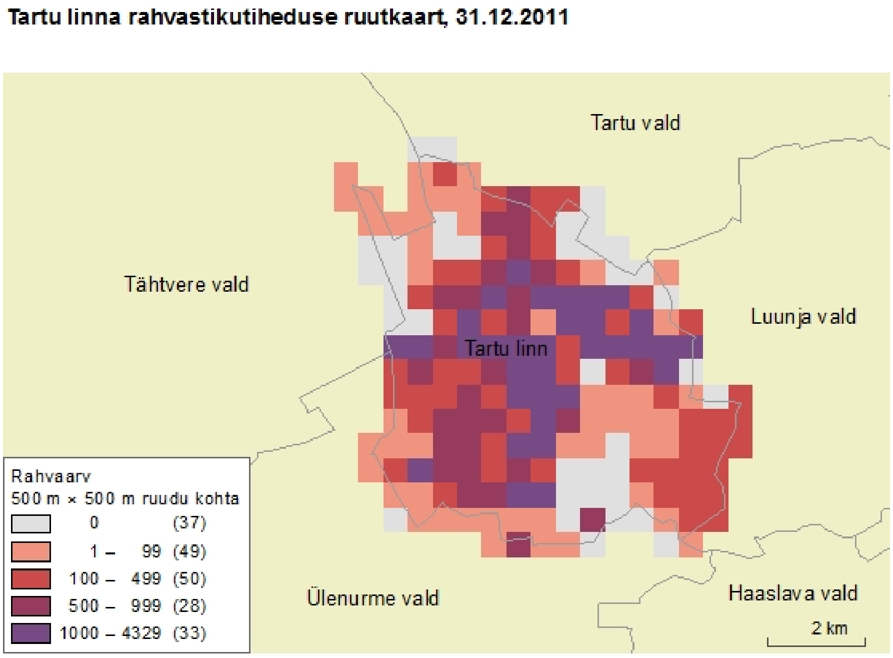


Figure 3.3: Location of cellular coverage area centroids in Tartu, Estonia.

Overlapping polygons do not give a good visualisation of people's locations, so this thesis uses Thiessen polygons [Yam16]. Polygon set is created using third party tool [vor]. This tool takes an array of two-dimensional points as input and using Fortune's algorithm [For87] outputs polygons that perfectly cover the given space



Allikas: Statistikaamet

Figure 3.4: Tartu population density in 2016.

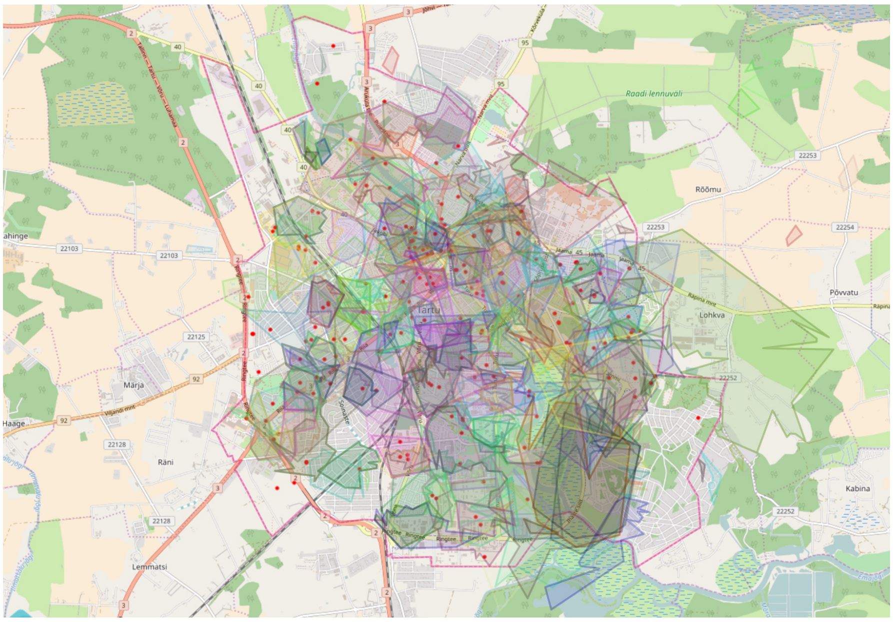


Figure 3.5: Cellular coverage areas in Tartu based on available information.

without overlaps. 262 smaller cellular coverage centres are used as an input, and the result is seen in *Figure 3.7*. Hereafter the term *cell* refers to a cellular tower's coverage area's representation on Thiessen polygons level.

Next step is to find events that occurred within the bounds of Tartu. A module

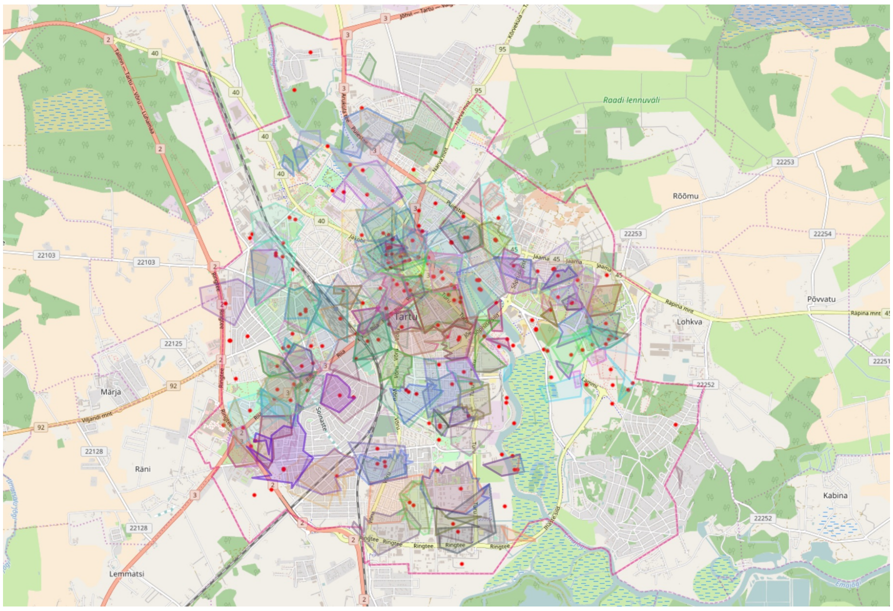


Figure 3.6: Cellular coverage areas in Tartu after removing cell towers with largest coverage areas.

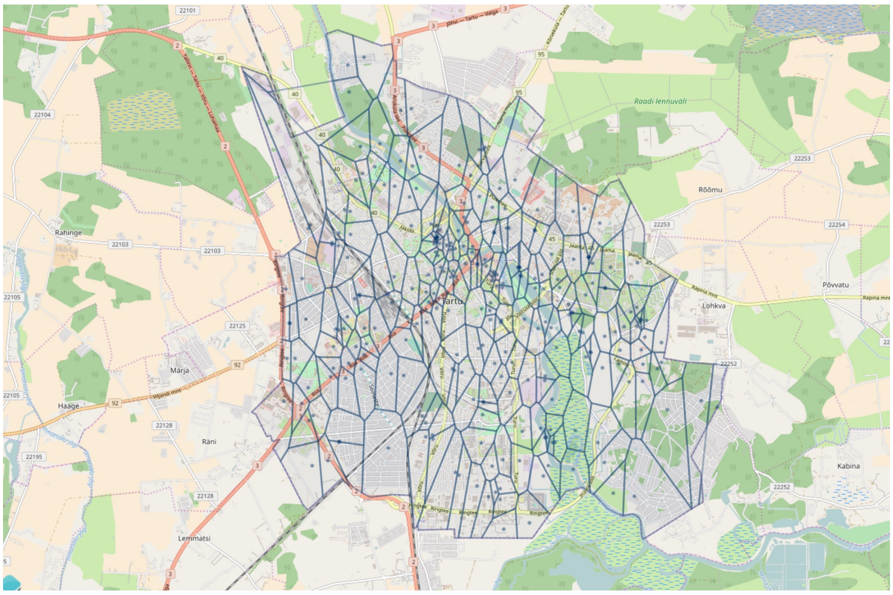


Figure 3.7: Thiessen polygons constructed from cell coverage area centres in Tartu.

[inP] that uses W. Randolph Franklin's algorithm [fra06] is used to determine if the cell's centre is inside Tartu's outline polygon. In total 77 098 282 events are parsed during one week of recording and 2 697 540 of those were in Tartu with 48 808 unique users.

There are many researches focusing on finding the most frequently used cells and using them as meaningful places. Those cells are then processed and using different methods, potential home and work places for each person are found. One of these methods, written by Ahas and others [ASJ+10], was described in the previous chapter. Unprocessed frequent cells do not give enough information, mostly because of the handover effect. During trying different approaches for this thesis, it was discovered that in many cases, two or more most frequent cells in which a person is observed, are overlapping cells, with almost the same count. It is not enough to take two most frequent cells and assign them to work and home locations. That was also what Ahas was trying to improve. This thesis offers a different approach to finding meaningful places like home and work.

To find work and home communities in Tartu, an assumption is made, that people move in the morning from home to work and in the evening they move the opposite way. The article [IBC+11] proposed an algorithm for finding important places based on different observable factors. They used real volunteers as their training set and also tested on volunteers so they could compare their results with actual data. As a conclusion, they found that for finding home location, the "Home Hour Events" factor was dominant in that algorithm. For finding work location, "Work Hour Events" was one of the dominant factors. The algorithm showed promising results achieving the median error of 1.45 kilometres from actual home location and 1.35 kilometres from work location. The "Home Hour Events" were tracked from 7 p.m. to 7 a.m. "Work Hour Events" were tracked from 1 p.m. to 5 p.m. The method proposed in this thesis takes into consideration the results concluded from the previously mentioned article.

In Estonia the usual time when work starts is between 7 a.m. and 9 a.m. and the duration is 8 hours plus lunch which means work usually ends around 4 p.m. to 6 p.m. As there is traffic before and after working hours, and not all people live close to their jobs, it might take as much as 1 hour to get to work for some people. This work considers 6 a.m. to 9 a.m. as morning period when people leave homes and arrive at work and 4 p.m. to 7 p.m. as evening period when work ends, and people go home. *Figure 3.8* shows how cellular activity depends on time. On weekdays the number of events starts to grow around 6 a.m., but mostly 7 a.m. This is the time when many people start waking up and leaving homes. 10 a.m. is the time when growth slows. From 4 p.m the frequency starts to decrease as a lot of people are finishing their work for the day and start heading back home. Evening information is not as good as morning's as people frequently do not go home after work or move somewhere else in the evening period like gym or shopping. Compared to the weekend (*Figure 3.9*), on weekdays people are less active after midnight. Usually, in Estonia, Saturday and Sunday are days off and there is no need to wake up early, so people tend to stay up late, and some of them go out to have fun. The number of cellular events still starts to rise around 7 a.m. on weekends but does not increase as fast.

For each person, the cell which processed the first event in the morning period is set as a morning home location candidate. Then a check is performed, whether previous n hours the events are also processed by the same or an overlapping cell. If that is the case, that cell is fixed as a morning home location for that day.

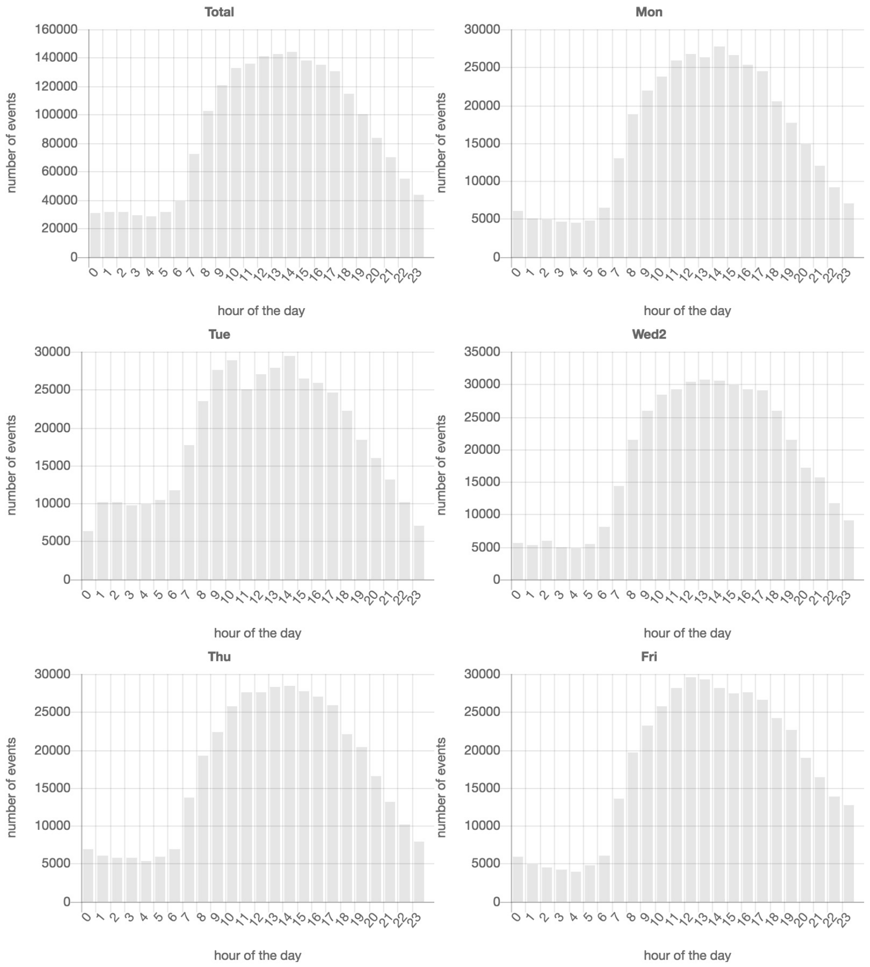


Figure 3.8: Number of received cellular event vs time on weekdays.

If not, we stop at the cell that is different, assign it as home location candidate and check n previous hours' events for that cells and see if the corresponding cells overlap. This is done until consecutive cells for n hours are overlapping or until 3 a.m. is reached, because we make an assumption that person sleeps at night and switching cells at night means that with high probability the person is moving around and probably is not at home.

Finding morning work location goes through a similar process. Last visited cell in the morning period is assigned as morning work location candidate. Next m hours are checked for movement, and if movement only occurred within overlapping cells, the work location candidate is accepted. If not, the first non-overlapping cell is assigned as morning work location candidate and is compared to next m hours of cell events until all cell visits in m hours are overlapping or until 1 p.m., the time when movement is at its peak as observed in *Figure 3.8*, is reached.

The same is done with the evening period, only the first cell occurrence in that period is a work location candidate and cells are checked back until the day time

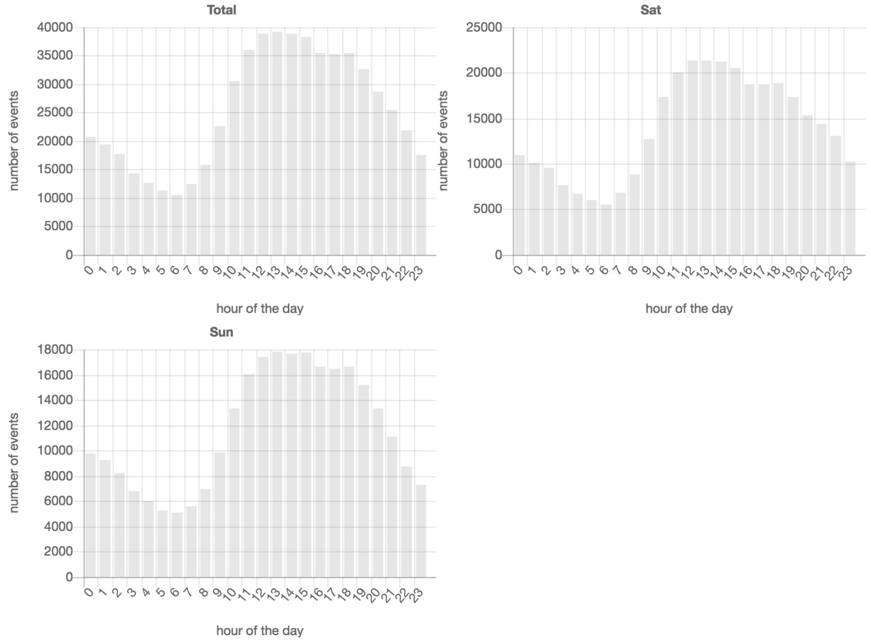


Figure 3.9: Number of received cellular event vs time on weekend.

peak at 1 p.m. Home location candidate is the last cell in the evening period and is checked until event frequency reaches its lowest point and stays consistently at about the same level, at 1 a.m. This should be the time when people have reached home and head to bed.

If after reaching the time limit no candidate satisfies the criteria, that location point is dropped. That means there are no locations, in which person stayed with certainty for enough amount of time to consider it person's home or work location. If parsing morning locations yields one home and one work location, a morning trajectory is constructed between those two points. Similarly, evening trajectory is found between evening work and home locations.

Ideally, each user would have two trajectories. However, as the data is not consistent all the time and some people do not always use their mobile phone, it may happen that some people do not have a morning or evening trajectory. To make data more informative, only users who have home and work trajectory in the morning and evening are considered for further analysis. Also, trajectories that start and end in the same cell are not of interest. After removing those people from the set, there were 34 107 left.

As finding common trajectories of different people was of interest, some algorithms are proposed for achieving the desired goal. To find matching paths, a list of all possible starting coordinates is extracted and used as a key to an array of all people who start from that position. The same is done with finish positions. The process is shown in *Algorithm 1.* This algorithm outputs two arrays, which are inputs to the next algorithm.

Now that the input data is organised, it is easier to find common paths based on intersections of people from starting and ending point. Each set of people from start coordinates is compared to sets of people from ending coordinates. If the number of the same people is equal or greater than the minimum group size, then a common path is found. To optimise the algorithm, people for whom the start and end coordinate was already checked, are extracted from ending coordinates because each person has only one path. The algorithm is shown in *Algorithm 2*.

Algorithm 1 Organise people by coordinates

```
procedure PeopleToCoordinates
Input:
    $start$ : list of coordinates from where the path starts
    $end$ : list of coordinates where the path ends
    both lists are of same length and their index represents a person
Output:
    $S$ : List of all start coordinates with a list of people starting from that point.
    $T$ : List of all end coordinates with a list of people finishing in that point.
    $S \leftarrow \{\}$
    $T \leftarrow \{\}$
    for $i = 0$ to $length(start)$ do
        $startPointAtIndex \leftarrow start[i]$
        if $startPointAtIndex$ in $S$ then
            $S[startPointAtIndex] \leftarrow S[startPointAtIndex] \cup [i]$
        else
            $S[startPointAtIndex] \leftarrow [i]$
        if $endPointAtIndex$ in $T$ then
            $T[endPointAtIndex] \leftarrow T[endPointAtIndex] \cup [i]$
        else
            $T[endPointAtIndex] \leftarrow [i]$
```

To check the precision of the morning and evening trajectories, those two paths are compared. As people move around a lot throughout the day and it is very hard to say precisely if the person was at home in the morning or at the gym, a small check is proposed to raise the certainty. For each person, that has a morning and evening home-work trajectory, we check if those two trajectories match. Meaning whether a person ends his or hers evening journey where he or she started in the morning and also starts evening path with where morning one ended. The search is extended a bit by not only comparing the equality of one point to another but comparing them with a nearby function. Because of the handover effect, and also overlapping coverage areas, we assume that if centres of the cells are not more than 0.5 kilometres apart, then these locations are the same.

Algorithm 2 Common paths

```
procedure CommonPaths
Input:
    $n$ : minimum number of matching trajectories
    $start$ : list of all path start points, with people who started in that point
    $end$ : list of all path end points, with people who finished in that point
Output:
    $S$ : list of all found paths with at least n people in common
    $S \leftarrow []$
    for $startPoint, startPeople$ in $start$ do
        if $length(startPeople) < n$ then
            continue
        for $endPoint, endPeople$ in $end$ do
            if $length(endPeople) < n$ then
                continue
            if $startPoint = endPoint$ then
                continue
            $I \leftarrow startPeople \cap endPeople$
            if $length(I) \geq n$ then
                $S \leftarrow S \cup [startPoint, endPoint]$
            $end \leftarrow end \setminus I$
    return $S$
```

Chapter 4

Results

This chapter describes obtained results that have been achieved using the methods and algorithms described above. The results are shown over one day period as the data was quite changing over the days, and more processing and complicated methods would be needed to make a better analysis for the whole week.

Figure 4.1 and *Figure 4.2* show common paths found on Monday morning and evening respectively in one part of Tartu. The minimum group size of people was 5 to form a common trajectory. There are quite many users moving between same cells. However, comparing morning and evening results shows that most of the common paths are in the same direction, so it is probably not the same group of people. It might be one group in the morning going from home to work and in the evening another group going from work to home or some other place. It can also mean that these cells pick up events more often from the overlapping areas than others. Some trajectories are quite short which could mean just handovers. Still, we see in the middle of the region in the morning a very long trajectory and almost the same path in the evening only the opposite way.

Figure 4.3 shows movements, for which the trajectory is the same in the evening and morning, which means those people moved in the morning from the first cell to second and in the evening period from the second back to first. There were 536 people with same trajectories on Monday. For most users, the trajectories are quite long meaning they actually moved, and the chances of handovers and overlapping polygons are quite small compared to shorter journeys. *Figure 4.4* and *Figure 4.5* show home and work locations respectively based on the same Monday data. The brighter the cell, the bigger density of those locations. Although home and work regions are very similar, *Figure 4.3* shows that there are still a lot of movements. This could mean that for some larger group of people one cell is a home region and for some other group it is where they work. Frequent cells that border the ouline of Tartu, could also be just the first cells for people who come or got to work outside of Tartu. Another small observation is that home locations are a bit more frequent in the suburban area, and work location densities are a

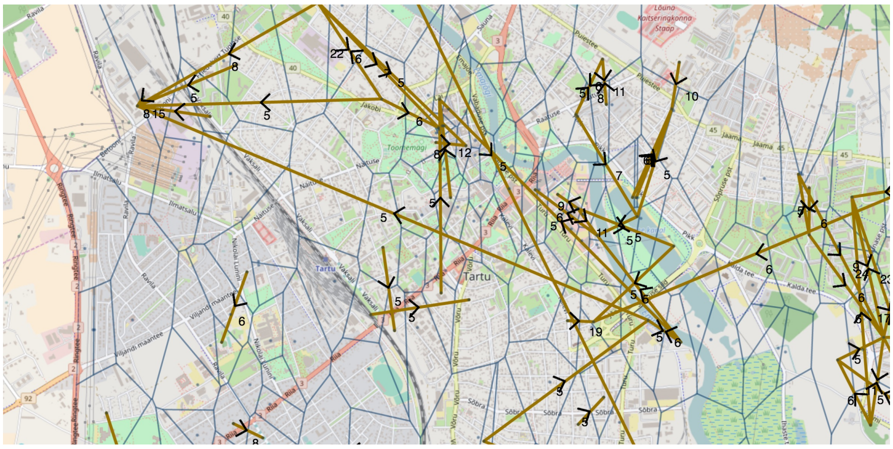


Figure 4.1: Trajectories that are in common for at least 5 people on Monday morning.

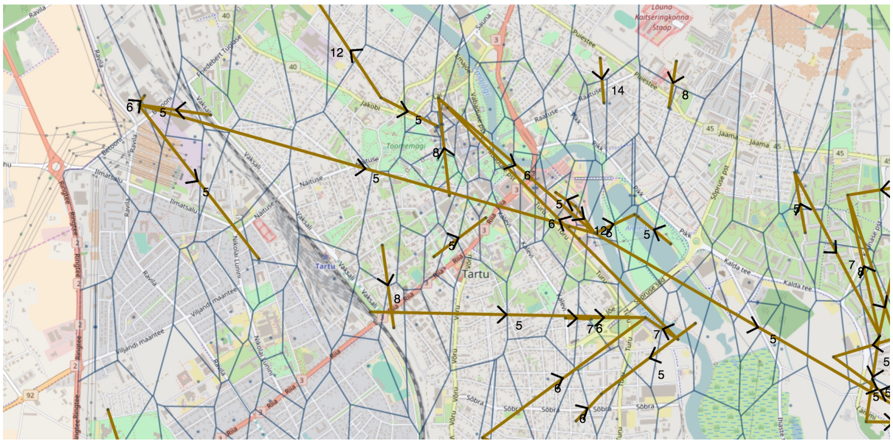


Figure 4.2: Trajectories that are in common for at least 5 people on Monday evening.

bit higher in the city centre. We can observe that regarding urban planning, the city of Tartu is encouraging to create offices building close to residential areas. In Tartu, there is a river flowing through the city centre. In the city there are only 4 bridges for automobiles connecting two sides. As seen from *Figure 4.3*, a lot of movement involves getting to the other side, so the bridges are very important regarding mobility. Closing down one bridge could result in big mobility issues, by increasing the traffic density that can lead to a large traffic congestion. The city has to carefully plan transportation to create a better urban environment.

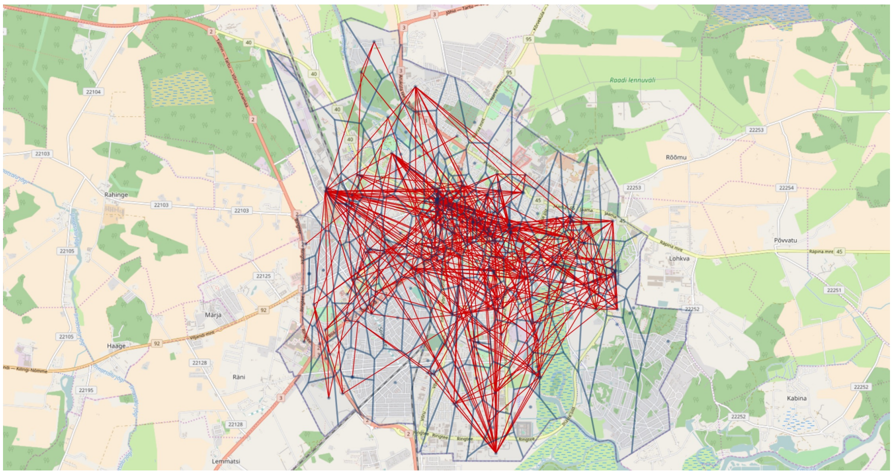


Figure 4.3: Each line represents a person, whose morning and evening trajectories were the same, but in the opposite direction. Derived from Monday data.

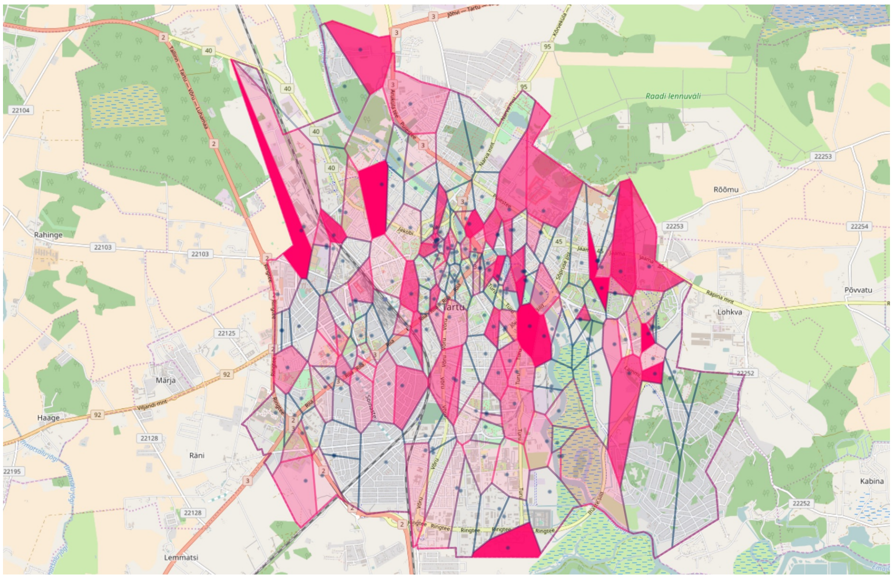


Figure 4.4: Pink cell shows that this cell is a possible home cell for some user. The brighter the cell, the more people have it as a home cell.

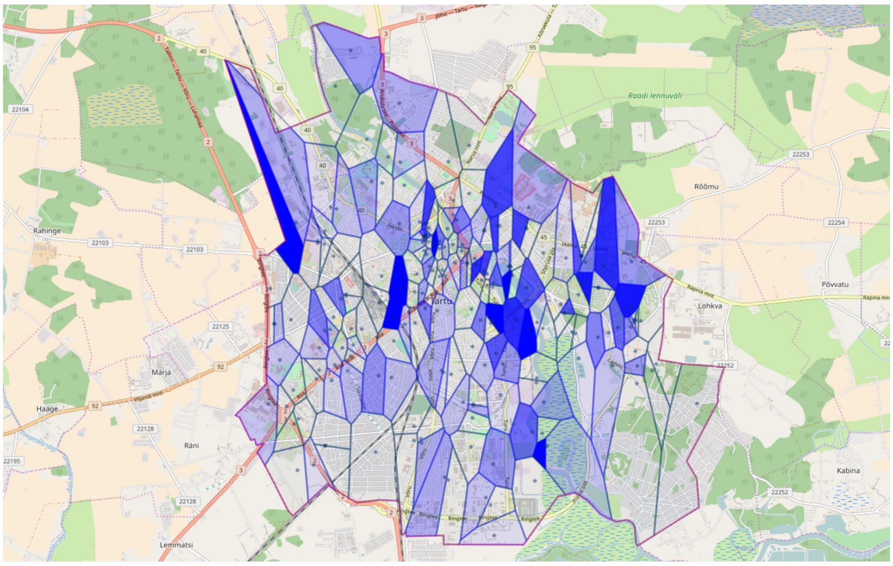


Figure 4.5: Blue cell shows that this cell is a possible work cell for some user. The brighter the cell, the more people have it as a work cell.

Chapter 5

Summary

5.1 Conclusion

Gathering data from CDR is quite difficult and needs a lot of analysing and research. This thesis gave only a small glimpse into the big world of big-data. Other researches show different approaches on how to extract meaningful data and observe human mobility patterns. It is possible to find home and work locations with quite big certainty. People usually do not follow strict patterns so predicting their movements is a challenge. However, taking into consideration their social ties and factors, some behaviours are predictable.

Processing such big data takes many attempts as some assumptions do not hold in real life scenarios and do not give informative results. When using mobile phone data to find locations, it is necessary to take into account that person's location could be anywhere in the coverage area and due to handovers he or she might seem to be moving while in reality staying in one place. However, with quite big certainty it is known that most people move in the morning period and also in the evening, as they need to get to work, school or another place, and then go back home. This movement is a pattern that many people follow on weekdays. Based on that knowledge we could extract possible home and work locations based on the cell towers that offer cellular service at that site.

Even though the method proposed in this thesis found many trajectories that could represent people's home to work paths, it is not possible to check with real results as the data is anonymous. However, there is a big similarity in found home and work location. It is possible that popular home locations in the city of Tartu are also popular work locations. These home, or work, locations did not form any distinguishable groups, so there were no urban and suburban communities found. However, we could still observe that many people do have patterns that are the same for morning and evening.

To sum it up, visualisations in this thesis showed that there is potential in working with CDRs and in the future we might see more and more services offering targeted marketing, rescue service using information for effective evacuation plans and much more.

5.2 Future Perspectives

As the human mobility patterns attract more and more companies from different fields, the number of different services that offer such pattern analysis are increasing. The quality of those services is improving as more CDR data is gathered every day, which enhances the accuracy through learning that data. There are still improvements to be made to position a person accurately.

This thesis showed how the places, where people stay during night and day time, can be gathered through observing their movement. Using probabilistic and statistics to eliminate handovers could result in outputting more people with observable patterns. This might give better results in finding urban and suburban communities.

One use of understanding human mobility is targeted advertisement. For that, analysing person's position based on available points of interest could give information about him or her. As the next step from this thesis, adding a semantic layer to the mobility patterns would build a better understanding of human activity behaviour.

Bibliography

[ASJ+10] Rein Ahas, Siiri Silm, Olle Järv, Erki Saluveer, and Margus Tiru. Using mobile positioning data to model locations meaningful to users of mobile phones. *Journal of Urban Technology*, 17:1,3–27, 2010.

[CPDL+10] Francesco Calabrese, Francisco C Pereira, Giusy Di Lorenzo, Carlo Ratti, and Liu Liang. The geography of taste: Analyzing cell-phone mobility and social events. In *Pervasive Computing*, volume 6030, pages 22–37. Springer Berlin Heidelberg, 2010.

[FGNG12] Anna Förster, Kamini Garg, Hoang Anh Nguyen, and Silvia Giordano. On context awareness and social distance in human mobility traces. In *Proceedings of the 3rd International Workshop on Mobile Opportunistic Networks (MobiOpp)*, volume 1, pages 5–12. ACM, 2012.

[For87] Steven Fortune. A sweepline algorithm for voronoi diagrams. *Algorithmica*, 2(1):153, 1987.

[fra06] Pnpoly point inclusion in polygon test. `http://www.ecse.rpi.edu/Homepages/wrf/Research/Short_Notes/pnpoly.html`, 2006.

[Hos09] William L Hosch. Zipf's law. *Encyclopedia Britannica*, 2009.

[IBC+11] Sibren Isaacman, Richard Becker, Ramón Cáceres, Stephen Kobourov, Margaret Martonosi, James Rowland, and Alexander Varshavsky. Identifying important places in people's lives from cellular network data. In *Pervasive Computing*, pages 133–151. Springer Berlin Heidelberg, 2011.

[inP] substack point in polygon. `https://github.com/substack/point-in-polygon`.

[itl16] Smart urban planning, Targa linnaplaneerimise uus tase. Eesti Infotehnoloogia ja Telekommunikatsiooni Liit, 01.08.2016.

[JKA11] RK Jain, Sumit Katiyar, and NK Agrawal. Hierarchical cellular structures in high-capacity cellular communication systems. *International Journal of Advanced Computer Science and Applications*, 2(9), 2011.

[KGSR14] Kevin S Kung, Kael Greco, Stanislav Sobolevsky, and Carlo Ratti. Exploring universal patterns in human home-work commuting from mobile phone data. *PLoS ONE*, 6(9), 2014.

[ope] OpenLayers homepage. `https://openlayers.org/`.

[staa] Statistics Estonia estonian population density. `http://www.stat.ee/ppe-327960`.

[stab] Statistics Estonia users of mobile phone aged 16-74 by group of individuals, 2012. `http://pub.stat.ee/px-web.2001/Dialog/varval.asp?ma=IC61&ti=USERS+OF+MOBILE+PHONE+AGED+16-74+BY+GROUP+OF+INDIVIDUALS%2C+2012&path=../I_Databas/Economy/20Information_technology/04Information_technology_in_household/&lang=1`.

[tar] Tartu outline polygon. `http://www.openstreetmap.org/relation/2153396#map=13/58.3750/26.7325`.

[vor] d3 voronoi. `https://github.com/d3/d3-voronoi`.

[Yam16] Ikuho Yamada. Thiessen polygons. *The International Encyclopedia of Geography*, 2016.

All the web references were valid on 11.05.2017.

License

Non-exclusive license to reproduce thesis and make thesis public

I, Anna Bilmaijer,

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,

   "Urban Mobility Sensing Using CDRs", supervised by Amnir Hadachi,

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

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

Tartu, 11.05.2017