University of Tartu
Faculty of Science and Technology
Institute of Ecology and Earth Sciences
Department of Geography

Master thesis in Human Geography 30 ECTS

Unmasking oscillation from mobile positioning data

Sander Pukk

Supervisors: PhD Anto Aasa
Erki Saluveer

Allowed to defence:
Supervisors:
Head of department:

Tartu 2019

Abstract

Unmasking oscillation from mobile positioning data

Passively gathered mobile positioning data has widely been used to study human mobility. All the analyses done with those datasets are dependent on the quality of the data. One of the issues affecting passively gathered mobile positioning data is a phenomenon called oscillation. These are recorded logs, which indicate the device's movement between cellular towers in abnormal manner. Generally, they show the movement to did not occur in real life. The purpose of this thesis is to create a method that can unmask oscillation from passively gathered mobile positioning data and to better approximate the location of the user. The method was tested on a test sample group from the main dataset and the results indicate that it does detect oscillation. The share of oscillation unmasked in an Estonian dataset of over 200 million network events is lower than the examples from the literature. There are some explanations for this, one being that cellular towers in Estonia are spatially more sparsely distributed than in the study areas of the examples from the literature. Although the method detects oscillation, a larger test group would greatly benefit any future works for improving the method.

Key words: mobile positioning data, call detail records, oscillation

CERCS code: P175 Informatics, systems theory; S230 Social geography

Annotatsioon

Mastiviskamise tuvastamine mobiilpositsioneerimise andmetest

Passiivselt kogutud mobiilpositsioneerimise andmeid on laialdaselt kasutatud inimeste mobiilsuse uurimiseks. Nende andmestike kasutamisel on analüüsid otseses sõltuvuses andmete enda kvaliteedist. Üks probleemidest, mis mõjutab mobiilpositsioneerimise andmeid, on mastiviskamine. Neid olukordi, kus kasutaja lühikese aja jooksul näiliselt liigub mitme masti vahel ja tõenäoliselt tegelikult ise asukohta ei vaheta päris elus, nimetatakse mastiviskamiseks. Käevoleva magistritöö peamiseks eesmärgiks on väljatöötada meetod, mille abil tuvastada mastiviskamised mobiilpositsioneerimise andmetest. Kõigepealt testiti meetodit katsealuste peal, kelle põhjal leiti, et antud meetod tuvastab mastiviskamist. Seejärel rakendati meetodit 200 miljonist kõnetoimingust koosnevale andmestikule, mille käigus tuvastati vähem mastiviskamist, kui näited kirjanduses on leidnud. Tulemusele on mõningaid seletusi, üheks peamiseks on mastide hõredam paiknemine Eestis võrreldes kirjanduse näidete piirkondadega. Kuigi meetod töötab ning tuvastab mastiviskamist, tuleks tuleviku töödeks laiendada esialgsete katsealuste valimit, mis aitaks põhjalikumalt arendada meetodit edasi.

Märksõnad: mobiilpositsioneerimise andmed, kõnetoimingud, mastiviskamine

CERCS kood: P175 Informaatika, süsteemiteooria; S230 Sotsiaalne geograafia

Table of contents

Introduction

In the last decade, the usage of mobile phones on a global scale has more than doubled. The estimated number of unique mobile phone subscribers in 2007 was around two billion. By 2017, it hit the five billion mark (GSMA Intelligence, 2017). At the same time, mobile networks and mobile broadbands have had their own breakthroughs in terms of wireless technologies, with the emergence of 4G in 2010 and now the implementation of 5G networks. With the new network developments and advances in smartphones, an increasingly higher number of people use mobile phones constantly and in one way or another, are constantly connected to a mobile network.

The current trend for Internet usage is the rise of the percentage of mobile devices of all the traffic in the web. Nowadays, over half of the visits to a website in United States are made from mobile devices (Enge, 2018). It shows that people use their phones for more than just calling and are perhaps more engaged with their device throughout the day. All those Internet usages, calls, SMS or any other mobile phone events generate logs for the operator to keep track of, so they can bill the customer and have a legal archive. Those logs are often referred to as call detail records (CDR) and logs which also contain mobile data communication as data detail records (DDR) (Horak, 2007).

Passively generated mobile phone location data (as CDR and DDR) has now been widely used to study human mobility and characteristics in different fields. From tourism (Ahas et al., 2008; Ahas et al., 2014; Girardin et al., 2009; Raun et al., 2016) to everyday commuters (Ahas et al., 2010a; Kung et al., 2014), to profiling people based on mobility (Bayir et al., 2010; Furletti et al, 2014) and of course studying people's general mobility patterns (Gonzalez, 2008; Lee & Hou, 2006) with many other examples from the literature. But as with any dataset, mobile positioning data has some flaws and they need to be addressed before it is possible to study all those things mentioned beforehand.

One of the issues affecting CDR (and DDR) is a phenomenon called oscillation (Bayir et al., 2010; Chen et al., 2016; Wu et al., 2014; Qi et al., 2016). In literature, it might also be called cell-tossing (Ahas et al., 2010b) or ping-pong effect (Gu et al., 2010; Iovan et al., 2013). In this thesis, oscillation will be used.

The mobile device is usually connected to the nearest, mostly with the strongest signal, cellular tower. But due to different reasons (briefly described in section 1.2) it might switch to another one (and back) without the device actually changing location in real-life, essentially creating noise in the dataset. This type of switching is called oscillation. Different results have been reported by researchers on the extent to which oscillation affects a CDR dataset. Iovan et al. (2013) eliminated 16% of logs, at the same time other researchers estimate their findings from 6% (Wu et al., 2014) to 13% - 15% (Qi et al., 2016). In general, the amount might be significant to raise the need to cleanse the dataset to improve the quality of the data.

If a device changes location and goes from one cell's coverage area to another's, the device's network events are transferred from the first to the second. This is referred to as a handover (Sauter, 2010). With the emergence of 4G and 5G technologies, the coverage area of a single cell is becoming smaller due to the technical aspects of the new generations and more cells will be needed to cover the same area as compared to 2G and to some extent 3G (Correspondence with operators). Due to closer proximity and the increase in the number of cells in the former space, more handovers will occur, which will most likely produce more oscillation.

This thesis will tackle the issue of unmasking the oscillation phenomenon from CDR/DDR logs. These kinds of the databases are usually tens or hundreds of gigabytes in size. If more than 10% of that data is excess noise (as Iovan et al. (2013) and Qi et al. (2016) detected) and it is possible to unmask it before completing any analytical calculations, it would save time and resources. Another reason to unmask oscillation is that in order to analyse and produce high quality statistics, mobility or any other results, the location of the user must be know. When detecting oscillation, the most likely location of the user for the oscillation logs is determined. This raises the location accuracy and therefore also raises the quality of the dataset.

The goal is to create a method that detects oscillation based on the examples of Wu et al. (2014) and Qi et al. (2016) and apply them on a test CDR/DDR dataset, whose users and their actual location at a given time during the study period of 2 months are known. Those test group users' real-life locations are compared to unmasked oscillation logs and it is then determined whether the captured logs are due to oscillation or not. **To answer how much oscillation occurs in a CDR/DDR dataset, and how oscillation is spatially and temporally distributed**, the method is also applied to a larger CDR/DDR dataset of over 500 000 users over the period of one month to assess the amount of oscillation in the case of Estonia.

1. Theoretical concepts of oscillation and overview of the subject

The number of people, who use mobile data communication has risen tremendously over the last decade. The Technical Regulatory Authority (TRA) of Estonia (2017) reports that the total amount of end-user mobile data consumption has increased more than ten-fold from 5873 to 80 455 terabytes between 2011 and 2016 (Figure 1).

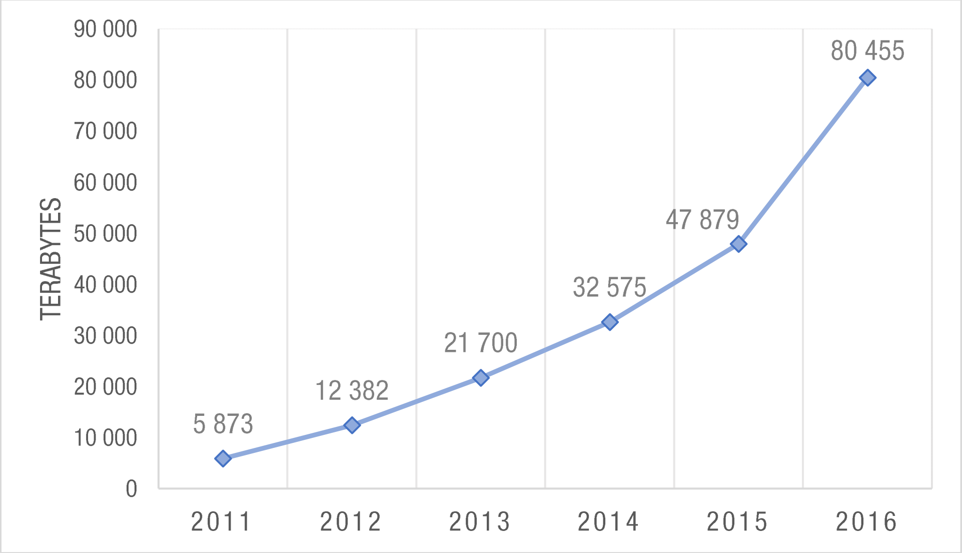


Figure 1. Mobile data consumption from 2011 to 2016 in Estonia (Data: TRA).

In the context of mobility and research in the field, the rise of mobile phone users is especially evident. This is particularly clear, when looking at the statistics from Eurostat for individuals who use mobile phone data on the move in the European Union (28 countries). Although it is found only for age group 16-74, it still shows the overall trend (Figure 2).

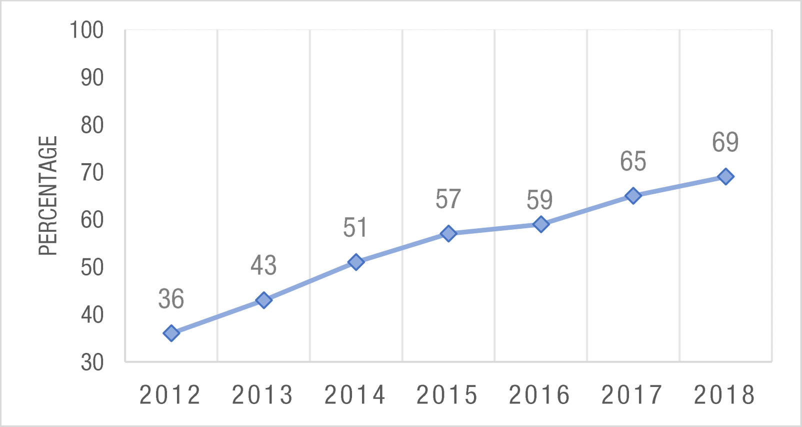


Figure 2. Individuals using mobile devices to access the Internet on the move in the European Union (28 countries) (Data: Eurostat).

At the same time, the number of voice calls has not experienced a sudden decline (Ofcom; GSMA Intelligence, 2019). It leads to a situation where the number of call activity logs is almost the same as before, but the records of mobile data usage have risen significantly over the past few years. For billing and legal purposes, the mobile network operators must spend resources and keep databases of those logs (Ahas et al., 2010b; Tiru, 2014; Wang & Chen, 2018; Wu et al., 2014). But that also brings an opportunity: that wealth of information can be used to study, for example, human mobility, spatial characteristics and aspects of people's lives, urban rhythms, transportation and so on. A brief overview of a mobile network and how the location of the device is determined in it, can illustrate how passive mobile positioning data is gathered.

1.1 Mobile network and passively gathered mobile positioning data

A cell is the smallest structural part of a cellular network (Figure 3). The term is quite loose and can mean both the coverage area of the antenna and the antenna itself. In this work, the cell is defined as the antenna. A cellular tower (base station, site) can have multiple antennas and therefor multiple cells. Every cell can be described by several attributes such as azimuth, sector angle, shape and size of the coverage area, type of antenna and location.

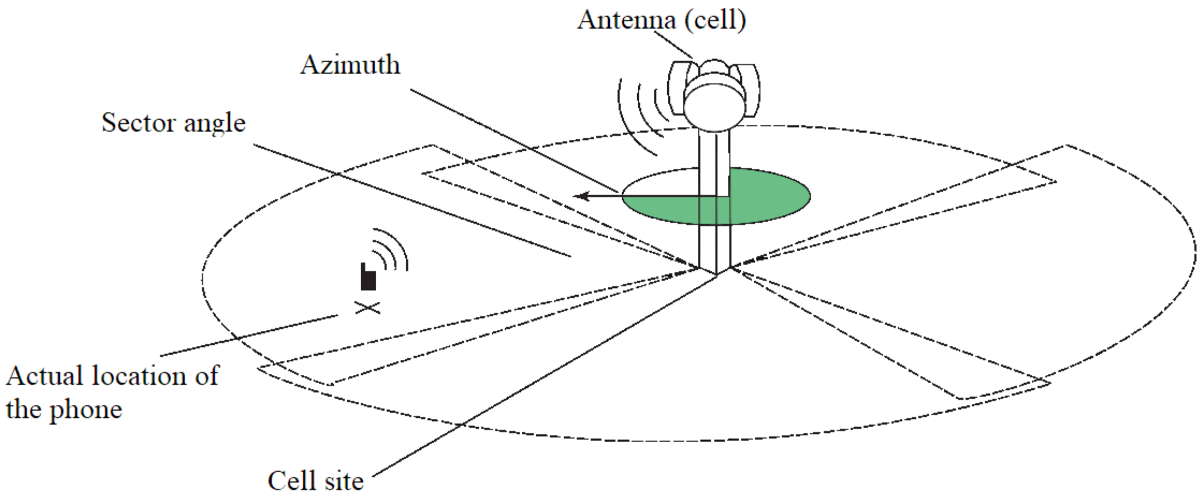


Figure 3. Cell, site and location area, and their relation to one another (Ahas et al., 2014).

A mobile operator's network consists of thousands of these types of cells and cover vast areas. They are usually spatially unevenly distributed as more clients are in urban areas and in the rural areas the coverage is usually sparser and involves fewer cells. As that is mostly the case, the coverage area for a single cell varies greatly from a couple of hundred meters to up to 35 kilometres in rural areas (Sauter, 2010).

Mobile phone location data can be divided into two broader categories: actively and passively gathered (Ahas et al., 2007). Without dwelling too much on the specifics, actively generated mobile phone data is when mobile network pings the location of the device trough triangulation or other means, where it actively seeks the device. Passively gathered mobile phone location data is generated through logging mobile device activities in the network. Each time a device does a network event (e.g. calls, sends SMS, uses Internet), it is logged and stored by mobile network operators for billing and legal purposes (Ahas et al., 2010b; Tiru, 2014; Wang & Chen, 2018; Wu et al., 2014).

Passive mobile positioning location data is mostly gathered by connecting **call detail records (CDR)** (Ahas et al., 2007; Horak, 2007; Tiru, 2014)**,** which are generated when a device does a phone call or sends/receives an SMS, with the antenna's location the call activity was made from. There is also **data detail records (DDR)**, which are logged when a device uses a certain amount of Internet data. In this thesis, the CDR is used to describe both unless specified otherwise. Depending on the mobile network operator, both sending and receiving can be stored. In this instance, only the outgoing information is used.

Typically, a CDR dataset includes a timestamp, the device's ID and the cell's ID, in which the event occurred. The mobile network operator keeps the logs for billing purposes (Ahas et al., 2010b; Wang & Chen, 2018; Wu et al., 2014). We can derive the approximate location of the device when connecting the cell ID with its point location. Although the device can be almost anywhere inside the coverage area, the location is usually aggregated to the cell's point location or the service area of the cell. The best service areas are usually procured from the mobile network operators, but otherwise there are theoretical ways of calculating areas from the cells point locations. One of the main methods for calculating the theoretical area of a cell is using Voronoi tessellation in Euclidean space, where every point gets a service area and they do not overlap (Ahas et al., 2010b; Gonzalez et al., 2008; Järv et al, 2014). Passively gathered mobile phone location data can also be generated by other network activities, such as sightings data, which is produced during handovers between neighbouring cells (section 1.2), but in thesis the CDR/DDR dataset is used.

1.2 Mobile device's location and handover

To receive or start a mobile network event, the device's location must be known to the network. That way the network can conclude in which cell the event is going to take place and connect to other devices inside the network through cells (Chen et al., 2016; Wu et al., 2014). The process of notifying the network of the device's location is called signalling (Sauter, 2010). Depending on the population and the network, there are usually hundreds of thousands to millions of devices in the network. If a single device would send regular updates to the network every time it changes cells, it would significantly increase the signalling load. Therefore, multiple adjacent cells are grouped together into location areas (Figure 4) and only the location area information is stored in the network as the device's location (Sauter, 2010). To connect one device with another, the network searches for the device inside the stored location area only, not in the entire network. This process is called paging (Sauter, 2010). If the device changes location, but remains in the same location area, no signalling data is sent to the network. Only if the device changes its location area, then it is updated. Depending on the mobile network operator, the location areas generally consist of 20 to 30 cells (Sauter, 2010).

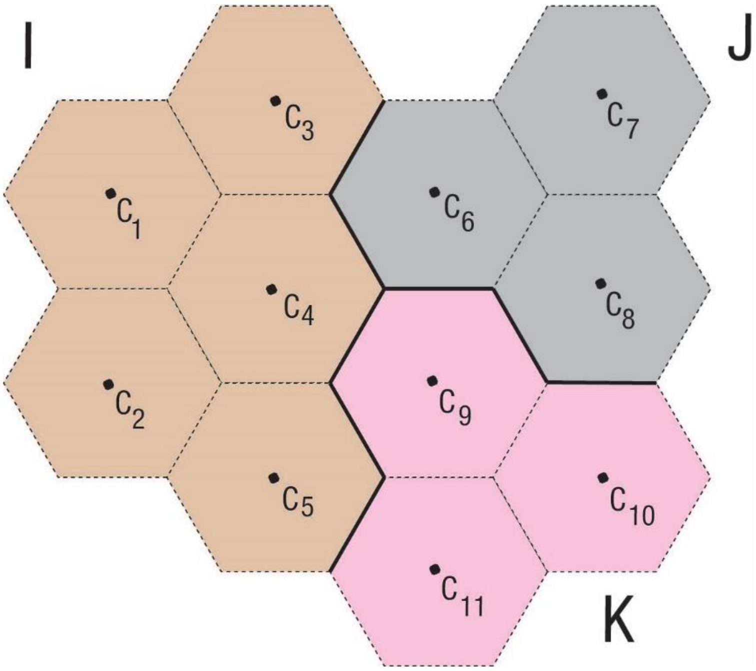


Figure 4. Cells (1-11) are each grouped into one of location areas I, J or K. If a device moves from one location area to the next, the area code for that device is updated

In theory, the mobile device should be connected to the closest antenna (cell) with the strongest signal strength. However, due to mobile networks' technical properties and cells' real spatial and topographical properties, it might not always be the case. A handover is the occurrence when the device's currently ongoing mobile network event is transferred from one cell to another (Corazza et al., 1994; Iovan et al., 2013; Sauter, 2010). In a mobile network, a handover usually happens when the device changes its location, e.g. is moving spatially from one cell to another (Iovan et al., 2013). The threshold for a handover is determined by the mobile network operator (Sauter, 2010).

In the case of SMS or Internet usage, the data is sent in packets and is not continuous in the same sense as a call is. Mobile devices generate non-continuous packet flows, which at some point are recorded in the mobile network (Sauter, 2010). With the rise of popularity and the wide-spread usage of on-the-go Internet access, records might be generated in short periods of time (minutes, even seconds). That can lead to the same situation as with call records, where the device itself has probably not moved spatially, but a handover event has occurred. The records with DDR are denser in terms of time and space compared to more traditional CDR. Data detail records (DDR) enrich the temporal and spatial resolution of the dataset. More logs mean that there are fewer gaps between network events as the mobile phone usage has shifted from more sparsely made phone calls to actively used mobile data communication (TRA, 2016; GSMA Intelligence, 2017; GSMA Intelligence, 2019). As the applications in the mobile phone update themselves and are almost constantly syncing with new information when mobile data communication is enabled, more records are being produced. The same applies to spatial coverage, as records are being logged in shorter intervals (more logs), giving better spatial resolution and movements of the device if its mobile data is being used.

1.3 Oscillation

A handover might occur with the device not changing its real spatial location. It might occur due to the load balancing of the network, not to overwhelm a single cell and distribute the load more evenly across the cells (Wu et al., 2014, Qi et al., 2016). Other aspects which can cause handovers are due to different weather events (like rain) or topographical elements like hills or building, which might block or intervene with the signalling strength from the currently connected cell. Or if the signal strength of two cells is almost equal (Iovan et al., 2013; Wu et al., 2014). All of that might invoke a handover situation. For a device to be able to do a handover, it must also constantly inform the network with the signal strength of nearby cells in its connecting radius, besides the one it is connected to (Miao et al., 2016). If a nearby cell’s signal becomes stronger than the current cell’s, a handover is invoked. The decision to invoke a handover to which tower to switch over, comes from the network as not to overwhelm the battery of the device. All of that can make the device hop between two cells without changing its real spatial location.

In the literature, there are different terms used for the phenomenon. Iovan et al. (2013) use “ping-pong records” as in order to illustrate the nature of the records based on table tennis, where the device is “bouncing” between two (or more) cells. It is described as “cell-tossing” or “switching” by Ahas et al. (2010) as the device being “thrown” between cells. The recent and more commonly used term to describe this sort of occurrence is oscillation (Chen et al., 2016; Laasonen et al., 2004; Wang & Chen, 2018; Wu et al., 2014).

For the purpose of this thesis, oscillation is defined by **Wu et al. (2014) as follows: “An oscillation occurs when a communication transaction oscillates between multiple cellular towers even though the mobile device is not moving”.**

When oscillation occurs, there might be records, which show that a person has travelled hundreds or thousands of meters in just mere seconds and then back. To gather and process people’s real movement, those kinds of entries are considered noise and slow down the overall data processing. Mapping unrealistic entries based on speed, is generally the first step undertaken. That in itself would not be insufficient, because standing on the edges of two cells’ coverage area (Figure 5), one might change and keep on moving to another cell with high speeds. In that example if a handover occurs in 2 seconds and the distance between cellular towers A and B is 5 kilometres, the speed would be 9000 km/h and might seem unrealistic. The key for detecting oscillation is locating certain patterns and using multiple parameters at the same time, not only looking each separately.

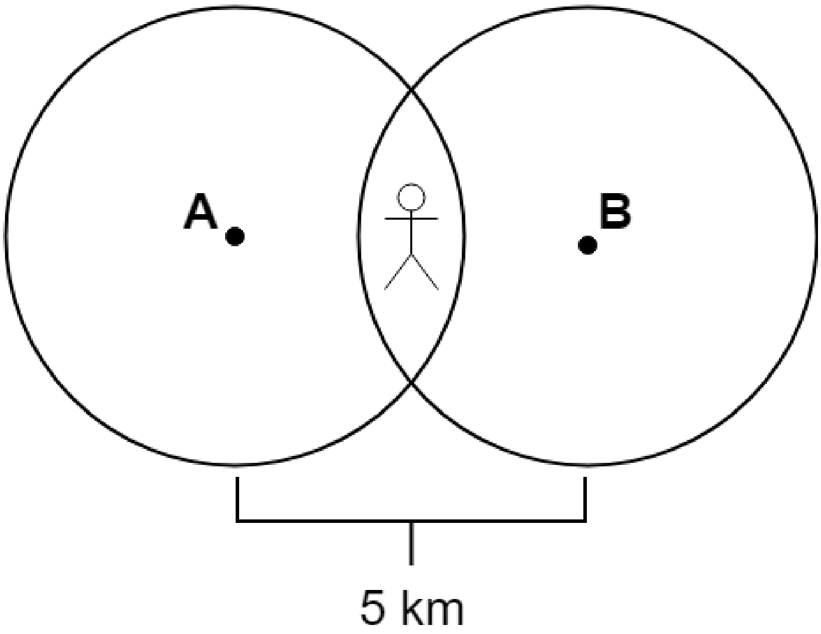


Figure 5. Standing on the edges of two cells' coverage areas

Iovan et al. (2013) use both the speed and velocity in their case and azimuth for the change of direction. They used a certain time window between trips to select suspected entries oscillating between some cells and then calculated the velocity between them. One of their assumption is that if a device is oscillating between two towers, the oscillation is moving in the directly "opposite" direction and the heading change is 180 degrees from the previous entry in the same window. They calculated the directional headings for each of these suspected pairs as true North was considered the basis. They removed entries which were over 200 km/h and with a directional change of 180 degrees. Approximately 16% of their 122 208 870 points were removed by detecting oscillation. Perhaps the weakness of this method is that it only considers oscillation between a pair, not multiple neighbouring cells and the difficulty of setting a proper speed threshold not to remove real movements from the dataset.

The same issue has been investigated for wireless access points (APs) by Lee and Hou (2006) where they used pattern-based logic to find oscillation. The coverage area of multiple APs in the campus area was simplified to a hexagon tessellation and is in a sense comparable to any cellular network, where the coverage is almost continuous across the entire area, same as for the mobile cellular network. They looked for user entries where the sequence followed a series of switches between APs. Two patterns were considered oscillating transitions:

1. Between two APs *i* and *j* with the following pattern: $i \rightarrow j \rightarrow i \rightarrow j$
2. Between three APs *i, j* and *k* with the following pattern: $i \rightarrow j \rightarrow k \rightarrow i$

An interesting find in that study was that over 30% of transitions between access points were considered oscillation and their general effect on mapping user mobility should not be ignored. The oscillating transitions were aggregated to a set, based on the above logic and the AP to which the user was connected to most of the time during the oscillation, was selected as the main AP. Though it is obvious that for their study they did not consider time and distance between the transitions, as that is a factor for mobile generated records. The strength of their method is that they considered not only oscillation between two, but multiple points (or cells).

Bayir et al. (2010) used a similar method, but with mobile generated data. The Reality Mining dataset used there was quite similar to passively generated CDR data. For that dataset, people were given mobile phones and each cellular tower change was registered. The location information is the same as for CDR data – not directly positioned mobile device, but through the recorded connection's cellular tower coordinates. The first step was to find proper candidates for oscillation. In their method, they only included records, which were made during movement/transition. End-locations were considered if the stay period was longer than 10 minutes and entries less than that were travelling. End-locations were always the first or last cellular towers of a trip.

Thus, a trip had a stay-location when the travel time between cellular towers was longer than the threshold of 10 minutes and then a new trip from there was calculated. In their method, Bayir et al. (2010) conceived that in urban areas, oscillation might not only be between two cellular towers, but multiple towers might be in-between the pair due to load balancing and other technical aspects of a mobile network. An oscillating pair in a trip must have at least 3 switches between the same cellular towers. So a sequence of a trip between cellular towers [*x,y,x,w,v,w,y*] gives only the pair {*x,y*}. With *x* and *y,* the mobile device switches three times – from indices 0 to 1, 1 to 2 and then 2 to 6. Cell towers *w* and *v* do not form an oscillating pair as there is not enough switching from one to the other and are considered just cellular towers between the oscillating pair. The strength of this method is the consideration of more than two cellular towers, in which the mobile device might be switching between in dense areas, while remaining static. As with the Lee and Hou (2006) method, this kind of pattern-based method might exclude actual real movements of people who really move between two cellular towers.

A more thorough approach is combining the speed and pattern-based method for a hybrid method. Wu et al. (2014) use a more practical solution with four heuristics to find oscillating logs. While finding oscillating logs between stable periods (10 minutes in a stay-location) they do not use speed per se, but time and distance between the logs. The logic behind it is, that the time difference of the two consecutive logs might be small and when calculating speed (*distance/time*), it produces abnormal speeds. For example, if one might stand on the border of two cells and a handover occurs, the user moves really fast from one cell location to another. The two cell points might only be couple of hundred meters apart, let's say 500 meters (Figure 6). The switch happens in 3 seconds. The speed for that would be 600 km/h and for all intents and purposes, considered abnormal speed. But that might be actual normal handover between two cellular towers when a device is moving. So, they use thresholds for both distance and time separately.

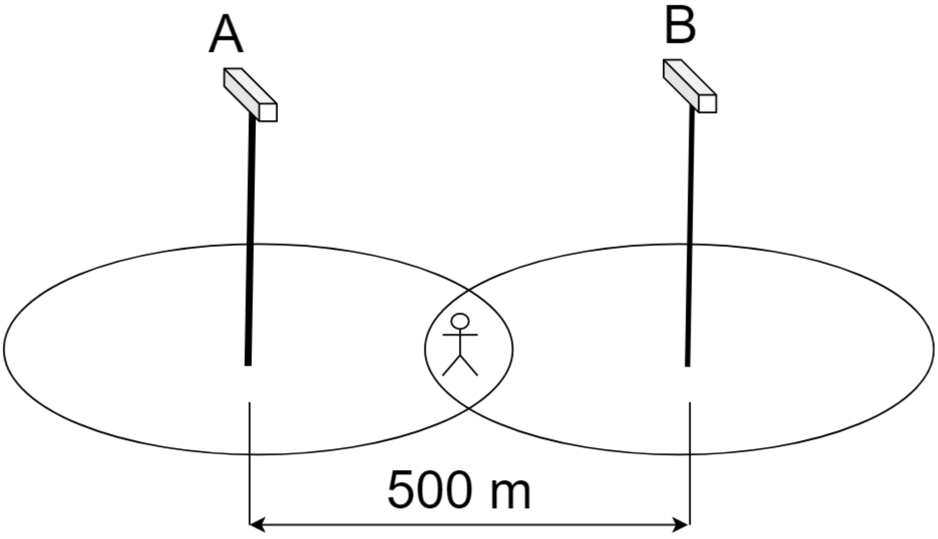


Figure 6. If a user has two network events in 3 seconds from two cells *A* and *B* with the distance between them 500 meters, the travelling speed of the user would be 600 km/h.

The four heuristics of Wu et al. (2014) might be described as such:

1. If there are two stable periods (more than 10 minutes) in the same cellular tower and there is a log between them under the time threshold (2 minutes), the log is considered oscillation. For example, a sequence of logs {*x*,*x*,*x*} *y* {*x*,*x*,*x*}, where the grouping shows two stable periods and the middle cellular tower *y* is in-between them (switching under the set time threshold), then it is oscillation.
2. If a log is shortly after a stable period and the distance is greater than the set threshold, it is considered oscillation. For example, a sequence of logs {*x*,*x*,*x*} *y*, where the grouping shows a stable period, if the *y* log happens under the set time threshold (2 minutes) and the distance between cellular towers *x* and *y* is unrealistic for a user to move, it is considered oscillation.
3. If in a sequence of three logs, the first and last log is the same cellular tower and the middle one occurs unreasonably fast and is far from the other two cellular towers, then the middle one is considered oscillation. As explained, speed itself might not be good enough if the time difference is small, time and distance should be observed separately. For example, there are records A, B and C, where A and C are the same cell or C is a close by one and the sequence is *A*-*B*-*C*. To consider the log B oscillation, the sequence must satisfy the following conditions (T meaning set threshold):

$$(\text{Speed (AB)} * (\text{Speed (BC)}) > \mathbf{T}_{\text{speed}}) \text{ AND}$$
$$(\text{Distance (AB)} > \mathbf{T}_{\text{distance}}) \text{ AND}$$
$$(\text{Distance (BC)} > \mathbf{T}_{\text{distance}}) \text{ AND}$$
$$(\text{Distance (AC)} < \mathbf{T}_{\text{distance}}/2)$$

4. The fourth heuristic contains additional steps compared to others. Firstly, they find candidates or as Wu et al. (2014) call them – suspicious sequences. They select logs, which are in a short period of time and have at least over three logs from more than two cellular towers. They expand the sequence, by looking a minute back and forward form the sequence and stops when a new cellular tower is encountered, which was not in the first selected suspicious sequence. If the cellular tower switching is in a circular manner, e.g. starting from tower *x* and ending in the same tower *x* and having other records in-between, the sequence is considered oscillation. For every such sequence, one cellular tower is chosen by a score-based algorithm to represent that user location and other logs are removed.

Wu et al. (2014) also measured the effectiveness of their method by comparing logs with the device’s actual ground truth. They first thought about using GPS locations the user might collect in the same time period (2 weeks), but decided that the GPS might not be reliable enough for indoors and opted for the user to keep a travel diary for a certain period of time. The user later corrected the oscillated records manually with their actual location for that given time. With their method, they removed around 6% of the records. A comparison with Bayir et al.’s (2010) method was also made and demonstrated that Wu et al.’s (2014) DECRE (Detect, Expand, Check, and Remove) algorithm removes more oscillating records and is closer to the true location of the user at that given time.

A similar methodical approach was implemented by Qi et al. (2016) who used their SOL algorithm, which uses three different time frames as periods. A stable period is when a device is connected to the same cell for a period of time, whereas an oscillating period is when during a short time interval, the device has multiple logs from multiple different cellular towers. The leap period is defined as when the device jumps from one cellular to another far away one and back to the first one or to a tower close to the first record during a short period of time. For every period, there are heuristics to find oscillating logs depending on the period. As with Wu et al. (2014), a pattern-based approach and time/distance is combined to find oscillating logs. Applying both approaches together most likely yields more comprehensive and trustworthy results than using them separately.

2. Data and methods

2.1 Data

CDR dataset usually consists of at least a timestamp, a cell (antenna) id and the device's unique id (Ahas et al., 2010b; Wang & Chen, 2018; Wu et al., 2014). For this current study, one of Estonia's largest operator's logs are used (TRA, 2016). The dataset has both CDR and DDR, which in return gives a better temporal and spatial resolution compared to only CDR, as the dataset just has more logs of user activities thanks to the logging of Internet usage. For easier referencing, CDR is used to describe both call and data records in this thesis. An excerpt of the dataset can be seen in Table 1, where "pos_usr_id" is the device's pseudonymous unique id, "pos_time" is the time of the event in epoch format, "ci" is the antenna's unique code and "pos_type" shows what kind of network event took place. The location of the antenna is known.

Table 1. An example of the CDR/DDR dataset used in this thesis.

POS_USR_ID	POS_TIME	CI	POS_TYPE
151998471734100596	1551526364	62734	12
151998471734100596	1552108644	63232	12
81911095162569217	1553878811	47181	3

A period of two months is chosen for the test group. It is four times longer than the time frame used by Wu et al., (2014). The idea is to see if the unmasked logs from these three users can be considered oscillation. The main CDR dataset consists of over 500 000 unique subscribers over the span of one month. There are over 200 million network events and covers the entire country for that operator during that month. In other words, no urban, rural or any other selection has been done besides the time period of one month. This way the results will best describe an Estonian CDR dataset.

Three people gave their consent to use their CDR dataset to determine if the detected oscillation logs are due to the user's actual real-life movements or if the methods of this thesis find oscillation. For these three users, two months of CDR data are compared to their actively gathered location data during the same time period. The rationale behind it is to look at an oscillation and determine where the user was at that given time and whether the detected oscillation log is due to user movement or not.

All test user during that time period gathered location points through Google's location history or Apple's location history. Google location history captures GPS points, cell tower information and nearby Wi-Fi connections to assess user location (Google Maps Help). Ruktanonchai et al. (2018) found in their study that the median difference between the traditional GPS tracker collected point location to Google's location history point location is around 65 meters in favour of the GPS. Additionally, two of the three users kept a travel diary for extra precision as well. Combining those methods (user memory, GPS points, travel diaries), we get as close as possible to the real location of the user at a given time and can compare the oscillated records to real-life locations. Wu et al. (2014) used a similar assessment over a period of two weeks from four users, who kept travel diaries for comparison.

One of test user's records are only compared to GPS locations and user's memory to see the difference from using travel diaries as an extra source. It might be possible that actively gathered GPS points are good enough for comparison and more time and resource consuming travel diaries might not be needed. Here, the user still gives his input based on memory if needed.

2.2 Methods

A hybrid approach similar to the one adopted by Wu et al. (2014) and Qi et al. (2016) is going to be used. Both pattern-based logic and time-distance will be applied to find suitable oscillating candidates. The methods in this thesis are greatly influenced by Wu et al. (2014) and Qi et al. (2016). Both of those works use heuristics and this thesis combines some of them, alters parameters and uses parts from each, which are deemed more suited for the Estonian CDR dataset. Their works could be considered a basis for this thesis' methods.

For this study, the location of a cell (antenna) is aggregated to cellular tower point location. In other words, multiple cells that are situated on a single cellular tower are considered as one aggregated location, called site. As mentioned in section 1.1 and shown in Figure 1, a cellular tower (site) usually has multiple cells (antennas). If you think of one cellular tower as a circle, they can be omnidirectional cells (whole circle as one cell), but half-sector, third-sector or four-sector cells that make up the entire site circle are more common (Sauter, 2006). Those directional cells have azimuth and sector angle, but in this study, they are aggregated to an entire cellular tower (site) and are figuratively considered as an omnidirectional cell. That way it gives a higher certainty that the logs to be found are indeed oscillation as the distance between cellular towers is much larger than between two cells from the same cellular tower. Although it comes with a cost - we lose on the spatial accuracy of the logs as we do not consider the extra location the direction of an antenna provides. At the same time, doing this already eliminates some of the oscillation that might have occurred between these neighbouring cells on a single cellular tower as they are aggregated to the same location and the cell id-s are represented as cellular tower id-s (site id).

In the Estonian context the coverage areas and cellular tower spatial resolution vary greatly between urban and rural areas. In less populated areas, the coverage area of a single cellular tower tends to be larger (Sauter, 2010), and therefore the same distance threshold should not be applied to both urban and rural areas. In this thesis, additional reference information is used to divide cellular towers into urban and rural towers (appendix 1) and different distance thresholds are used for heuristics based on the location of the cellular towers in which the network event took place. The parameters for the heuristics in this thesis are static throughout this work to provide a reference point for any future works.

The average mean distance to their nearest neighbour was found for both groups and the distance thresholds for heuristics is set by doubling that:

- Urban distance threshold 2 000 meters
- Rural distance threshold 12 000 meters
- If cellular towers from both involved, then 5 000 meters

The division into urban and rural cellular towers is done based on Estonian administrative units where cellular towers in the cities are selected as urban and some manually selected cellular towers that lie in the usual hinterlands of Tallinn and Tartu are also included in the urban group. Other cellular towers are classified as rural. As this work does not try to define the difference between urban and rural areas, this grouping is subjective and should be taken as such.

2.2.1 Stable periods heuristics

Firstly, stable stay-periods are found, e.g. when the device is connected to a single cellular tower. For this study, based on the findings from Wu et al. (2014), the period must be more than 10 minutes to be classified as a stable period. Qi et al. (2016) use the same logic to find stable periods, but they do not specify what constitutes a stable period. They just say that the time difference between the first log and the last log in the same site must be greater than the set time threshold.

The first oscillating logs are found using the stay-periods. If between two stable periods, which are from the same site, there is a log from a different site in-between them in a short time interval, that log is considered an oscillation. This heuristic comes from the concept that if a mobile device is situated near one cellular tower for more than 10 minutes, jumps to another tower and then back to the first one, that device most likely did not move in real life, but the network, due to the different reasons mentioned in section 1.2, switched between sites. This is illustrated in Figure 7, where ΔT shows the maximum time difference allowed between the stable periods.

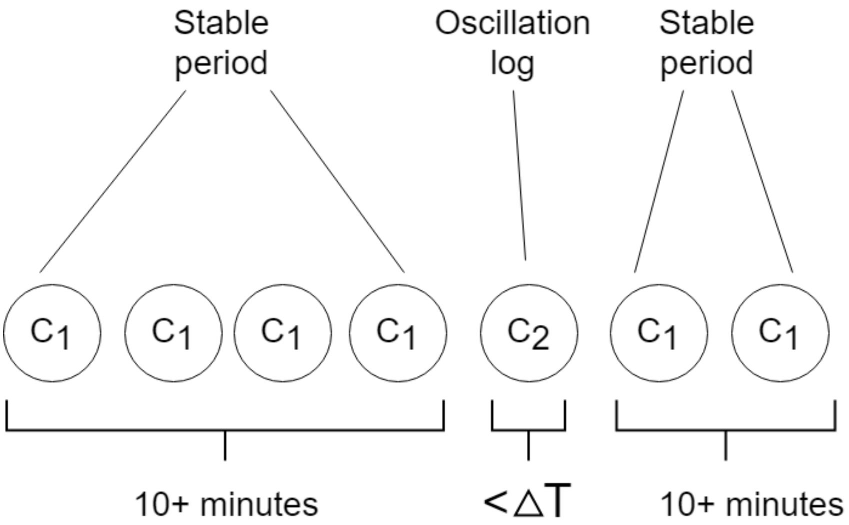


Figure 7. An oscillation log between two stable periods.

Setting a time window is quite subjective. Wu et al. (2014) use a two-minute one, but Qi et al. (2016) only us ten seconds. Both methods have their merits: Qi et al. (2016) most likely only remove oscillation, keeping all other records between two periods, which actually might be movement. On the other hand, mobile logs are temporally not continuous and depending on the dataset (if DDR is included) might be quite sparse. But setting that low time threshold only gets the most definitive oscillating logs and might not find others. Wu's approach casts a wider net to capture more. However, it might be unlikely that a device moves between two stable periods during a two-minute window, but in some fringe cases it can happen. Then a real movement might be falsely identified as oscillation. For this thesis, a compromise between the two is chosen and a time frame of 60 seconds is used as a threshold. One minute is considered short enough for a user not to move like that in real life while staying near one site for a longer period. In this instance, distance and speed are not looked at.

As a second step, stay periods are combined with time, speed and distance. If there is a log straight after a stable period and that site is too far, then the log is oscillation. The logic behind is that if a device is in one place for a longer time, it is unreasonable that a device jumps, for example, 5 kilometres to another site. The key here is to set a proper time and distance threshold to capture these kinds of logs. Time divided with distance is speed, but that is not used alone, but with an extra parameter with it. The reasoning behind it is that when a device actually does switch between sites like that, it can happen in seconds, while the distance between the sites is might only be a couple of hundred meters and then the calculated speed might seem unreasonably high for an actual regular handover between two cellular towers. For example, if the distance between sites is 300 meters and the time difference between logs is 2 seconds. The speed would be 540 km/h, which is too high for a person to travel and would be perhaps falsely considered to be oscillation or another form of noise if speed would be the only parameter.

But if we look at them separately, as Wu et al. (2014) did, or still find the speed but use an extra threshold on the distance between the switching cellular towers (Qi et al., 2016), then the aforementioned log would not be classified as oscillation as there needs to be a minimum distance between the two records. For not removing actual movements or a device entering a new coverage area, the distance threshold should be well above the average distance between the cellular towers. Most of the literature on the subject has used around 3-5 kilometres as the distance threshold in this kind of heuristic. As the location characteristics cannot be perfectly deducted from the mentioned literature, it seems that their datasets are mostly urban (Qi et al. (2016) use Chinese data). Considering the found average mean distance to the nearest neighbour, the distance thresholds mentioned in section 2.2 are used. **The abnormal travel speed is defined as over 200 km/h.**

Similarly, some logs before a stable period can be oscillation as well. The rationale here is that before a device commences a longer stay time in one cellular tower, it might have log from said tower, then in a short time window from another tower further away and back to the original one, where the device is then connected to for a longer period of time. Here, time and distance are used as parameters.

2.2.2 Sudden jumps further away and back

Thirdly, logs which switch unrealistically far away and back between cellular towers in a short period of time are found. The heuristic to capture logs between two stable periods kind of follows the same pattern of cellular tower handovers (*A*-*B*-*A* where *B* is the oscillation log), but the key difference is that the stable period one only looks at a 60 seconds time interval between the two stable periods, does not consider speed or distance as extra parameters and both first and last log need to be stay periods. The sudden jumps further away and back heuristic captures oscillation in the same *A*-*B*-*A* pattern, but does not require any of the logs to be stable periods and speed and distance are used as additional parameters.

For the reasons mentioned in previous chapter, speed by itself would not be enough as a parameter. In the literature, it is common to use the same approach (Wu et al., 2014; Qi et al., 2016). In both works the rationale is that two handovers occurring between a cellular tower, away and back to that cellular tower (or close to it), might not be reasonable user movement and most likely due to mobile network behaviour. Jumps from cellular tower A to further away cellular tower B and back to cellular tower A are considered as oscillation if they exceed set thresholds for speed and distance. For this thesis, close by towers will not be investigated. It is both for simplicity sake and to be sure that the log found is oscillation. When using close by towers, it becomes more subjective, what is close and, in the end, it cannot be determined whether the log was due to oscillation or not. To be certain that the found log is due to this phenomenon, close by towers are not investigated and only the sequences with the same start and end sites are. Setting the right parameters is quite tricky. The question here is at which point is the device’s speed unrealistically fast. Both works by Wu et al. (2014) and Qi et al. (2016) do not specify which speed threshold had been set. This heuristic follows the same distance thresholds as set in section 2.2 and uses 200 km/h and more as abnormal speed threshold.

2.2.3 Suspicious sequences and oscillation sequences

The fourth step is to find logs where in a short period of time the device has moved between multiple cellular towers, but it contains a cycle (e.g. *A*-*B*-*A*). Both aforementioned research studies approach this step similarly. They first find a suspicious sequence of logs and from that selection determine if any of the logs in the sequence are due to oscillation. A sequence is considered suspicious when it satisfies the following conditions:

- In a short period of time (e.g. couple of minutes) there are multiple logs
- More than one cellular tower
- One distinct cellular tower needs to be recorded twice in the sequence

Wu et al. (2014) set the criteria to at least three logs in one minute from at least two different cellular towers. On the other hand, Qi et al. (2016) use more than five logs and at least three different cellular towers. But as mentioned before, not all suspicious logs are oscillation. Here comes the main difference to how both researches determine oscillation from them. Wu et al. (2014), after determining a suspicious sequence, expand it up to a minute before and after until it encounters a new cellular tower that was not in the original sequence or the time threshold ends. On that expanded sequence, they determine if there are any logs which are in a circular manner, e.g. let's say an expanded sequence is like this *A – B – C– A– E*. For their algorithm, there would be oscillation for cellular tower *A.* Keeping in mind that it happens during a maximum of three minutes (one minute for original suspicious log and one minute to each side of the log). As there might be multiple circular cycles, a score-based method is applied to determine which cellular tower is most suitable to represent that sequence. Only the logs from that tower remain, others are removed.

As Qi et al. (2016) already set themselves up to have at least 5 logs in a short time frame (which they unfortunately do not specify), they do not expand their selection of a suspicious sequence. To be considered oscillation, the sequence must have a distinct cellular tower more than once and the distance between each pair of cellular towers in the sequence does not exceed a set threshold. If those conditions are satisfied, the sequence is due to oscillation. To find which site represents that sequence the best, a weight is set for each site on how many times it is encountered. After that a weighted centroid is calculated for that sequence. The closest cellular tower to that centroid is selected and logs from other towers in that sequence are classified as due to oscillation.

In this thesis the aforementioned approaches are somewhat combined. A suspicious sequence needs to have at least 2 distinct cellular towers and four or more logs. The time window here is in a sense as Wu et al. (2014), where the first selection (2 towers, four logs) is expanded up to any logs that are less than 180 seconds from the last until the next one is more than that. To determine the main cellular tower for each sequence, a weighted mean point is calculated and as with Qi et al. (2016), the tower closest to that point is set as the main tower and logs not originating from that, are classified as oscillation.

2.3 Analysis steps

The data processing was done using PostgreSQL (ver. 10.5), its extension PostGIS (ver. 2.4.4), GIS software QGIS (ver. 3.6) and ArcMap (ver. 10.4). The heuristics are written in Structured Query Language (SQL).

Before applying any heuristics, some pre-processing of the CDR dataset is done. The network event logs are aggregated from the antenna level to the cellular tower. After that the cellular towers are divided into two groups of urban and rural. For every unique device in the CDR dataset, stable periods of 10 minutes are calculated. A device’s logs are divided into sequences based on cellular tower switches ordered by timestamp. A sequence is created when the device switches to a cellular tower and ends when the device switches to another one. There can be multiple logs from the same cellular tower in a sequence. When the device switches to another cellular tower, a new sequence is created. For these sequences time intervals are found between the first and the last log in the sequence. The sequence is classified as a stable period, if that difference is over 10 minutes.

The heuristics are independent and mostly find different cases of oscillation, meaning that they can be applied separately, and besides the pre-processing steps, can be run in any order. Although it would be recommended to run the suspicious sequences and oscillation sequences (section 2.2.3) heuristic last as it only finds the weighted mean point for oscillation sequences where there were no previous oscillations present. It is done so as not to double the same work some previous heuristic had already done. In some cases, multiple heuristics might capture the same oscillation. That does not change the overall amount of oscillation unmasked but might over-emphasise the number of logs a single heuristic captured.

An overview of the workflow is shown in Figure 8. The general approach is to pre-process the CDR/DDR dataset to be able to run heuristics on it. After that, heuristics are applied, and the suspicious sequences and oscillation sequences heuristic will be run last. Then the detected oscillation logs of the three test group users are compared to their real-life location at the same time the oscillation took place, in order to assess whether the unmasked log might have been due to user movement and not due to a sequence of network handovers. Based on the results of the test group, the heuristics are assessed if they are applicable or not. There might be cases where some heuristic captures for example too many actual movements as oscillation or the results include some unusual findings. The unusual part might be that some heuristics do not capture any oscillation or too few. As heuristical methods are more on the practical side, the assessment of the heuristics will be conducted in the same vein.

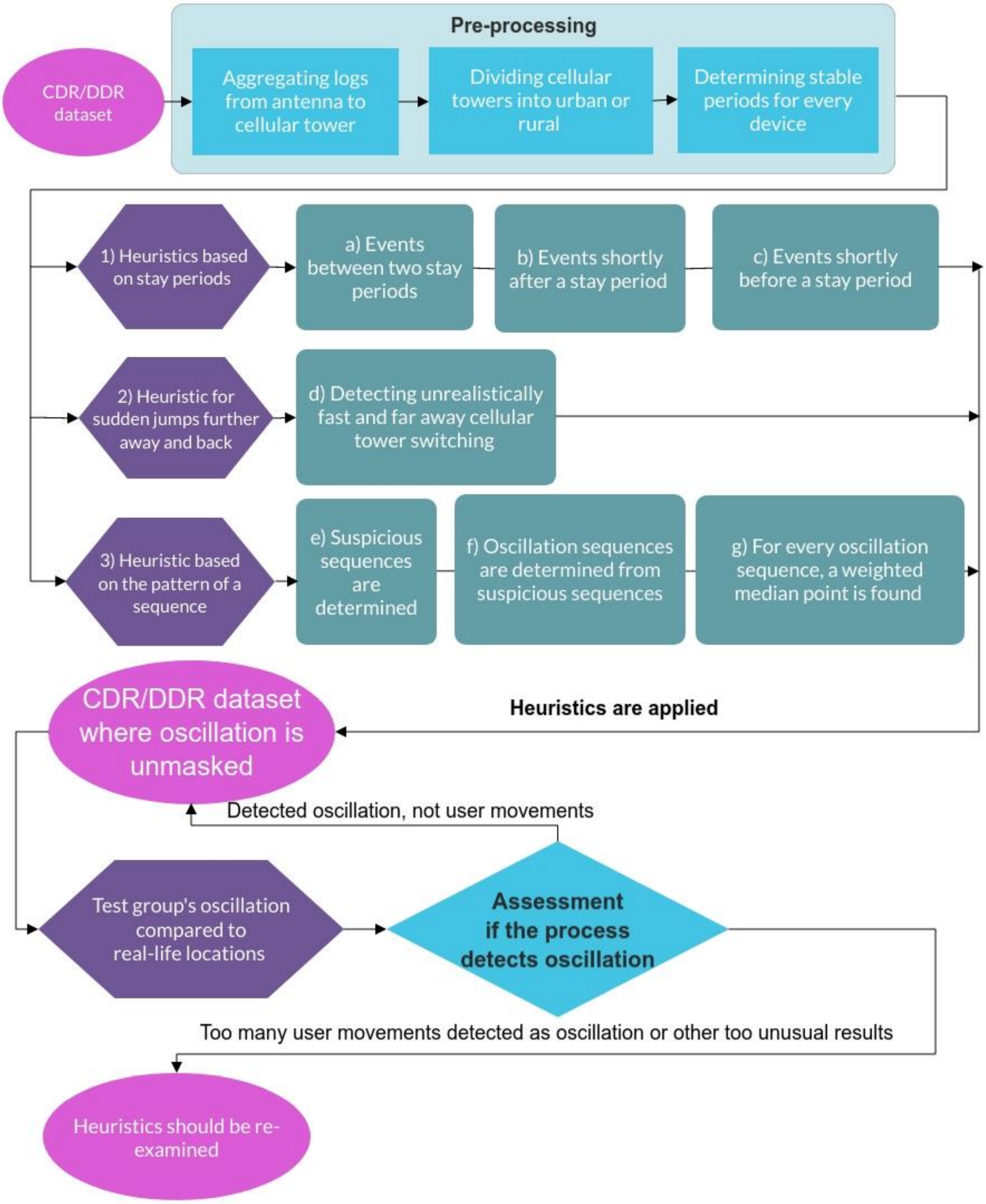

Pre-processing
CDR/DDR dataset
Aggregating logs from antenna to cellular tower
Dividing cellular towers into urban or rural
Determining stable periods for every device
1) Heuristics based on stay periods
a) Events between two stay periods
b) Events shortly after a stay period
c) Events shortly before a stay period
2) Heuristic for sudden jumps further away and back
d) Detecting unrealistically fast and far away cellular tower switching
3) Heuristic based on the pattern of a sequence
e) Suspicious sequences are determined
f) Oscillation sequences are determined from suspicious sequences
g) For every oscillation sequence, a weighted median point is found
Heuristics are applied
CDR/DDR dataset where oscillation is unmasked
Detected oscillation, not user movements
Test group's oscillation compared to real-life locations
Assessment if the process detects oscillation
Too many user movements detected as oscillation or other too unusual results
Heuristics should be re-examined

Figure 8. Workflow for unmasking oscillation from the CDR/DDR dataset.

3. Results

The results are given based on their heuristic, where first the assessment of the heuristic is given based on the test users and then the amount of oscillation captured with that heuristic on the main CDR/DDR dataset is shown.

The test group's three subscribers' CDR/DDR data for the time period of two months consisted of 2784 subscriber network events, where 47 of them were detected as oscillation. That's about 1.7% of the test group's dataset. The main dataset consists of over 200 million events over a period of one month. Out of those, 2.2 million were detected as oscillation. This is about 1% of the entire CDR/DDR dataset for one operator in Estonia.

3.1 Oscillation related to stable periods

A stable period is defined in this thesis as a ten-minute period, during which the device does not change cellular tower. Three heuristics are applied to find oscillation, which are related to stay periods. The first one follows the logic that if there are two stay periods from the same cellular tower and in a short time window there is a log in-between those periods (Figure 7). In this work, one minute is set as that threshold. With this heuristic, two logs are found. An example from the test group's dataset is shown in Figure 9. There are two stable periods originating from cellular tower 658, but in-between those periods is an event that happens fast and is further away than the set threshold of 2000 metres for urban setting. Both oscillations are reasonably assumed to be due to network handovers and not deemed as user movements. Here, the approximate location of the user would be the cellular tower, in which the two stable periods took place as the middle log is due to oscillation. The heuristic was then applied to the main dataset and more than 75 000 events were detected as oscillation (Figure 13).

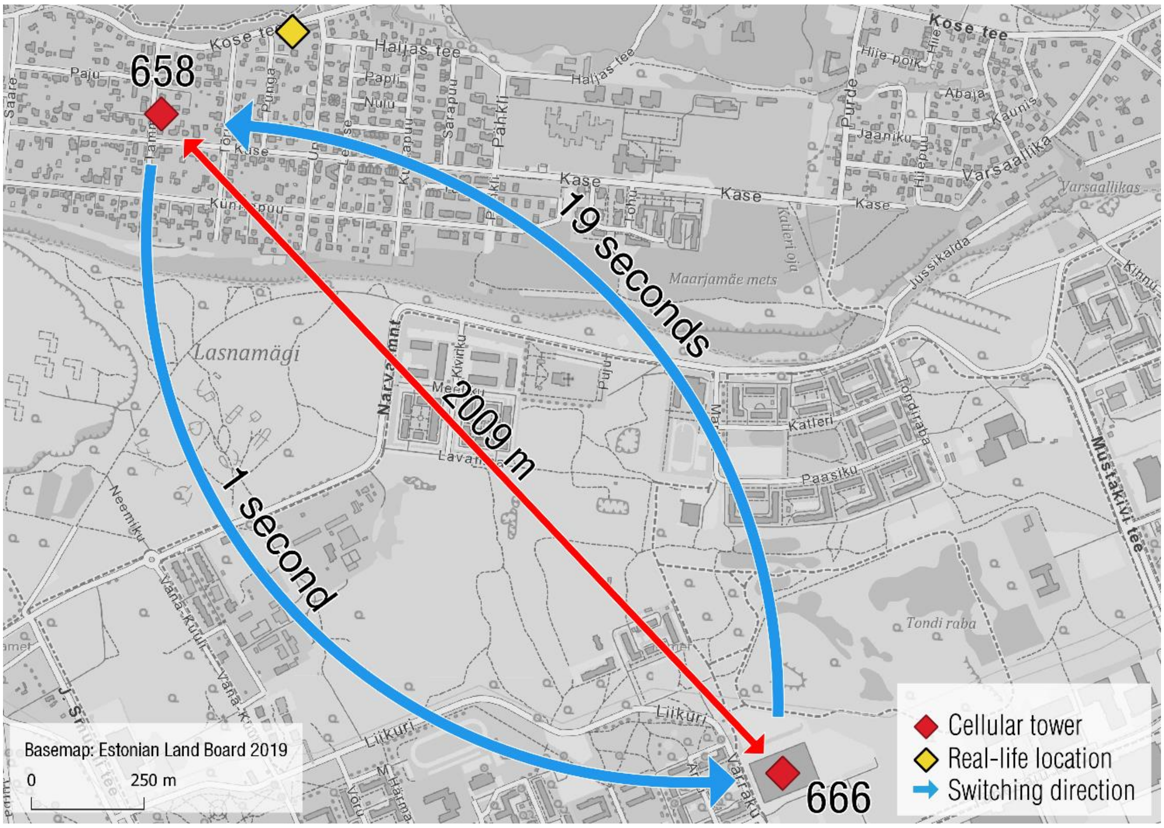


Figure 9. A log from tower 666 in-between two stay periods from 658. It all happens in less than 60 seconds.

The second heuristic involving stay periods looks for logs before the stable period begins. According to Qi et al. (2016), it is common for a device to connect to a cellular tower, then another cellular tower, which is further away and then back to the first one where a stable period then commences. The rationale is that the device did not move, and the first log should be included for the stable period and the middle one is due to oscillation. Again, a threshold of time is used here with distance and speed as extra parameters. This heuristic only finds one log, which is the same as shown in Figure 9. Although the other heuristic finds the same one, it satisfies the conditions for this one too. Here, the approximate location of the user would be the first and last cellular tower, as the middle one is due to oscillation. When applied to the main dataset, 50 700 events are detected as oscillation (Figure 13).

The third heuristic, which involves stable periods, is determining if the follow-up log after the stay is oscillation or not. Here speed, time and distance are combined as parameters. If a log happens shortly (e.g. 1 minute) after a stable period and the device's distance travelled is over the set thresholds and the speed it travelled with is over a reasonable amount (200 km/h), then the entry is considered oscillation. With this heuristic, 13 logs are found, which is almost 0.5% from the test group’s dataset. An example can be seen in Figure 10.

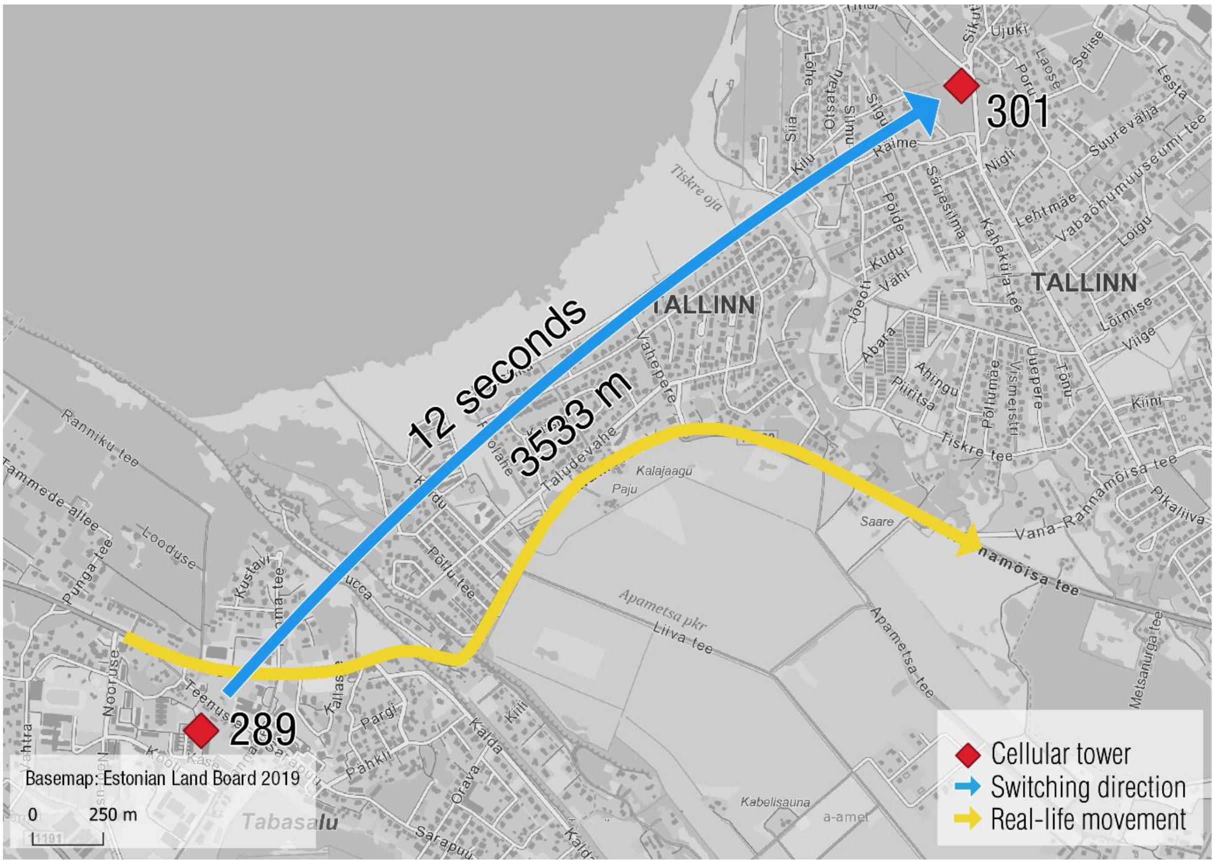


Figure 10. A stable period in the coverage area of cellular tower 289 is followed up with a log in 12 seconds from cellular tower 301, which is 3533 meters away. The user travelled by car in the shown direction.

As seen from the above Figure 10, this heuristic captured an actual movement the user did by travelling with a car on a highway. Although the user spent over 10 minutes in the coverage area of cellular tower 289, after starting their trip, it switches to the other tower 301. Out of the 13, only 7 can be considered oscillation. The other 6 are similar to the case in Figure 10, which are caused by travelling with a car at higher speeds. In this case, the approximate location of the user for these 2 logs would be the first cellular tower, where the stay period took place.

This heuristic does not capture only oscillation but was still used on the main dataset to get a reference point. When applied to the main dataset, 409 702 network events were detected as oscillation (Figure 13).

3.2 Oscillation involving sudden jumps further away and back

Compared to oscillation related to stable periods, these cellular tower switches from one tower to another and back to the first one are really common. 160 logs in our test dataset follow this kind of a pattern, but it does not necessarily make them oscillation. The key here is to determine if the jump happened too fast and too far to be considered real-life movement. So, to find oscillation from those logs, speed and distance are used as parameters that were defined in section 2.2. Out of those 160 logs, only 3 fit this heuristics criterion. The same switching of cellular towers illustrated in Figure 9 is detected by this heuristic: a jump to a tower further away and then back to the first one while exceeding the set speed and distance thresholds during a short time frame. This heuristic does not need any of the logs to be stable periods. As the middle log is determined to be oscillation, the approximate location of the user is the first and last cellular tower for those three 3 logs. When applied to the main dataset, over 185 000 network events were detected as oscillation (Figure 13).

3.3 Oscillation involving suspicious sequences and oscillation sequences

To capture oscillation, which happens in a short time window and involves multiple different cellular towers, suspicious sequences are found that have at least two distinct cellular towers and four or more logs in that time span. Not every suspicious sequence contains oscillation. From that selection, a cycle of cellular towers is identified and only those sequences are considered containing oscillation. 61 suspicious sequences are found, which consist of 318 logs. Out of those 61 sequences, 13 contain oscillation with 78 logs. In those sequences, the ones not originating from the main cellular tower are unique oscillation logs. If a suspicious sequence did not previously contain a log due to oscillation found by previous heuristics, a weighted mean point between the cellular towers was found and the closest cellular tower in that sequence was chosen as the main one. Weight comes from the number of times a cellular tower appears inside the sequence. Every log that takes place in that sequence and is not from the main cellular tower, is considered an oscillation log. For 11 sequences a main site was found, and 35 logs were determined as oscillation (Table 2). The main dataset’s sequences follow the same pattern, where the share of oscillation sequences make up a smaller minority of the suspicious sequences.

Table 2. Number of sequences and logs by steps to detecting oscillation from suspicious sequences in the test group's dataset

	Number of sequences from the test group dataset	Number of sequences from main dataset
Suspicious sequences	61	2 300 00
Oscillation sequences	11	670 000

Comparing these final oscillation sequences with the real-life location, it can be determined that in general it captures oscillation. An example of a detected oscillation sequence can be found in Figure 11.

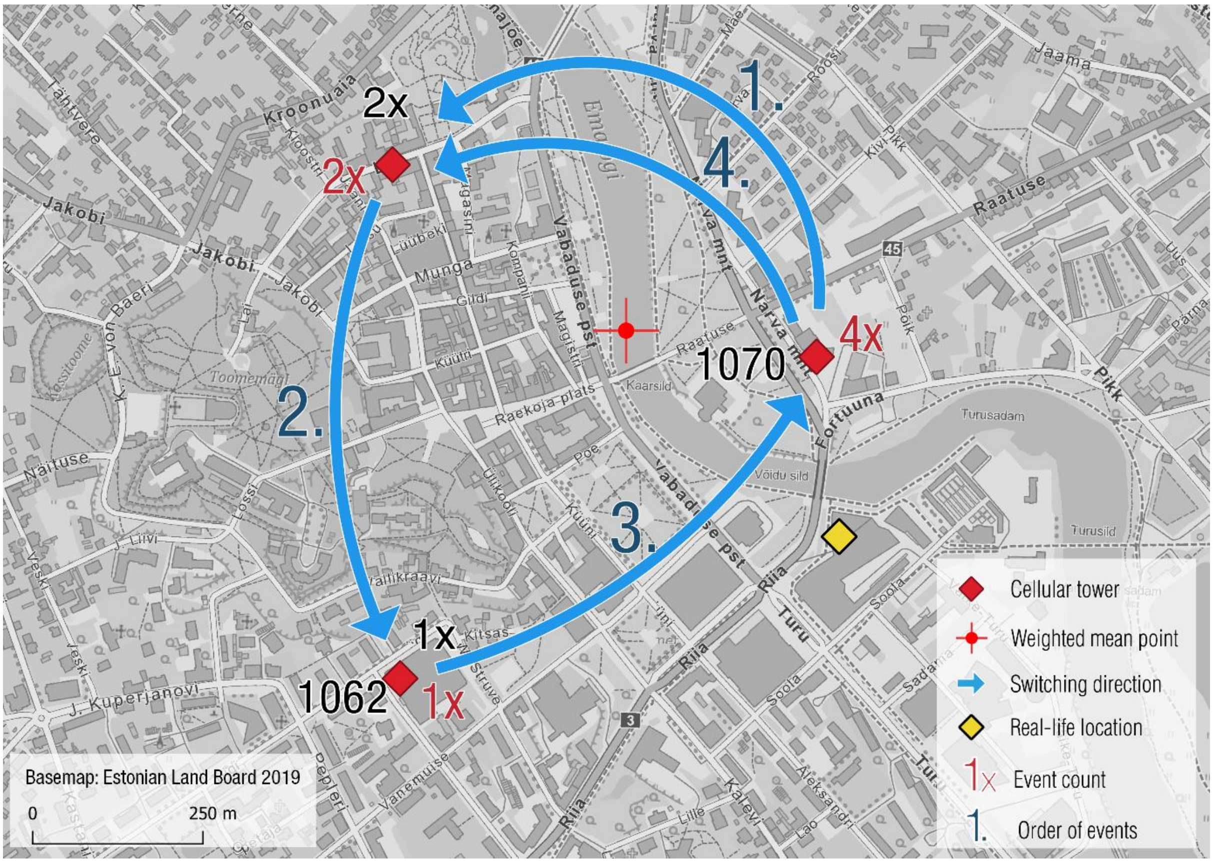


Figure 11. Starting from cellular tower 1070, the device switches to tower 1065 then to 1062 and back to 1070 all in under 2 minutes. With red text, the number of events in each tower is shown and how the weighted mean point is skewered towards 1070 from that.

As seen from the above example (Figure 11), the cellular tower switching does not correspond to the user's location or movement at that time interval and oscillation does occur. With the weighted mean point, the cellular tower 1070 is chosen as the main site for this oscillation sequence, which indicates the actual location of the user during those 2 minutes the best and is used as the approximate location of the user. Other oscillation sequences are similar in their pattern and time frame, except one.

One of the oscillation sequences (Figure 12) contains both logs due to oscillation and user movement as well. The sequence contains 13 logs during the period of 8 minutes, while the user was travelling by car on a city street during rush hour. The following order of cellular towers illustrates the handovers:

- *337→338→327→338→343→374→405→407*

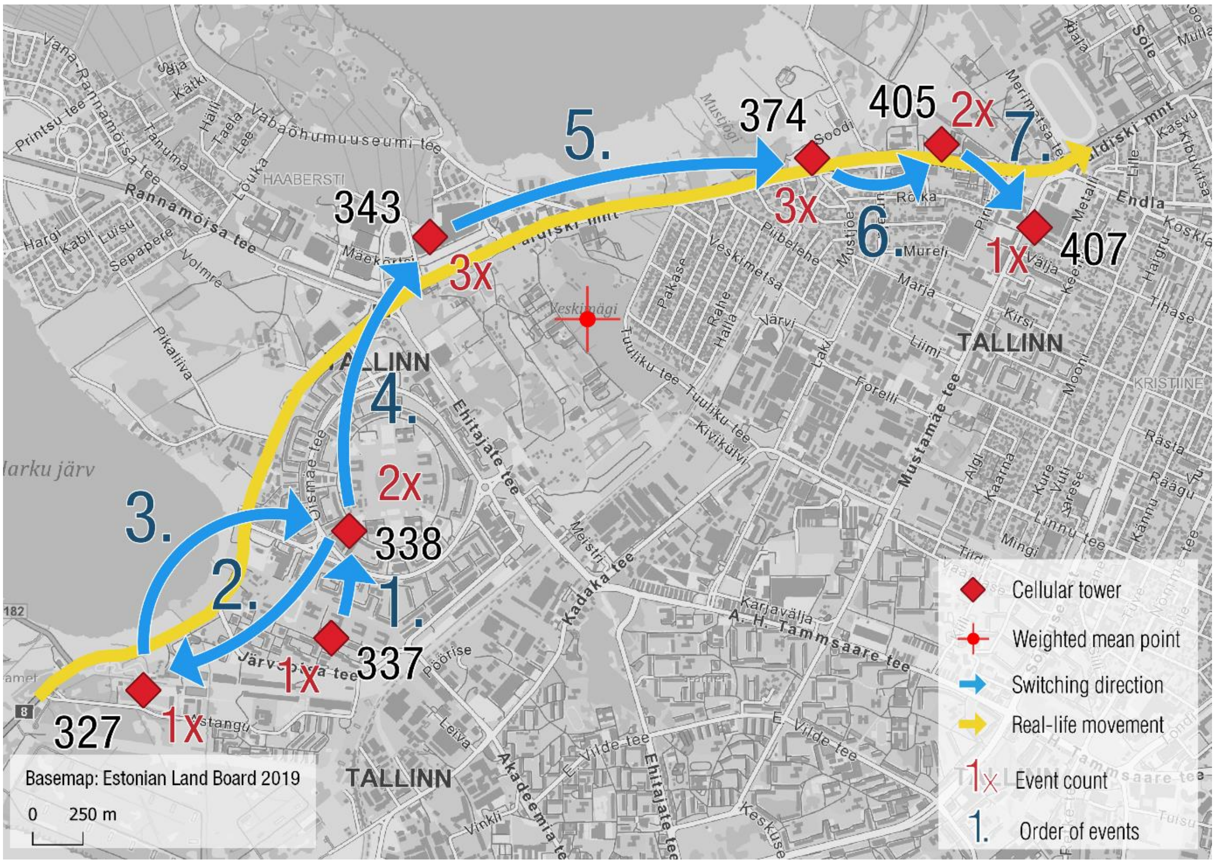


Figure 12. A detected oscillation sequence that does contain oscillation but also actual user movements.

The criterion for an oscillation sequence is the presence of a cycle of cellular towers. The first part of the sequence (*337→338→327→338*) does indeed contain one, where the cellular tower 338 would be considered the main site and the log in-between from 327 would be due to oscillation. Logically, the logs after the last log from 338 should not be in the sequence and the cut off should be made not to include them. The heuristic should be improved not to capture all of them, only the oscillating part. But for this thesis, an exception is made, and the heuristic will be used as such. The reasoning is that it did capture oscillation, just more logs than necessary were captured while doing so and for future use, the heuristic will be re-examined.

When applied to the main dataset, 1 600 000 network events were detected as oscillation (Figure 13).

3.4 Characteristics of oscillation events

The main dataset consists of over 200 million network events, which can be divided into three groups of outbound events: calls, SMS and mobile communication data. The division of the entire dataset is 13%, 1% and 86%, respectively (Table 3).

Table 3. Distribution of events in a dataset based on network event

	Calls	**SMS**	**Mobile Data**
Test group’s oscillation logs	19%	0%	81%
Main dataset’s oscillation logs	21%	1%	78%
Main dataset	**13%**	**1%**	**86%**

The majority of network events are mobile phone data usages (86%), which unsurprisingly also involve the most oscillation. But oscillation involving calls is higher than the overall percentage from the main dataset might suggest. In both the test group’s (19%) and main dataset’s (21%) oscillation logs, calls constituted a bigger part on the expense of mobile data. SMS network events remain the same percentage through the main dataset and oscillation logs. Different generations of mobile network are sometimes used for calls and data. Some phones do not even support 4G calls and must switch to another cellular tower for calls, which might not be the closest one to the device or be the previously connected cellular tower and oscillation might occur. It might be one of the reasons to explain the rise of oscillation related to calls compared to the overall dataset.

To fully capture oscillation, all the heuristics must be applied as they mostly capture different aspects of the phenomenon. After applying them on the main dataset, the total amount of oscillation logs captured was 1% out of the main dataset. The heuristic involving suspicious and oscillation sequences captures the most (73%) from the total amount of oscillation logs and the heuristic to find oscillation logs before stay periods detected the least amount (2%) of oscillation per heuristics (Figure 13).

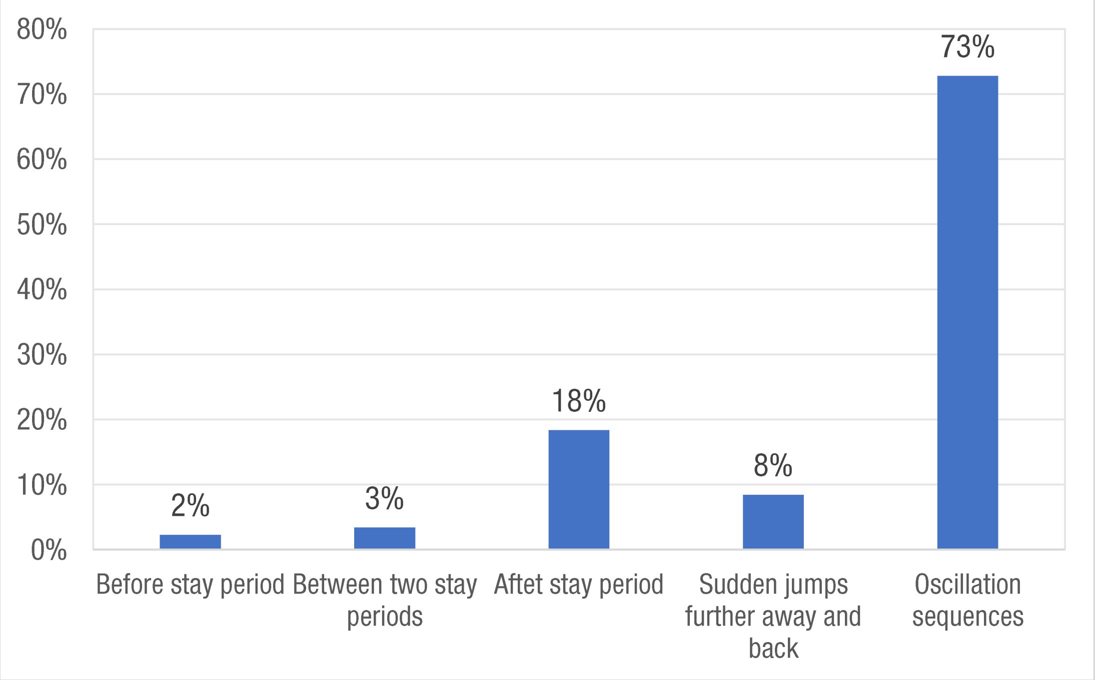


Figure 13. Each heuristic’s share of detection from the total amount of oscillation unmasked.

Because some heuristics overlap, the overall share is 105%, which show that 5% of oscillation logs are found by multiple heuristics.

3.4.1 Spatial and temporal distribution of oscillation

In order to visualise how oscillation is spatially distributed, the percentage of oscillation from the total network events per cellular tower was found. For better insights, the cellular towers were aggregated to the Estonian administrative settlement level (Figure 14). The areas in grey do contain some oscillation in different settlements, but to highlight more prominent areas, really low values are filtered out. As expected, urban areas have more oscillation than rural areas. The same goes for more populous cities. One smaller town, Räpina, pops out from the figure. That area has 3 cellular towers in quite close proximity, which in turn might cause more handovers and therefore oscillation.

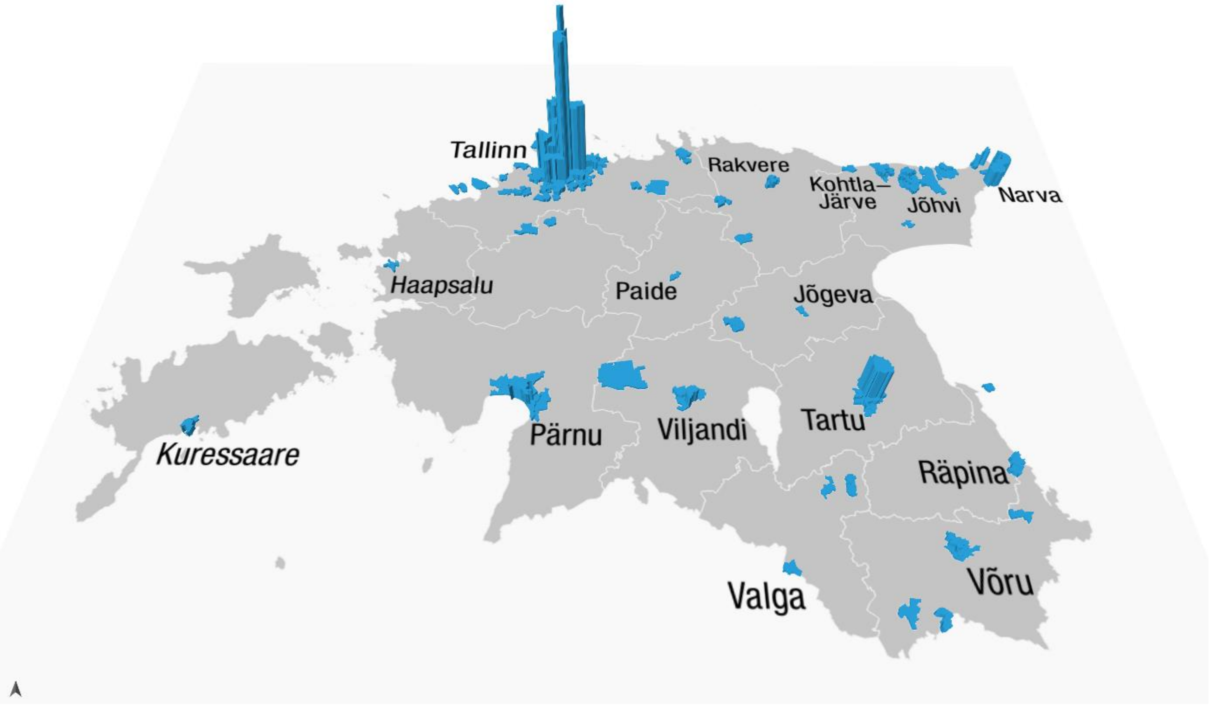


Figure 14. Percentage of oscillation from total network events per cellular tower. Aggregated to the settlement administrative level. Grey areas contain a minimal amount of oscillation, blue areas significantly more.

Overall, the visualisation supports the conclusion that more oscillation happens in urban areas, where the population is denser. In order to satisfy the needs of the subscribers, more cellular towers are in closer proximity and overlap in their coverage areas, thus provoking more handovers. As well as there are more people densely populated in urban areas, there is more load balancing on the mobile network side, which needs to be considered.

In the temporal dimension, oscillation has some visible patterns. If the main dataset's oscillation logs are divided by days of the week, a clear peak can be seen on Friday (20.9%). It is followed by low values during the weekend (Saturday 12.9%, Sunday 10.4%).

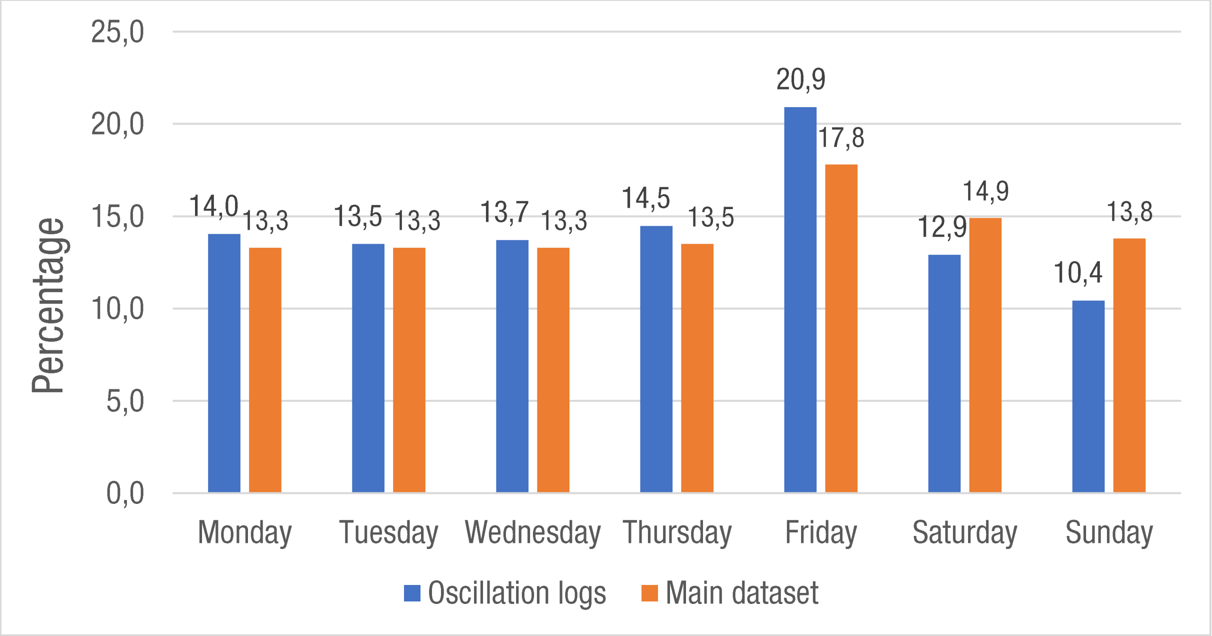


Figure 15. Share of oscillation for every day of the week from the main dataset and the distribution of main dataset's network events for each day.

When daily oscillation share is compared to the main dataset one, working days have slightly higher percentages with both having a clear peak on Friday. But on the weekends the percentage of oscillation falls compared to the amount of network activity.

If the share of oscillation in the diurnal aspect is looked at, a steady incline from morning hours up to 16:00 can be observed (Figure 16). The same kind of pattern that is present in the oscillation diurnal cycle, can also be observed in the main dataset where there is an incline from morning hours to the early hours of the evening. But the main dataset is more evenly distributed throughout the day and the peak (6.5%) is less than the peak of oscillation logs.

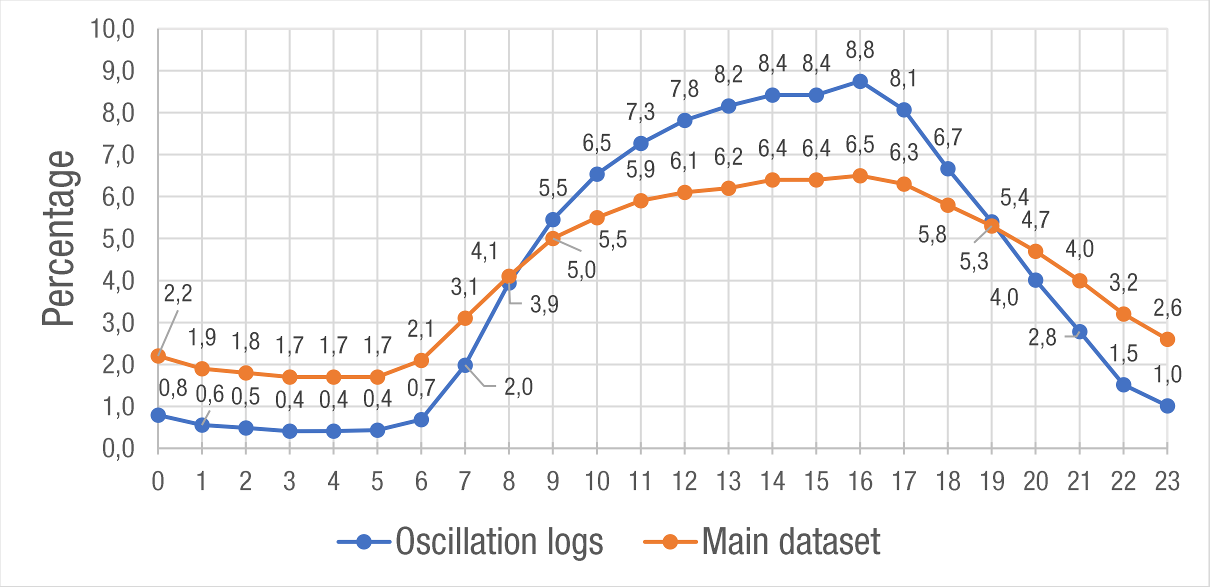


Figure 16. Diurnal pattern of oscillation from the main dataset and the diurnal pattern of the main dataset.

There might be two explanations for this kind of temporal distribution. In both daily and diurnal pattern, the oscillation logs follow the general pattern of the network load (the number of total events). An assumption can be that as a certain amount of network activity is achieved, more load balancing happens in the mobile network and a by-product of the balancing is oscillation.

In diurnal aspect, the main dataset is more evenly distributed over its time period compared to the oscillation logs, which has a steeper incline and decline with a higher peak. It is reasonable to assume that during night-time and late at night, people are most likely at home and less movement occurs in the mobile network. As most oscillation does not involve stay periods (Figure 13), the hours of the day where people are perhaps more mobile (Järv et al., 2007; Järv et al., 2012) produce more oscillation logs, too. In the late afternoons, and early evenings people are usually on their way home from work or school and are most likely more engaged with their mobile phone as compared to the morning period when they go to work or to school.

The percentage of total network events does not fall during the weekend when compared to Monday-Thursday, but the percentage of oscillation logs do. Mobility might be the explanation for the difference between them. As researchers suggest (Järv et al., 2007; Järv et al., 2012), people are more mobile in Estonia during the working days and especially on

Fridays, with less movement on the weekends. In general, movements in the mobile network bring uncertainties and the network must deal with more handovers and balancing than with stationary devices.

4. Discussion and conclusions

Passively generated mobile phone location data is a useful tool to study human mobility. But due to mobile network load balancing, buildings blocking the signal or other reasons such as poor weather conditions, handovers between cellular towers might occur without the device really moving in real life. Unmasking these kinds of handovers, where the device did not move (oscillation), can be one of the steps to raise the quality of a CDR/DDR dataset and thus improving any analysis done based on that dataset.

The main goal of this thesis was to create a method, which can unmask oscillation from a CDR/DDR dataset. The works of Wu et al. (2014) and Qi et al. (2016) were the foundations of this thesis. Both approached the problem similarly by applying practical heuristics to a CDR dataset to uncover oscillation. Compared to other research studies mentioned in the theoretical chapter (1.3 Oscillation), they combined a pattern-based approach with other parameters such as speed, time and distance. In this thesis the same concept was applied, and practical heuristics were developed based on the research of Wu et al. (2014) and Qi et al. (2016) and on testing three users' CDR data. The parameters for the heuristics in this thesis were defined in the methods section and were static throughout this work to provide a reference point for any future works.

Three broad categories of oscillation were defined as:

- oscillation related to stable periods
- oscillation related to unrealistically far away cellular tower handovers
- oscillation related to a short time interval with multiple logs from multiple different cellular towers

Every heuristic developed for this thesis captures a different type of oscillation. In other words, they can be applied separately. Nevertheless, it should be noted that some sequences of logs can be multiple types of oscillation and are captured by different heuristics. However, it does not change the overall amount of oscillation found from a CDR/DDR dataset.

Almost all the heuristics captured oscillation and not user movement when applied to the test group's dataset, except for one heuristic involving stable periods. To detect oscillation straight after a stable period, logs that were too far away in a short time window with travelling speed unreasonable to be actual movement, were selected from the test users' dataset. The comparison of these logs and real-life location of the user indicated, that in over half of the cases the heuristic captured user movements and the log was not due to oscillation. From the remainder of the selection, some were also captured by other heuristics as oscillation. The suggestion for any future works is to raise the values of the parameters of this heuristic for it to be more viable and to expand on the size of the test group.

Most of the oscillation logs were evident enough without even comparing with real-life locations. The travel diary entries were useful, but so were the gathered GPS points. To further expand on the test group's size, the collection of GPS points is sufficient enough to evaluate the detected oscillation logs. Without using travel diaries, it is easier to collect the comparing data, which also makes the process less time and resource consuming and less dependent on the participant.

To determine, how much oscillation affects an Estonian CDR/DDR dataset, the method was used on the country's entire user base of a single operator for a period of one month. This way, the most objective sample of users and spatial distribution of Estonia was obtained. The literature on oscillation suggests that there is at least 6% and up to 16% (Iovan et al., 2013; Qi et al., 2016; Wu et al., 2014) of oscillation in a CDR/DDR dataset. The findings of this thesis on both the test group's and the entire dataset suggests that the number might be lower in Estonian setting - 1% from the total CDR/DDR dataset. Although the parameters were different, they were not stricter than Qi et al. (2016) use and mostly on par with Wu et al.'s (2014), so that can not explain the difference that well.

Bayir et al. (2010) suggest that a considerable amount of oscillation occurs in metropolitan areas. Iovan et al. (2013) use only call detail records from the Parisian region in France, Qi et al.'s (2016) dataset comes from a Chinese telecom and it is reasonably assumed that it is also most likely urban. As for Wu et al. (2014), they do not specify their dataset's location, but from one of the figures it also seems to be urban.

The population density of Estonia is only 30.3 people per km^2 and in the cities it is 1328.4 people per km^2 (Statistics Estonia). The Parisian region in France that Iovan et al. (2013) use has 971 people per km^2 (Wikipedia: Demographics of Paris) and the Chinese population density is around 150 people per km^2 (Li et al., 2018) with bigger cities like Shanghai having 4200 people per km^2 (Wikipedia: Shanghai). A denser urban environment means that a larger portion of subscribers need to be serviced in an area and more cellular towers in closer proximity are necessary when compared to, for example, the rural environment. It is reasonable to assume that more cellular towers in closer proximity provoke more network load balancing due to the amount of people or buildings blocking the signal to the closest tower. This induces more handovers and therefore possible oscillation. As seen from the population density, the Estonian dataset covers a less densely populated area than Iovan et al. (2013), Qi et al. (2016) and Wu et al. (2014) had. This might be one of the reasons the amount of oscillation is higher in their works than in the Estonian CDR/DDR dataset. As seen from the visualisation of spatial distribution (Figure 14), the share of oscillation is concentrated to urban regions in Estonia, which does indeed suggest that the phenomenon is more prevalent in metropolitan areas.

Another aspect is that in heuristics close by cellular towers were not investigated as was in Qi et al. (2016) and Wu et al. (2014). For example, in a sequence of cellular tower switches *A*-*B*-*A* the first and last log come from the same tower. If close by towers would be considered, the sequence might be *A*-*B*-*C*, where cellular tower *C* is deemed close to tower *A* and would be used as if the first and last logs came from the same cellular tower. That might skew the overall percentage of oscillation to higher values as more possibilities are opened that way. Nevertheless, determining whether a cellular tower is near enough to another is subjective and quite dependent on the spatial distribution of cellular towers. It would require more testing and was not implemented in this thesis.

The literature on the subject suggested a considerable amount of oscillation is present in a CDR/DDR dataset, which would have enabled to adjust the methods on plenty of examples from the test group's data even with the small sample size (Wu et al., (2014) also only used 4 people). As the results indicate, it is not the case in the Estonian setting. Unfortunately, time and access to people's CDR/DDR data and their actual location data for the same period is limited and dependent on their consent. To improve the method of unmasking oscillation, a more comprehensive sample group would greatly benefit any future development of the method.

The temporal dimensions (daily and diurnal) of oscillation in the Estonian CDR/DDR dataset might suggest that oscillation follows, to some degree, the mobile network load (Figure 15, Figure 16). The reasoning might be that if more network events are happening at the same time, there is a stronger need for network balancing, thus more handovers between multiple cellular towers might occur in a short time interval. If these handovers happen in a manner where the movement can be determined not to have occurred in the way the logs suggest, then oscillation occurred.

The temporal patterns of oscillation also correspond to people’s mobility (Järv et al., 2007; Järv et al., 2012). A clear peak on Fridays can be seen from the oscillation logs. During the weekends the overall network activities do not fall below the working days, but oscillation logs do. As people are generally less mobile during the weekend (Järv et al., 2007; Järv et al., 2012), there might be fewer handovers occurring due to movement and thus less chance for oscillation. Mobility likely plays a role how oscillation occurs. As an example, the connection to the nearest cellular tower might be temporarily blocked by a building or other obstruction during the movement and a handover to another cellular tower might occur. After the blockage has cleared it might connect back to the closer one. If it happens in a short time interval, then an oscillation occurred. As this is just a preliminary and not in-depth analysis of the spatial and temporal dimensions of oscillation, it would be intriguing to dive deeper into the patterns and causes of them.

Overall, the thesis achieved its goals: a practical method for unmasking oscillation from a CDR/DDR dataset, determined how much oscillation is present in an Estonian dataset and how does oscillation vary in temporal and spatial aspects. There is no doubt that oscillation is present in passively gathered mobile positioning data as the thesis’ data analysis also shows. Oscillation might cause misinterpretation of human mobility as it captures movements that did not occur in real life. By unmasking oscillation and approximating the location of a user more precisely, it is possible to improve the quality of the dataset and any analyses done with it.

Mastiviskamise tuvastamine mobiilpositsioneerimise andmetest

Sander Pukk

Kokkuvõte

Viimasel aastakümnel on mobiiltelefonide arv enam kui kahekordistunud. Kui 2007. aastal oli unikaalseid kasutajaid hinnanguliselt umbes 2 miljardit, siis 2017. aastaks on see arv tõusnud juba üle 5 miljardi (GSMA Intelligence, 2017). Märkimisväärselt on kasvanud mobiilse interneti kasutamine ainuüksi Eestis, kus Tehnilise Järelevalve Ameti raporti (TRA, 2017) järgi on mobiilse interneti kasutamine tõusnud praktiliselt 13 korda. Lisaks tavapärastele kõnedele ja sõnumitele, hoiustatakse ka mobiilse andmeside toimingute logisid operaatori poolt. Nendest logidest on võimalik määrata seadme asukoht toimingu aja ning antenni identifikaatoriga, võttes aluseks mobiilimasti asukoha, millel see antenn asus. Selliseid andmeid nimetatakse passiivseteks mobiilpositsioneerimise andmeteks, mida on laialdaselt kasutatud inimeste ruumiliste ning ajaliste liikumiste uurimiseks (Ahas et al., 2008; Ahas et al., 2010a; Ahas et al., 2014; Bayir et al., 2010; Furletti et al, 2014; Girardin et al., 2009; Gonzalez, 2008; Kung et al., 2014; Lee & Hou, 2006; Raun et al., 2016).

Enamasti on seade ühendatud lähima mobiilimastiga, mille signaal on kõige tugevam. Mastiüleandmine toimub siis, kui teise mobiilimasti signaal on teatud määral üle hetkel ühenduses olevast (Corazza et al., 1994; Iovan et al., 2013; Sauter, 2010). Tavapärases situatsioonis on see põhjustatud sellest, et kasutaja lihtsalt liigub ühe masti levialalt teise ning toimub mastiüleandmine (Sauter, 2010). Aga on mastiüleandmisi, kus päris elus liikumist ei toimu. See võib olla tingitud mobiilivõrgu koormuse jagamisest erinevate mastide vahel, ilmastikust tingituna või näiteks on signaal blokeeritud lähimast mastist hoonete või muude takistuste tõttu. Neid olukordi, kus kasutaja lühikese aja jooksul näiliselt liigub mitme masti vahel ja tõenäoliselt tegelikult ise asukohta ei vaheta päris elus, nimetatakse mastiviskamiseks (Bayir et al., 2010; Chen et al., 2016; Positium, 2017; Wu et al., 2014; Qi et al., 2016). Antud teema olulisus seisneb andmekvaliteedi tõstmises, puhastades andmestikku ning määrates inimeste tegelikud asukohad mastiviskamise tuvastamisel paremaks analüüsiks.

Käesoleva magistritöö eesmärgiks on meetodi väljatöötamine mastiviskamise tuvastamiseks passiivsetes mobiilpositsioneerimise andmetes, võttes aluseks Wu et.al. (2014) ja Qi et.al. (2016) uurimusi antud teemal ja rakendada seda Eestis kogutud passiivsetel mobiilpositsioneerimise andmetel. Lisaks uurida lühidalt võimalikke ajalisi ja ruumilisi aspekte antud nähtuse puhul. Töö alusandmestikuks oli ühe Eesti suurima operaatori passiivse mobiilpositsioneerimise andmed terve kliendibaasi kohta ühe kuu jooksul (TRA, 2016). See sisaldab endas ka mobiilseid andmeside toiminguid ning koosneb rohkem kui 200 miljonist võrgutoimingust. Antud andmebaasist tehti kirjaliku nõusolekuga kolme kasutaja väljavõtted kahe kuu kohta, kelle mobiilpositsioneerimise põhjal leitud mastiviskamist kõrvutati nende päris asukohaga sellel ajahetkel. Sel viisil tuvastati, kas leitud võrgutoiming oli tõesti mastiviskamise tõttu või oli selleks mingi muu põhjus, näiteks kasutaja tavapärane liikumine ühest masti levialast teise. Vastavalt teoreetilisele raamistikule ja eelnevalt tehtud uurimustele, töötati välja erinevad praktilised heuristikad mastiviskamise tuvastamiseks, testiti kolme kasutaja põhjal ning rakendati kuu aja andmestikul.

Kolme kasutaja tulemused näitasid, et peaaegu kõik heuristikad tuvastasid mastiviskamist, mitte kasutaja enda liikumist võrreldes nende päris asukohaga antud ajahetkel. Üks heuristika, mis on seotud viibimisperioodidega, püüab kinni aga liiga palju tegelikke liikumisi ning tuleks järgnevateks töödeks üle vaadata. Antud kasutajate mastiviskamise osakaal oli 1.7% kogu nende kõnetoimingute arvust.

Kuu aja terve Eesti ühe operaatori kasutajate peal leitud tulemus näitas, et mastiviskamised moodustavad ainult 1% kogu andmestikust. Töö aluseks olevad uurimused Wu et al. (2014) ja Qi et al. (2016) leidsin selleks osakaaluks olevat vastavalt 6% ja 13-15% kogu nende andmestikust. Üks võimalike seletusi sellele tulemusele on see, et Eesti andmestik ei ole nii urbaniseerunud piirkondade kohta nagu oli neil. Linnalistes asulates on mobiilimastid üksteisele lähemal ning levialad kattuvad rohkem kui näiteks maalistes piirkondades (Sauter, 2010). Lähestikku paiknemine toob tihti kaasa mastivahetamisi, sest erinevatel asjaoludel, näiteks halb ilmastik või ehitised ja muud objektid blokeerivad signaali, võib signaali tugevus telefoni ja masti vahel muutuda. Sellised mastide signaalitugevuste vaheldumised toovadki kaasa seadme ja mitme erineva masti ühendused lühikeste ajaakende jooksul ning põhjustavad andmestikus näilisi liikumisi. Arvestades, et antud magistritöös on märkimisväärne osa võrgutoiminguid väljaspool Tallinnat ja selle lähiümbrust, siis võib olla see üks oletus madalale mastiviskamise osakaalule, sest uuringuala on hõredama asustusega ning seega on ka mastid hõredamalt.

Mastiviskamise ruumiliseks vaatluseks leiti igale mobiilimastile tema mastiviskamise protsent kogu kõnetoimingute arvust antud mastis. Paremaks visualiseerimiseks agregeeriti tulemused Eesti asustusüksuse tasemele (Figure 14). Ruumiliselt tulevad selgelt esile linnalised asulad, eriti Tallinn ja selle lähiümbrus, mis samuti viib järelduseni, et mastiviskamine on eelkõige tihedama asustusega alade nähtus.

Ajalise dimensiooni hindamiseks leiti mastiviskamise jaotumine nädalapäevade lõikes ning ööpäevane muutumine. Reedeti esineb mastiviskamist kõige rohkem (20,9%) ja kõige vähem pühapäeviti (10,4%). Ööpäevases rütmis tuli selgelt esile pidev tõus alatest hommikust kuni pärastlõunani, mis päädis õhtuse tipptunni algusega kella 17:00 paiku. Üks võimalike seletusi on inimeste suurem mobiilsus antud kellaaegadel või päevadel (Järv et al., 2007; Järv et al., 2012), mis toob kaasa rohkem mastivahetusi ühest levialast teise liikudes. Samas võib liikudes olla signaal kohati blokeeritud lähimast mastist ning ühendatakse mõnda teise, kaugemale masti. Signaali taastumisel aga tagasi lähimasse. Sedasi kasutaja asukohta pendeldades nagu mobiilpositsioneerimise näitab, päris elus ei toimu ning on tuvastatav mastiviskamise kaudu. Lisaks järgib teatud määral ajaline rütm kõnetoimingute koguarvu. Sealt tulenevalt võib oletada, et kui kasvab kõnetoimingute arv, siis on mobiilivõrgus rohkem mastivahetamisi seoses võrgukoormuse tasakaalustamisega. Tasakaalustamise tulemusena võib esineda logides näiliselt liikumisi, mida seade ilmselt ei teinud ning neid on võimalik tuvastada kui mastiviskamisi.

Töö käigus loodud meetod on eelkõige tähtis passiivsete mobiilpositsioneerimis andmete kvaliteeti tõstmiseks. Paljud autorid on välja toonud oma töödes, et mastiviskamise probleem esineb sellistes andmestikes ning võib mõjutada analüüsi tulemusi (Ahas et al., 2010b; Bayir et al., 2010; Chen et al., 2016; Iovan et al., 2013; Wang & Chen, 2018). Ka hindab kirjandus seda osakaalu üpris kõrgeks 6%-st kuni 16%-ni, mis on märkimisväärne osa koguandmestikust. Selles magistritöös ei leidnud sellised numbrid kinnitust, mis tõstatab vajaduse mastiviskamist veel lähemalt uurida, laiendades katsegrupi suurust.

Kokkuvõtvalt saab öelda, et käesoleva magistritöö eesmärgid said täidetud. Töötati välja praktiline meetod mastiviskamise tuvastamiseks, leiti mastiviskamiste osakaal Eesti kontekstis ning uuriti põgusalt mastiviskamise ajalisi ja ruumilisi aspekte. Pole kahtlustki, et mastiviskamist esineb passiivsetest mobiilpositsioneerimise andmetes, mida näitab ka töö analüüsi osa. Mastiviskamise tõttu on sellistes andmestikes ekslikult liikumisi, mis päris elus tegelikult ei toimunud. Mastiviskamise tuvastamise abil on võimalik kasutaja asukohta täpsemalt määratleda, seeläbi tõstes andmete kvaliteeti ja analüüse nende põhjal.

Acknowledgements

My biggest thanks go to Kaisa Vent, who suggested the topic and gave me feedback whenever asked for and whose support throughout the writing was always encouraging. The work would have not been possible without Kaire. She has been the most supportive and motivational during the writing of this thesis and my studies. I am grateful to Ferru, who with his fluffiness has always been able to cheer me up and keep me positive, which in turn helped me to recharge and stay on task.

I am thankful to my two supervisors, Anto Aasa and Erki Saluveer, who provided their own viewpoints, which made it possible to see things in different aspects.

I would also like to thank the company Positium, who provided me with the access to the data. A special thanks goes out to my co-workers at Positium – Egle, Jaak and Patrick, who had to, sometimes forcefully, think with me to solve some coding issues or listen to me complaining not being able to solve an issue.

Bibliography

Ahas, R., Laineste, J., Aasa, A., & Mark, Ü. (2007). The spatial accuracy of mobile positioning: some experiences with geographical studies in Estonia. In Location based services and telecartography (pp. 445-460). Springer, Berlin, Heidelberg.

Ahas, R., Aasa, A., Roose, A., Mark, Ü., & Silm, S. (2008). Evaluating passive mobile positioning data for tourism surveys: An Estonian case study. Tourism Management, 29(3), 469-486.

Ahas, R., Aasa, A., Silm, S., & Tiru, M. (2010a). Daily rhythms of suburban commuters' movements in the Tallinn metropolitan area: Case study with mobile positioning data. Transportation Research Part C: Emerging Technologies, 18(1), 45-54.

Ahas, R., Silm, S., Järv, O., Saluveer, E., & Tiru, M. (2010b). Using mobile positioning data to model locations meaningful to users of mobile phones. Journal of urban technology, 17(1), 3-27.

Ahas, R., Armoogum, J., Esko, S., Ilves, M., Karus, E., Madre, J. L., O., Potier, F., Schmucker, D., Sonntag, U & Tiru, M. (2014). Feasibility study on the use of mobile positioning data for tourism statistics-consolidated report. Tech. rep., Consortium.

Bayir, M. A., Demirbas, M., & Eagle, N. (2010). Mobility profiler: A framework for discovering mobility profiles of cell phone users. Pervasive and Mobile Computing, 6(4), 435-454.

Chen, C., Ma, J., Susilo, Y., Liu, Y., & Wang, M. (2016). The promises of big data and small data for travel behavior (aka human mobility) analysis. Transportation research part C: emerging technologies, 68, 285-299.

Corazza, G. E., Giancristofaro, D., & Santucci, F. (1994, June). Characterization of handover initialization in cellular mobile radio networks. In Vehicular Technology Conference, 1994 IEEE 44th (pp. 1869-1872). IEEE.

Furletti, B., Gabrielli, L., Giannotti, F., Milli, L., Nanni, M., Pedreschi, D., ... & Garofalo, G. (2014). Use of mobile phone data to estimate mobility flows. measuring urban population and inter-city mobility using big data in an integrated approach. In Proceedings of the 47th Meeting of the Italian Statistical Society.

Girardin, F., Vaccari, A., Gerber, A., Biderman, A., & Ratti, C. (2009). Towards estimating the presence of visitors from the aggregate mobile phone network activity they generate. In Intl. Conference on Computers in Urban Planning and Urban Management.

Gonzalez, M. C., Hidalgo, C. A., & Barabasi, A. L. (2008). Understanding individual human mobility patterns. nature, 453(7196), 779.

Gu, J., Bae, S. J., Chung, M. Y., Cheon, K. Y., & Park, A. S. (2010). Mobility-based handover decision mechanism to relieve ping-pong effect in cellular networks. In 2010 16th Asia-Pacific Conference on Communications (APCC) (pp. 487-491). IEEE.

GSMA Intelligence. (2017). Global mobile trends 2017. GSMA, September.

GSMA Intelligence. (2019). The mobile economy 2019. GSMA, February.

Horak, R. (2007). Telecommunications and data communications handbook. John Wiley & Sons.

Iovan, C., Olteanu-Raimond, A. M., Couronné, T., & Smoreda, Z. (2013). Moving and calling: Mobile phone data quality measurements and spatiotemporal uncertainty in human mobility studies. In Geographic information science at the heart of Europe (pp. 247-265). Springer, Cham.

Järv, O., Saluveer, E., & Ahas, R. (2007). Analysis of traffic distribution and frequencies in Kose-Võõbu and Võõbu-Mäo units of E263 Tallinn-Tartu-Luhamaa highway (E263 Tallinna-Tartu-Luhamaa maantee Kose-Võõbu ja Võõbu-Mäo lõikude liiklusuuring mobiilpositsioneerimise abil). Tartu: Positium LBS.

Järv, O., Ahas, R., Saluveer, E., Derudder, B., & Witlox, F. (2012). Mobile phones in a traffic flow: a geographical perspective to evening rush hour traffic analysis using call detail records. PloS one, 7(11), e49171.

Järv, O., Ahas, R., & Witlox, F. (2014). Understanding monthly variability in human activity spaces: A twelve-month study using mobile phone call detail records. Transportation Research Part C: Emerging Technologies, 38, 122-135.

Kung, K. S., Greco, K., Sobolevsky, S., & Ratti, C. (2014). Exploring universal patterns in human home-work commuting from mobile phone data. PloS one, 9(6), e96180.

Laasonen, K., Raento, M., & Toivonen, H. (2004, April). Adaptive on-device location recognition. In International Conference on Pervasive Computing (pp. 287-304). Springer, Berlin, Heidelberg.

Lee, J. K., & Hou, J. C. (2006). Modeling steady-state and transient behaviors of user mobility: formulation, analysis, and application. In Proceedings of the 7th ACM international symposium on Mobile ad hoc networking and computing (pp. 85-96). ACM.

Li, M., He, B., Guo, R., Li, Y., Chen, Y., & Fan, Y. (2018). Study on population distribution pattern at the county level of China. Sustainability, 10(10), 3598.

Miao, G., Zander, J., Sung, K. W., & Slimane, S. B. (2016). Fundamentals of mobile data networks. Cambridge University Press.

Raun, J., Ahas, R., & Tiru, M. (2016). Measuring tourism destinations using mobile tracking data. Tourism Management, 57, 202-212.

Ruktanonchai, N. W., Ruktanonchai, C. W., Floyd, J. R., & Tatem, A. J. (2018). Using Google Location History data to quantify fine-scale human mobility. International journal of health geographics, 17(1), 28.

Sauter, M. (2006). Communication systems for the mobile information society. John Wiley & Sons.

Sauter, M. (2010). From GSM to LTE: an introduction to mobile networks and mobile broadband. John Wiley & Sons.

Technical Regulatory Authority of Estonia. (2016). Annual report. https://www.tja.ee/sites/default/files/content-editors/TJA/Aastaraamat/tra_annual_report_2016_eng_web.pdf . Last reviewed 15.02.2019

Tiru, M. (2014, October). Overview of the sources and challenges of mobile positioning data for statistics. In Proceedings of the International Conference on Big Data for Official Statistics, Beijing, China (pp. 28-30).

Wang, F., & Chen, C. (2018). On data processing required to derive mobility patterns from passively-generated mobile phone data. Transportation Research Part C: Emerging Technologies, 87, 58-74.

Wu, W., Wang, Y., Gomes, J. B., Anh, D. T., Antonatos, S., Xue, M., ... & Decraene, J. (2014, July). Oscillation resolution for mobile phone cellular tower data to enable mobility modelling. In 2014 IEEE 15th International Conference on Mobile Data Management (Vol. 1, pp. 321-328). IEEE.

Qi, L., Qiao, Y., Abdesslem, F. B., Ma, Z., & Yang, J. (2016, June). Oscillation resolution for massive cell phone traffic data. In Proceedings of the First Workshop on Mobile Data (pp. 25-30). ACM.

Internet sources

Enge, E. (2018). Stone Temple Insights: Mobile vs Desktop Usage in 2018: Mobile takes the lead. https://www.stonetemple.com/mobile-vs-desktop-usage-study (Reviewed 05.04.2019)

ESRI. How Average Nearest Neighbor works. https://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/h-how-average-nearest-neighbor-distance-spatial-st.htm (Reviewed 21.04.19)

Eurostat. General Database: Science, technology, digital society. https://data.europa.eu/euodp/data/dataset/Ge5r8AKYkHXK70lT8PrJA (Reviewed 15.02.2019)

Google Maps Help. https://support.google.com/maps/answer/2839911?co=GENIE.Platform%3DAndroid&hl=en (Reviewed 23.05.2019)

Ofcom. Communications Market Report. https://www.ofcom.org.uk/research-and-data/multi-sector-research/cmr/cmr-2018/interactive (Reviewed 26.04.2019)

Positium. (2017). Tallinna ja Tallinnaga seotud liikumiste lähte- ja sihtkohtade korrespondentsmaatriks (ODM) mobiilpositsioneerimise andmetel. https://www.mnt.ee/sites/default/files/tallinna_ja_harjumaa_mobiilpos_aruanne_20180123.pdf (Reviewed 16.05.2019)

Statistics Estonia. (2017). Population number, area and density by administrative unit or type of settlement. (Reviewed 05.05.2019)

Wikipedia: Demographics of Paris. https://en.wikipedia.org/wiki/Demographics_of_Paris (Reviewed 05.05.2019)

Wikipedia: Shanghai. https://en.wikipedia.org/wiki/Shanghai (Reviewed 05.05.2019)

APPENDICES

Appendix 1: Nearest average neighbour for cellular towers by group

To determine distance thresholds for various heuristics, average nearest neighbour was calculated for urban, rural and total group. The assumption is that rural cellular towers are less densely situated compared to towers in urban areas and different distance thresholds should be looked at for each group. A nearest neighbour ratio was also found, which shows if the cellular towers are clustered, randomly or dispersedly situated (Table 4.) If the value is less than 1, then the spatial pattern indicates clustering. If the value is more than 1, then the spatial pattern indicates dispersion of the cellular towers (ESRI).

Table 4. Average mean distance in meters to nearest neighbour and spatial distribution pattern based on group.

Cellular tower group	Average mean distance to nearest neighbour (m)	Nearest neighbour ratio	Spatial pattern
URBAN	930	0.695 *	Clustered
RURAL	6735	1.000**	Random
TOTAL	2824	0.204*	Clustered

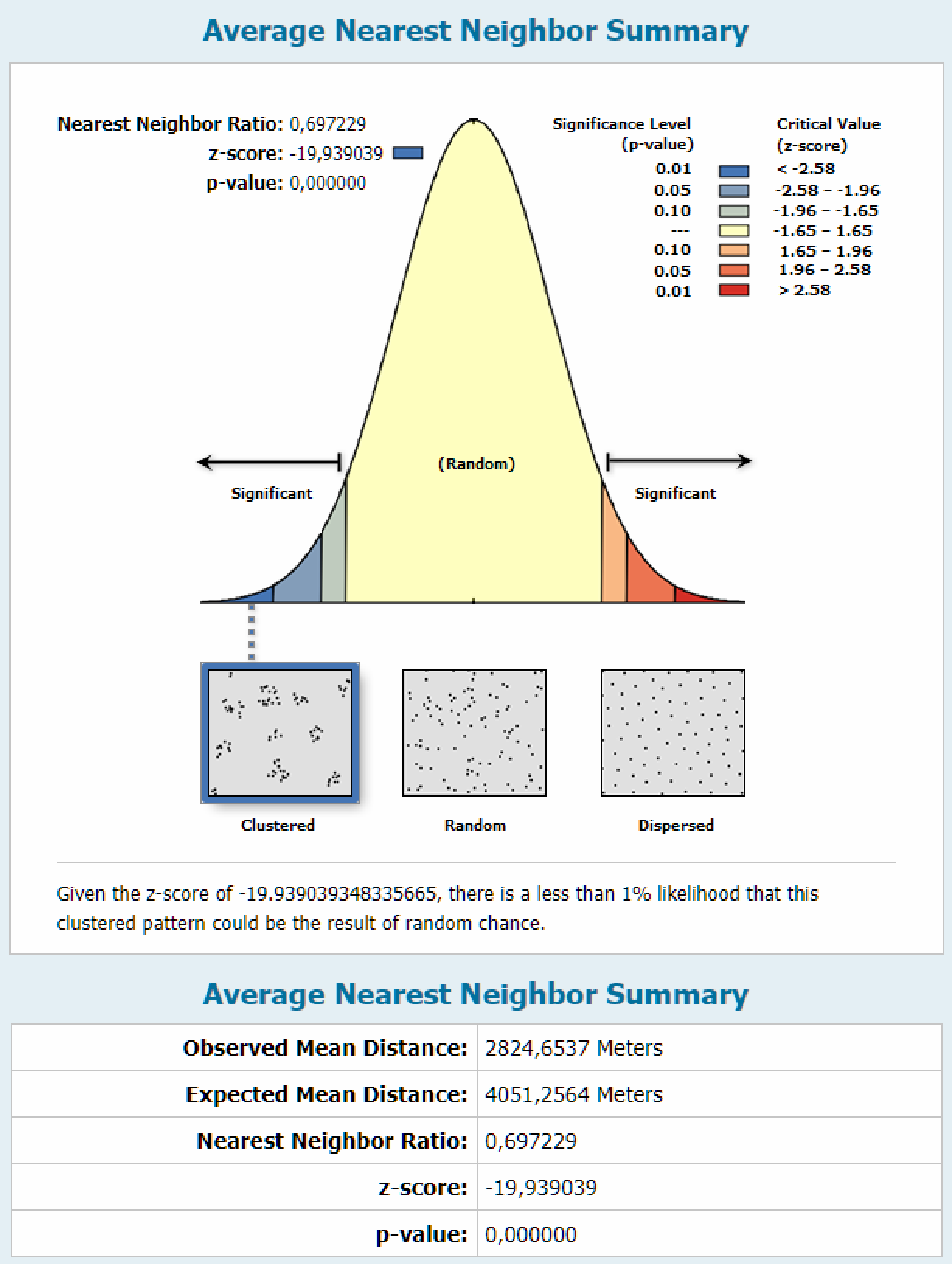
** p < 0.01*

*** p>0.05. No significant clustering or dispersion. Indicates randomness*

For all the cellular towers, the mean distance from one cellular tower to said tower's nearest neighbouring cellular tower is 2824 meters. The calculated nearest neighbour ratio 0.697 (*z > 2.58; p < 0.01*) indicates that the cellular towers are spatially clustered (Figure 17).

For cellular towers that are classified as urban, the mean distance from one cellular tower to said tower's nearest neighbouring cellular tower is 930 meters. The calculated nearest neighbour ratio 0.204 (*z > 2.58; p < 0.01*) strongly indicates that the cellular towers are spatially clustered as would be expected (Figure 18).

For cellular towers that are classified as rural, the average distance from one cellular tower to said tower's nearest neighbouring cellular tower is 6735 meters. The calculated nearest neighbour ratio 1.00 (*z = 0.320; p > 0.05*) indicates that the cellular towers are not significantly different from random and therefor no clustering or equally dispersed spatial pattern can not be determined (Figure 19).



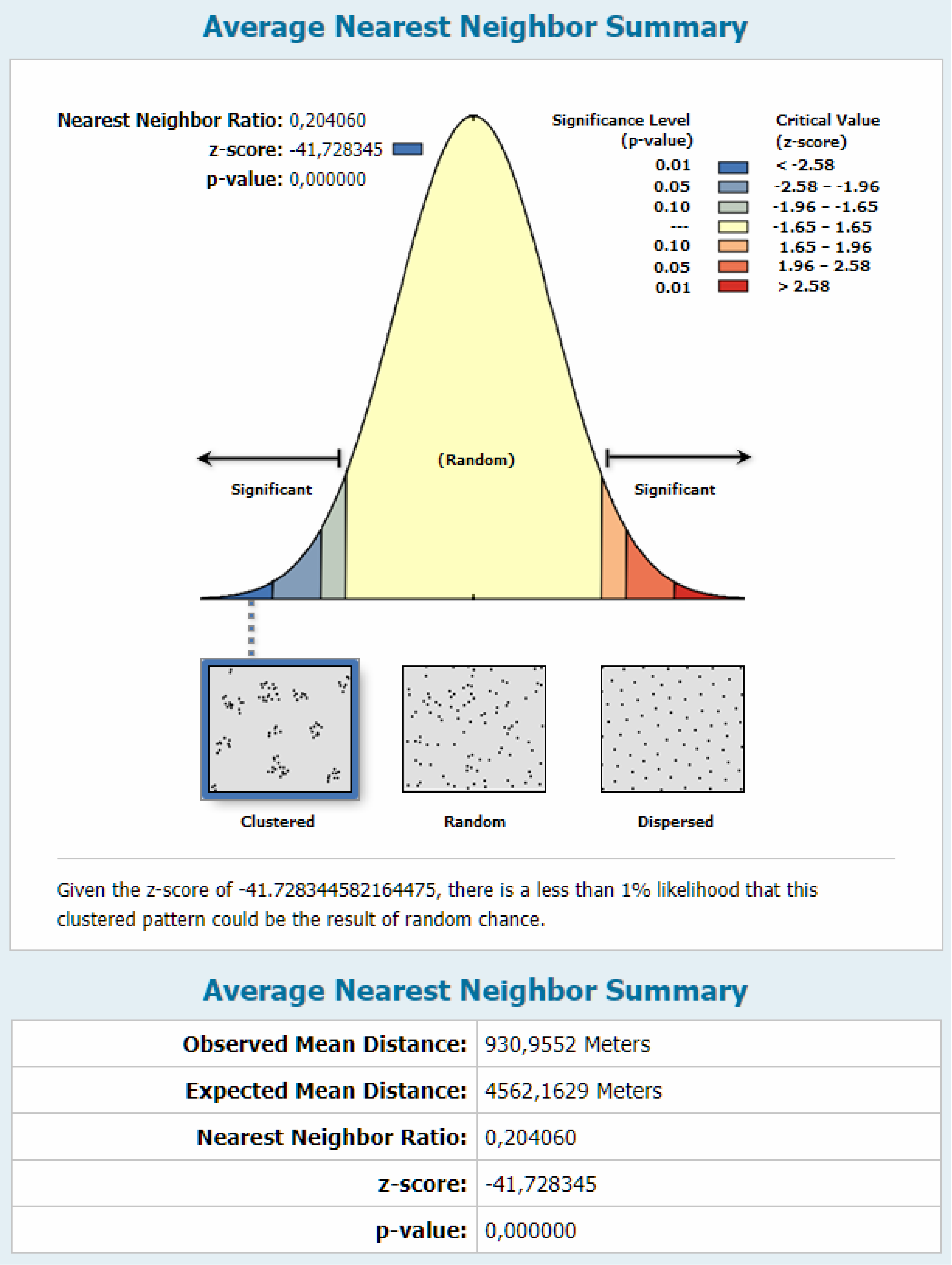
Average Nearest Neighbor Summary

Given the z-score of -19.939039348335665, there is a less than 1% likelihood that this clustered pattern could be the result of random chance.

Average Nearest Neighbor Summary

Observed Mean Distance:	2824,6537 Meters
Expected Mean Distance:	4051,2564 Meters
Nearest Neighbor Ratio:	0,697229
z-score:	-19,939039
p-value:	0,000000

Figure 17. Average nearest neighbour summary for all the cellular towers.



Average Nearest Neighbor Summary

Nearest Neighbor Ratio: 0,204060

z-score: -41,728345 🟦

p-value: 0,000000

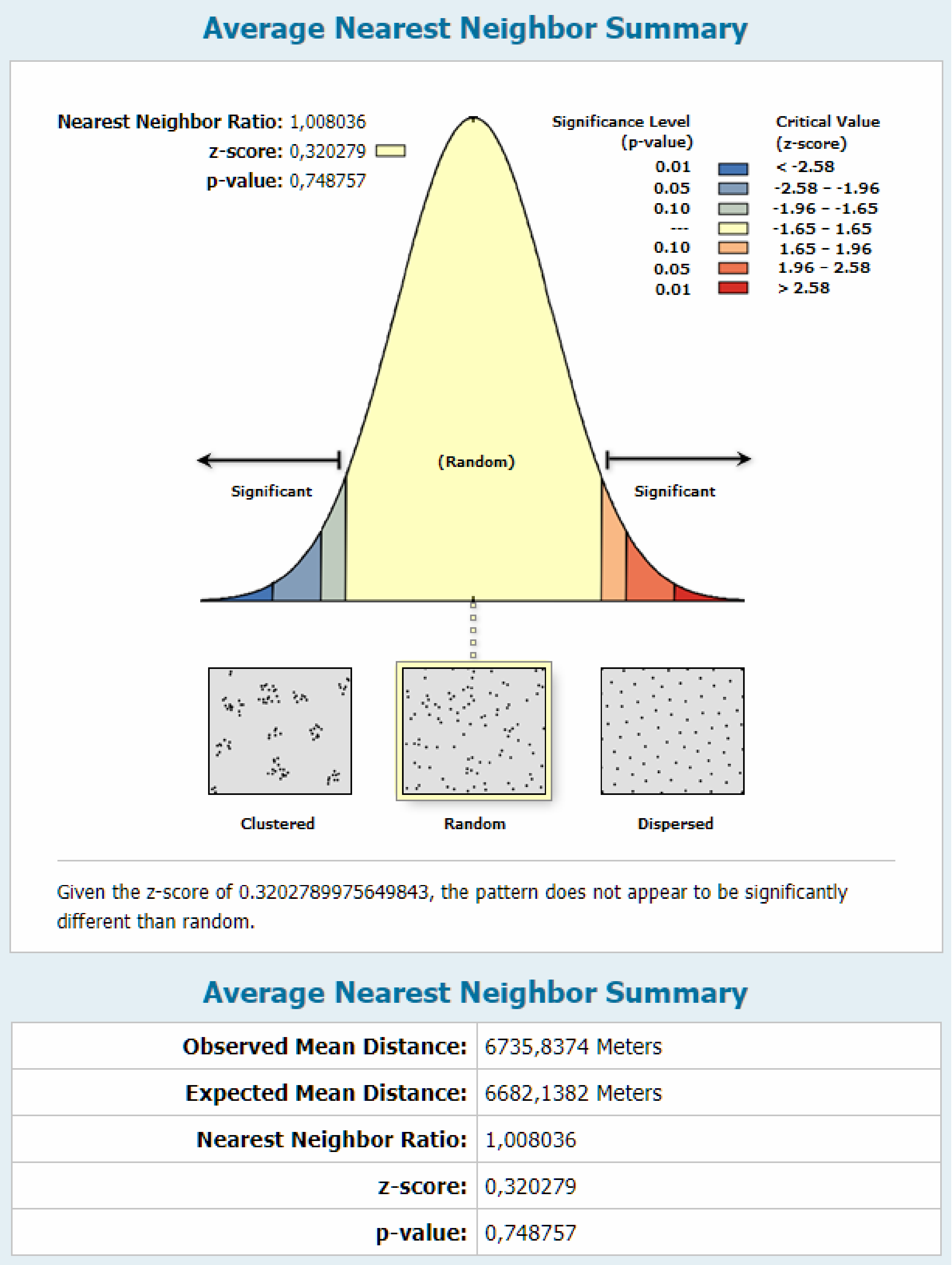
Significance Level (p-value)		Critical Value (z-score)
0.01	🟦	< -2.58
0.05	🟦	-2.58 - -1.96
0.10	▨	-1.96 - -1.65
---	🟨	-1.65 - 1.65
0.10	🟧	1.65 - 1.96
0.05	🟧	1.96 - 2.58
0.01	🟥	> 2.58

Given the z-score of -41.728344582164475, there is a less than 1% likelihood that this clustered pattern could be the result of random chance.

Average Nearest Neighbor Summary

Observed Mean Distance:	930,9552 Meters
Expected Mean Distance:	4562,1629 Meters
Nearest Neighbor Ratio:	0,204060
z-score:	-41,728345
p-value:	0,000000

Figure 18. Average nearest neighbour summary for urban cellular towers.



Average Nearest Neighbor Summary

Nearest Neighbor Ratio: 1,008036

z-score: 0,320279 🟨

p-value: 0,748757

Significance Level (p-value)		Critical Value (z-score)
0.01	🟦	< -2.58
0.05	🟦	-2.58 - -1.96
0.10	⬜	-1.96 - -1.65
---	🟨	-1.65 - 1.65
0.10	🟧	1.65 - 1.96
0.05	🟧	1.96 - 2.58
0.01	🟥	> 2.58

Given the z-score of 0.3202789975649843, the pattern does not appear to be significantly different than random.

Average Nearest Neighbor Summary

Observed Mean Distance:	6735,8374 Meters
Expected Mean Distance:	6682,1382 Meters
Nearest Neighbor Ratio:	1,008036
z-score:	0,320279
p-value:	0,748757

Figure 19. Average nearest neighbour summary for rural cellular towers.

Sander Pukk

24/05/2019