UNIVERSITY OF TARTU
Institute of Computer Science
Computer Science Curriculum

Hendrik Hiir

The impact of external factors on Estonian mobile call activity

Bachelor's Thesis (9 ECTS)

Supervisors: Rajesh Sharma, PhD
Anto Aasa, PhD

Tartu 2019

The impact of external factors on Estonian mobile call activity

Abstract:

The aim of this thesis is to show how different external factors affect mobile call activity. Based on 628 716 Call Data Records from Estonia, a social network is formed and analysed. Call connections between Estonian counties are observed, with emphasis on population and popularity differences and finding the most and least in-area centered counties. Call activity over time is analysed and the reasons behind the differences in activity are discussed. In addition, the impact of different events on call activity is studied. From natural events, this work focuses on weather, full moon and solar eclipse. From non-natural events, the impact of parliamentary elections, a major football match and infamous Friday the 13^{th} on call activity is observed. Based on the results of this thesis, it can be indicated that human calling activity depends on the time period of calling and it also gets impacted by events happening around it.

Keywords: CDR, mobile interactions, network science, social network analysis, call activity

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

Väliste tegurite mõju Eesti mobiilkõnede aktiivsusele

Lühikokkuvõte:

Käesoleva töö eesmärk on näidata, kuidas erinevad välised tegurid mõjutavad mobiilkõnede aktiivsust. Kasutati 628 716 rida kõneandmete kirjeid, et koostada Eestis tehtud kõnede põhjal sotsiaalvõrgustik ja selle põhjal analüüs. Vaadeldi, kuidas erinevad kõneühendused maakondade kaupa, uurides ka maakondade rahvaarvu ja populaarsuse vahesid. Näidati, millistes maakondades on kõige rohkem ja millistes kõige vähem maakonnasiseseid kõnesid. Lisaks uuriti, kuidas erinevad looduslikud ja mitteloodusliku sündmused kõneaktiivsust mõjutavad. Looduslikest sündmustest uuriti ilma, täiskuud ja päikesevarjutust. Mitteloodusliku sündmustest uuriti parlamendivalimisi, vaatlusperioodi suurimat jalgpallimatši ja 13. kuupäeva mõju reedele. Töö tulemuste põhjal on inimeste kõneaktiivsus märgatavalt mõjutatud ajast ning ümbritsevatest sündmustest.

Võtmesõnad: CDR, mobiilsidevahelised suhted, võrguteooria, sotsiaalvõrgustiku analüüs, kõnede aktiivsus

CERCS: P170 Arvutiteadus, arvutusmeetodid, süsteemid, juhtimine

Acknowledgements

I am very grateful to Positium and an anonymous data provider for providing the dataset that was used for this work.

Contents

1 Introduction

Call Data Records (CDR) are information collected by mobile network operators during the course of service in order to assemble billing data. CDR usually contain of timestamp, calling party and called party in pseudonymised form, call type, duration and coordinates of corresponding parties for localisation [1]. With an estimated of 7.683 billion mobile cellular subscriptions worldwide [2] and more than 1.8 million active SIM cards in Estonia as of 2017 [3], CDR provide valuable metadata for behavior analysis through network science. This implies that there are much data that can be combined with other datasets to solve problems, including the problem of this thesis: observing the impact of external factors on people's activity.

CDR are often combined with other datasets for research purposes. For example, one study conducted in Norway in 2018 [4] describes how mobile phone data can be used in combination with railway infrastructure and train traffic data to analyse the number of train travellers. Using mobile phone data can be an alternative way to gather information about the statistics of passengers. This can be applied to many situations to discover patterns in social behavior as discussed in this thesis.

One of the biggest challenges when working with big data such as CDR is finding a balance between privacy and utility. The more detailed and identifiable the used dataset is, the more opportunities there is to analyse subjects' behavior on specific characteristics level. However, even accumulated data can pose a privacy risk due to the fact that groups are still identifiable based on locations, provided that location coordinates are not pseudonymised [1].

Pseudonymised data is data where most of the personal attributes are replaced with reference numbers. Pseudonymisation is frequently practiced method in data science since it protects subjects' privacy, as the data cannot be attributed to a specific data subject. Although it is possible to retrieve personal information from an assigned reference number with additional information, technical and organisational measures must be in place in order to ensure that this additional information is not accessible to anybody but the data provider [5]. Pseudonymised data is still personal data in contrast to anonymised data, which cannot be used to identify individuals. Therefore, data protection regulations apply to them [6]. Personal data is any information or combination of factors that are related to an identifiable person, for instance, name, genetic information or location data [5]. In the context of this thesis, names in the used dataset are pseudonymised to project privacy of subjects.

In addition to having to find a balance between privacy and utility, there are a few other limitations while working with CDR. The location from users is tracked only when the user carries and communicates through the phone, which does not guarantee all movement patterns. Moreover, when analysing data from a social network perspective, a few assumptions are usually made which are probable but not definite for all users. The most frequent assumption is home and work locations based on localization data, which cannot be validated without harming the privacy of the subjects [7].

The aim of this thesis is to analyse the dataset of real users' mobile interactions, which were recorded over a period of 31 days from March 1st, 2015 until March 31st, 2015 to study how different external factors affect call activity. CDR are combined with other data such as weather data or Estonian 2015 parliamentary elections statistics in order to find results. In addition, a network of the callers is created and activity over Estonian counties and over time is analysed. The used dataset contains communication data of 722 724 records and is provided by one of the largest mobile operators in Estonia.

It is to be noted that the author submitted an article based on this thesis work to 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining with the title "Impact of Natural, Sociocultural, Political Events on Mobile Call Data Records – A Case Study".

The structure of the thesis is following. The second chapter gives an overview of related work by describing the usage of CDR in different research fields. The third and fourth chapters describe the used methodology and dataset, respectively. This is followed by a descriptive analysis that is divided into network properties, call activity over Estonian counties and call activity over time. Next, the effects of external events on call activity are analysed. These events are divided into natural events and non-natural events. This is finished with a conclusion and future perspectives of this work.

2 Related work

This chapter gives an overview of some of the works that have explored mobile phone data to analyse mobile interactions on a longitudinal basis. The aim is to give an overview of different fields of application where a similar approach has been used with datasets used and results discovered while covering different countries. CDR used in this chapter is an abbreviation for Call Data Records, which contain data such as calling parties, positioning timestamp and co-ordinates.

2.1 Human Behavior Analysis using CDR

Since the early 2000s with the massively increasing popularity of mobile phone usage, which nowadays plays important role in people's actions and life, data, social and geographic scientists have seen great potential in analysing the accompanying data. Already in 2003, the company Positium carried out a Social Positioning Method (SPM) survey [8] in Estonia with 30 people in order to find movement patterns in central Tallinn with data that was recorded over a course of one week. The movement patterns were differentiated in three age groups: up to 35 years old, 36-60 years old and older than 60 years old. Although no specific analysis based on age groups was made and only a few similar surveys were completed by that time, it was noted that social positioning would be life-changing in the future by giving us an overview of crisis situations and helping to prevent problems related to the movement of people. It is noteworthy that the analysts knew the examined people and their characteristics.

One year later, in 2004, as described in paper [9], a larger case study was held in Milan, Italy to discuss the possibility of more widespread usage of Location-Based Services (LBS) in the context of urban research. Even though the mobile communications market was booming, cell phone data was scarcely used in urban analysis back then. In that study, Global Positioning System (GPS) was used to gather location data in aggregated form. The chosen timeframe was sixteen days from April 19th to May 4th, which seemed to be enough for finding patterns and including some anomalies such as Liberation Day on April 25th and Milan soccer team's victory in Italian 2004 League Championship on May 2nd. Both of these events saw more than 50 000 people gathering in specific sites. In addition to analysing cell phone activity by time and density, three-dimensional graphs were made to show the geographical distribution of base-stations by time. The assumption was that cell phone station activity was related to the

number of people in the neighborhood. The study led to the conclusion that location data analysis was an underappreciated method in urban research as it gives an opportunity to monitor changing urban dynamics and has potential to gather information about the city in many details.

In [10] is described how between 2004 and 2005 in Cambridge, United Kingdom, a total of 330 000 hours of behavioral data from 94 subjects was collected and compared with self-report rational data. Although the number of subjects was not large, it was stated that the used methods could be applied to hundreds of millions of cell phone users. The subjects were students and faculty members from a major research institution, who also reported the proximity of friendship with others and satisfaction with their workgroup. The data from mobile phones was gathered with the help of pre-installed software, which helped to send researcher the data and call logs. The first part of the analysis was analysing the relationship between self-report and actual behavior, where observed proximity between friends and non-friends was estimated by time and location and it was correlated with reported data. Multiple Regression Quadratic Assignment Procedure (MRQAP) was used to analyse social network data. The second part of the analysis was to investigate the friendship characteristics using MRQAP and factor analysis. Repeated proximity and communication on Saturday nights were taken as an indication of friendship. In the third part, individual satisfaction of was compared with proximity and communication, with a standard assumption that individuals are more satisfied if they have more friends. The study predicted 95% of reciprocated friendships and such a method can predict individual-level outcomes.

In fact, behavioral analysis is a common subject of interest when it comes to working with mobile phone data. A paper [7] describes a comprehensive Human Behavior Understanding (HBU) research, which was made in 2010 based on one million anonymous mobile phone users data in the state of Massachusetts in the United States. The data included around 130 million location coordinates, which were used to determine eating, shopping, entertainment and recreational activities. Only weekdays were considered because the weekend pattern was believed to be different due to the considerably smaller amount of working people. People were divided into four different groups depending on their work cell's profile. Based on the information, it was possible to create an activity distribution map for different timeslots and find daily activity patterns. It was discovered that there is strong a correlation in activity patterns among people sharing a common work area's profile. Furthermore, within the groups, the activity patterns

similarity is related to the distance between the members. The results of this analysis can be used in urban planning.

Travel behavior research is also a field where mobile phone data is used. One study about travel demand estimation as described in paper [11] was made in 2015. That study used data from cities from several continents to find road usage patterns and proposed a system, which estimated travel demand with Call Data Records (CDR) in conjunction with open-source geospatial data, records and surveys. The analysed cities were Boston, San Francisco, Lisbon, Porto and Rio de Janeiro – a variety of cities from different continents to show the flexibility of the system. Among the results was found that most of the cities have a very high correlation in Origin-Destination (OD) matrices with the exception of Lisbon, which had the smallest units of aggregation.

2.2 Criminal Investigations

In addition, there have been attempts to use CDR for criminal investigation. A paper [12] describes how CDR are a source of data used by investigation agencies to investigate the case and proposes a system where CDR are stored in a centralised database for analysis. Cellar numbers and their relationships were presented in graphical representation based on graph theory. An example was given where a robbery took place at 11th Cross, Malleswaram, Bangalore on 20th November 2015 at 10:30 PM and suspected criminals were shortlisted based on phone calls made near the crime location. In that particular case, the calls within 1-hour interval from the crime occurrence would be suspicious and easily tracked from the database using a query, which includes phone calls only from n of the nearest towers. As a solution for many results, the used CDR can be merged with CDR of old cases and associated criminals. As a result, it is possible to analyse whether the result has relations with associated criminals. The paper also demonstrates that graph visualisation makes it easier to solve such cases.

A paper [13] proposes another system, presented in 2018 in India, which is able to generate reports consisting of maximum call durations and most frequently called users, which can help officers in investigation. A Criminal Database is used to extract data, which contains CDR of criminals. Additionally, the proposed system will be able to display caller profiles with frequently changing International Mobile Equipment Identity (IMEI) numbers in order to detect suspicious activity. In conclusion, although the real-life usage of the system was theoretical at

the point of its presentation, it has a potential to save time of investigators by creating automatic statistics about suspicious activities.

2.3 CDR in Health

Likewise, CDR provide beneficial information for health research as they can track the movement of people and infectious disease transference. To give an illustration, a paper [14] mentions a research published in Canada in 2015, which found that it was possible to predict flu infection with 30% precision among evaluated 72 people within a period of 17 weeks. However, a study [15] published in 2019 in the United Kingdom revealed that there is a lack of public engagement when it comes to using CDR in healthcare, as only 3% of questioned participants were aware of their usage in health research. Before explaining the safeguards of using their anonymous CDR, only 62% of people agreed with its usage, which later increased to 80%. This led to the conclusion that the conditions of using CDR for research purposes should be clearer to the users. The study mentioned that the value of the gained information through CDR can be increased further in health research when combining them with other datasets such as traffic data and malaria surveys. This applies to all fields of applications where mobile phone data is used.

3 Methodology

This chapter gives an overview of used programs and packages and the reasoning behind their usage.

3.1 Used programs

RStudio [16], which is an integrated development environment for R programming language, was used to process the data in order to gather statistical results for the analysis part. Also, network properties were found and plots were made in RStudio. The reason behind the usage of RStudio and R language is that they offer powerful functions for data processing, data combining and statistical measures with a variety of visualisation opportunities. Gephi [17] was used to visualise the undirected and unweighted graph of callers network due to its powerful abilities to process a large number of graph properties. The network graph is displayed in chapter 5.1 (Network properties).

3.2 Used libraries

Most of the used functions were from R Base Package [18]. Igraph [19] library was used to create a graph object from the edges containing of used Call Data Records' callers and call receivers. Igraph library provided functions that can read and create a simplified graph from a .csv file that contain graph edge attributes (nodes). These functions are graph.data.frame() [20] and simplify() [21]. In addition, igraph provided methods to calculate network properties. In addition, xlsx [22] library was used to create sub-files of the dataset that contained only necessary information for connections maps and elections activity maps. Dplyr [23] library was used to filter and arrange the data. Libraries sf [24] and tidyverse [25] were used to combine the shapefile of Estonian counties with data, find centroids for connection maps and create call connection figures which described call activity over the Estonian counties and during the parliamentary elections.

4 Dataset description

This chapter gives an overview of data size, data attributes and data cleaning.

4.1 Data description

The data used in the analysis part of this thesis was provided by one of the biggest mobile operators in Estonia and includes call records of 10% of caller IDs that made calls in Estonia during the period of March 1st, 2015 to March 31st, 2015. In total, the data included records of 722 724 calls. The dataset has 9 attributes:

1. **pos_usr_id_from** – ID of user who made a call;
2. **pos_usr_id_to** – ID of user who was called to;
3. **pos_time** – timestamp when the positioning was made, describes call start time;
4. **X_from** – X-coordinate of caller location;
5. **Y_from** – Y-coordinate of caller location;
6. **X_to** – X-coordinate of user who was called to;
7. **Y_to** – Y-coordinate of user who was called to;
8. **MKOOD_from** – code that describes the county where the call started;
9. **MKOOD_to** – code that describes the county where the call was received.

Pos_usr_id_from attribute is used to determine the caller in pseudonymised form. For the given period, each caller keeps the same ID, which can be used to determine the network of the caller. **Pos_usr_id_to** is used to find the person who was called to. **Pos_time** includes the timestamp containing of date (year, month and day) and time (hour, minute and second). This attribute can be used to group call records by days or narrower time slots. **X_from** and **Y_from** describe the caller location coordinates while **X_to** and **Y_to** describe the call receiver coordinates. **MKood_from** and **MKood_to** describe the county names where the call was made and where it was received. These codes are derived from positioning coordinates.

4.2 Data cleaning

In the described dataset, one ID had a significantly higher amount of received calls than any other ID. It received 94 008 calls (13.01% of total calls made), with no outgoing calls. As it was suspected to be a service center, it was removed for further analysis. The second most

frequent calling point received just 584 calls and was included. After data cleaning, 628 716 Call Data Records were left, which were used for analysis in chapters 5 (Descriptive analysis) and 6 (Effect of events on call activity).

5 Descriptive analysis

This chapter gives an overview of Estonia's call network and describes call activity over counties of Estonia. In addition, call activity over time (days and hours) is analysed.

5.1 Network properties

To understand the nature of Estonia's call network, an undirected and unweighted network graph $G = (N,\ E)$ was created, where N is the number of callers and E is the number of unique interactions. In the case of this network, the number of callers is 130 093 and the number of interactions is 137 809. For simplicity, all multiple edges and loops were removed. Table 1 describes the properties of the created graph that represents the network of calls made in Estonia. The average degree of 2.12 (Row 3) means that on average, each person is connected with just two persons. Although the number of disconnected components is 10 176 (Row 4), which indicates that the network is very disconnected, 73.16% of callers are present in that component (Row 5). Edge density of 1.63e-05 (Row 6) describes the ratio of network edges (interactions) and possible interactions, meaning that there are 61350 times fewer undirected interactions than possible within the network. The diameter of the network is 39 (Row 8), which is the longest path of all network shortest paths, with an average path length of 13.37 (Row 9). This means that the network is highly spread out.

Table 1. Properties of used dataset

Row	Attribute	Notation	Value
1	Nodes (Callers)	N	130 093
2	Edges (Interactions)	E	137 809
3	Average Degree		2.12
4	Disconnected Components		10 176
5	Nodes in the Biggest Component		95 182
6	Edge Density		1.63e-05
7	Clustering Coefficient		0.036
8	Diameter		39
9	Average Path Length		13.37
10	Communities		10 334 (0.94)

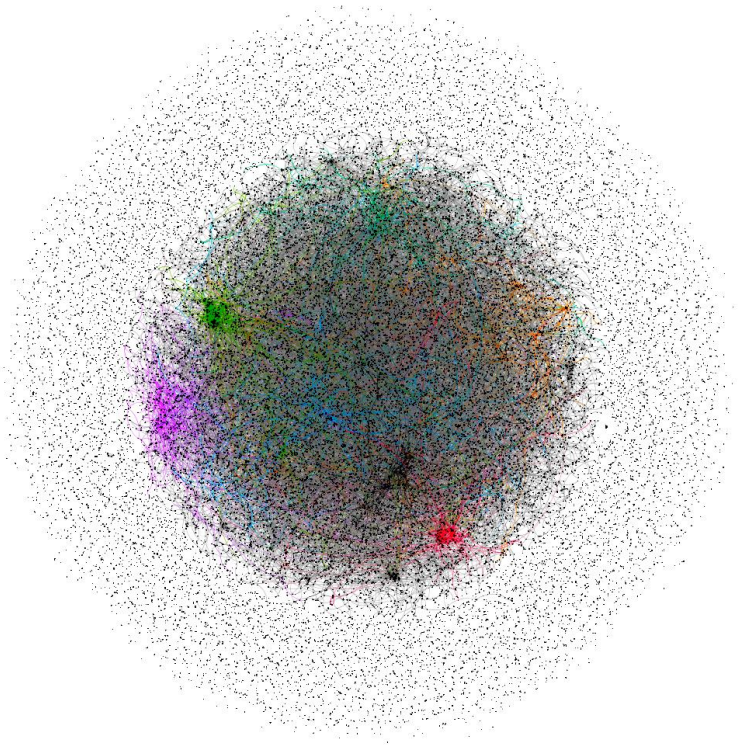

Figure 1. Visual representation of Estonia's calls network as an undirected and unweighted graph. Each community is displayed in different tone

Figure 1 visualises Estonia's calls network with its 10 334 communities (Row 10) as an undirected and unweighted graph. In the case of this network, only 5 out of 10 334 communities involve more than 1% of all callers.

5.2 Call activity description over Estonian counties

In Estonia, there are 15 counties, with Harju county, which includes capital Tallinn being the most populous and Hiiu county being the least populous. The following subchapter analyses connection between the counties.

Table 2. Abbreviations and populations of the counties

Abbreviation	**County**	**Population** [26]
Ha	Harju	575 601
Ta	Tartu	151 377
Id-V	Ida-Viru	147 597
Pä	Pärnu	82 349
Lä-V	Lääne-Viru	59 039
Vi	Viljandi	47 010
Ra	Rapla	34 436
Võ	Võru	33 172
Sa	Saare	31 706
Jõ	Jõgeva	30 841
Jä	Järva	30 109
Va	Valga	29 944
Põ	Põlva	27 438
Lä	Lääne	24 070
Hi	Hiiu	8 582

Table 2 gives an overview of abbreviations for each county used on network maps and their populations as of January 1st, 2015 in descending order.

5.2.1 Bidirectional call connections

Figure 2 shows bidirectional connections between each of the Estonian counties. In-area calls are the calls where both nodes are in the same county.

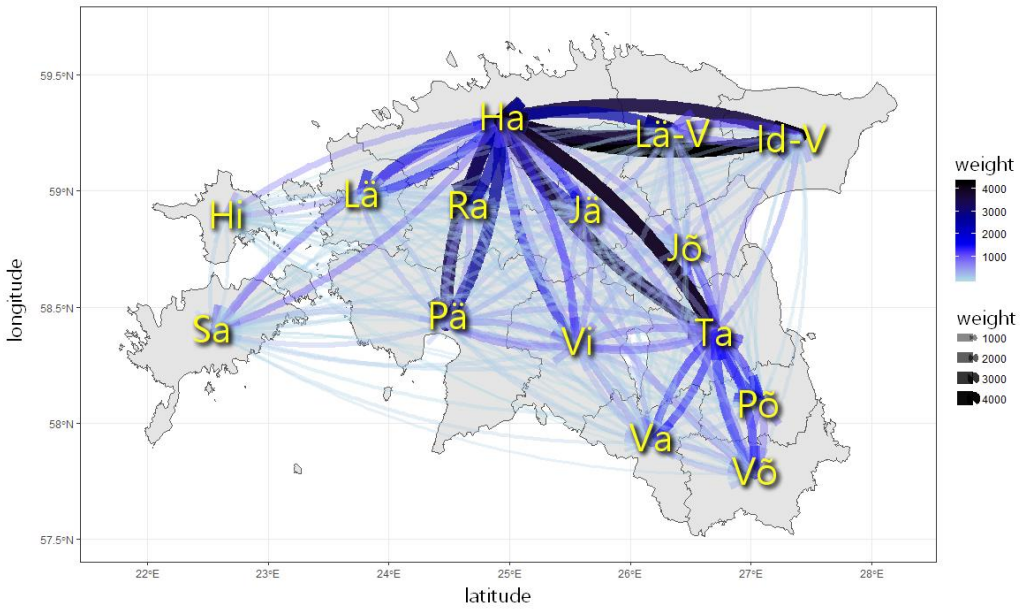

Figure 2. Bidirectional call connections between Estonian counties

In 10 out of 14 possible cases (as in-area calls were excluded), Harju county is the most popular destination for other counties outside the county's borders. For Harju county itself, the most popular destination is Ida-Viru county. It is notable that both counties have the biggest percentage of Russian community in Estonia, which is the likely reason for this bidirectional relation between these two areas. In 4 out of 14 possible cases, the second most populous Tartu county is the most popular destination outside the area's border. This is the case for the counties of Jõgeva, Valga, Põlva and Võru. All of them are either neighboring areas of Tartu (Jõgeva, Valga and Põlva) or having Tartu as the significant nearby center and working area (Võru).

Hiiu county is the least popular destination for 7 counties (all of them being in the eastern part of Estonia) while Põlva county is the least popular destination for 4 counties (all of them being in the western part of Estonia). Another county that was the least popular in more than one case is Saare county, which received the least amount of calls from the counties of Jõgeva and Põlva. The islands of Hiiumaa and Saaremaa (Hiiu county and Saare county) have unique least popular destinations: Võru county and Jõgeva county correspondingly. Therefore, the islands do not follow the trend of other western counties that have Põlva county as the least popular destination. Although the margins were not big, it is one of the ways to demonstrate how counties' communities differ.

5.2.2 Undirected call volume

Figure 3 shows the volume between each of the Estonian counties.

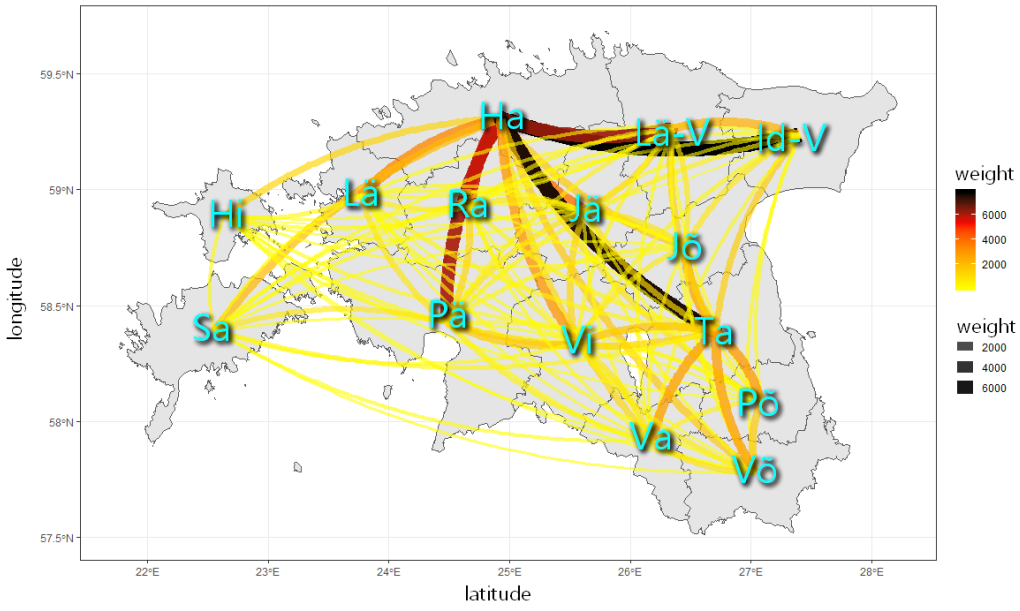

Figure 3. Undirected call volume between Estonian counties

The top 8 most frequent connections include Harju as one of the nodes. The most frequent connection is between Harju and Ida-Viru counties with 7807 calls (8.22% of all calls excluding in-area calls), likely because of the large Russian community in both of these counties. The second most connection frequent was between two most populous counties: Tartu and Harju with 7581 calls (7.99% of all calls excluding in-area calls). The least frequent connection with just 7 calls (0.007% of all calls excluding in-area calls) was between the least populous Hiiu county and Ida-Viru county. As the connections between the counties were bidirectionally similar like Figure 2 demonstrates, with some small exceptions such as more calls from Tartu county to Harju county than vice versa, despite Harju having 3.81 times bigger population, undirected call volume (demonstrated in Figure 3) could be a good measure to analyse network between the counties.

5.2.3 Population and popularity differences

Figure 4 describes the difference of population rank and the popularity rank in terms of incoming calls for each Estonian county. Table 3 shows the popularity rank and population rank differences in alphabetic order of the counties which are displayed in Figure 4. In addition, Table 3 shows the average rank of how popular destination the county is over all the counties.

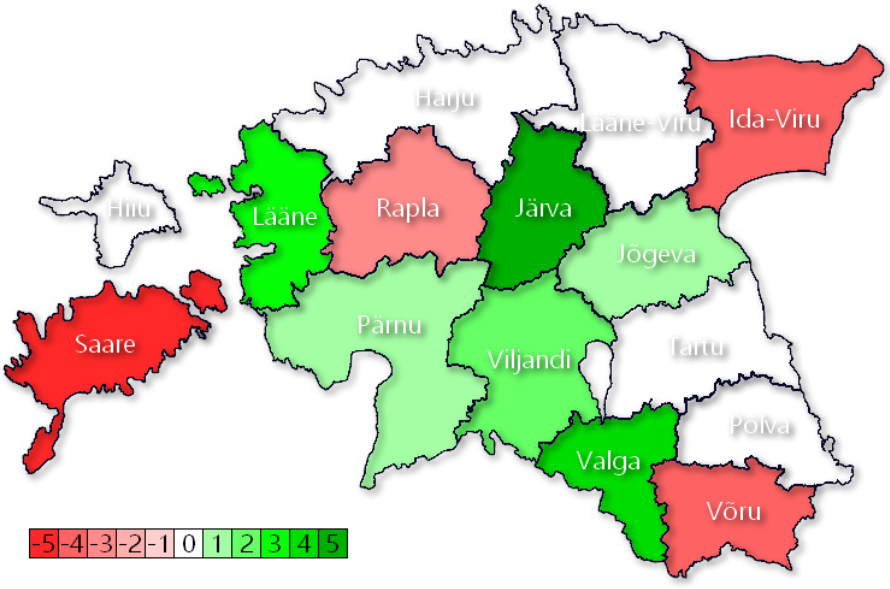

Figure 4. Difference between population rank and popularity rank of each county. The greener the color, the more calls it receives compared to its population. The more red the color, the less calls it receives compared to its population

Table 3. Population rank and popularity rank differences. Average rank shows average popularity as a call destination over all counties excluding in-area calls

County	Average rank	Popularity rank	Population rank	Difference
Harju	1.4286	1	1	0
Hiiu	12.1429	15	15	0
Ida-Viru	7.7143	7	3	-4
Jõgeva	8.2857	9	10	1
Järva	6.9286	6	11	5
Lääne	9.0714	11	14	3
Lääne-Viru	6.2143	5	5	0
Põlva	9.7143	13	13	0
Pärnu	5.1429	3	4	1
Rapla	9	10	7	-3
Saare	10.5	14	9	-5
Tartu	3.0714	2	2	0
Valga	8	8	12	4
Viljandi	5.7143	4	6	2
Võru	9.3571	12	8	-4

The difference of 0 means that the amount of incoming calls is corresponding to its population. Such areas are Harju county, Tartu county, Lääne-Viru county, Põlva county and Hiiu county. The counties with a negative difference are Saare county (-5), Ida-Viru county (-4), Võru county (-4) and Rapla county (-3). Ida-Viru county has a popularity rank of 7, which is less than Järva county's popularity rank of 6, despite of exceeding Järva's population rank by 8. Järva county has also the most positive difference (+5), which means that it receives the most calls compared its population. This is a case of a good average rank instead of an anomalously high amount of calls to the area. Other areas with a positive difference are Valga county (+4), Lääne county (+3), Viljandi county (+2) and Pärnu county (+1) and Jõgeva county (+1).

Ida-Viru county's big negative difference comes mostly from the fact that it receives most of its calls from neighboring areas (second most popular for Lääne-Viru county and fifth most popular for Jõgeva county, excluding in-area calls). Also, it is the most popular calling destination for Harju county, not counting in-area calls. Other areas do not have Ida-Viru in their top 5 most popular calling destinations. Saaremaa's big negative difference is the opposite case of Järva county's popularity: low popularity's average rank as only the least populated Hiiumaa has a lower average rank.

5.2.4 Most and least in-area centered counties

Table 4 shows the percentages of calls that are made and received in the respective county. Figure 5 describes how many calls of each Estonian county stay inside the county boundaries.

Table 4. Percentage of in-area calls for each Estonian county

Rank	County	In-area calls (%)		Rank	County	In-area calls (%)
1	Ida-Viru	90.7728		9	Valga	75.9028
2	Harju	90.1492		10	Lääne	73.1515
3	Pärnu	81.9078		11	Võru	72.9669
4	Lääne-Viru	81.8295		12	Järva	71.6937
5	Saare	81.7023		13	Rapla	69.2658
6	Tartu	80.4753		14	Põlva	69.0716
7	Hiiu	79.9483		15	Jõgeva	63.9830
8	Viljandi	77.0099				

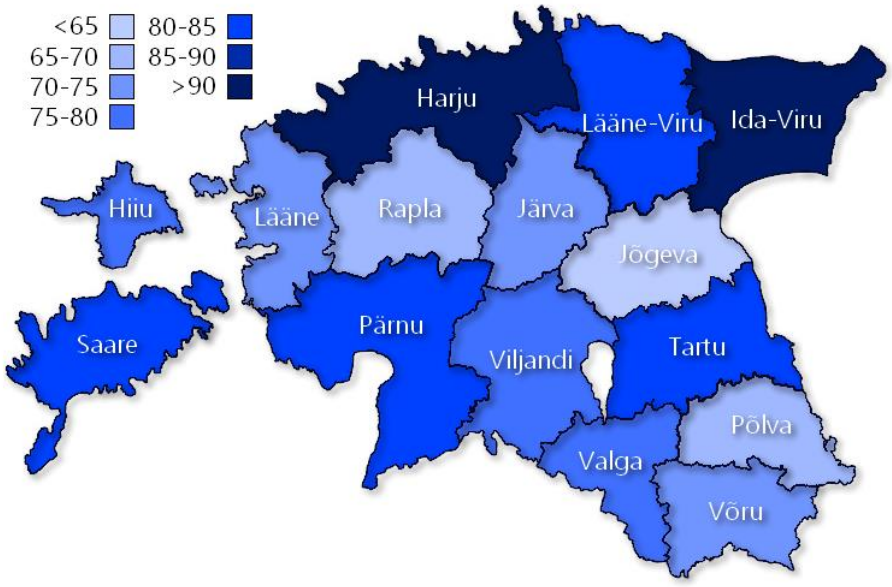

Figure 5. Percentage of calls that stay inside the county's area

The most in-area centered county is Ida-Viru with over 90.77% of calls staying inside the county boundaries. The most likely reason for this is that there is a large amount of Russian population who form a large network together. The second most in-area centric county is Harju with the indicator of 90.15%. The reason for such a high percentage of Harju county including

Tallinn is that it is the biggest center of Estonia and contains most of the points of interest such as service centers, institutions and working places which give fewer reasons to make calls outside the county. A relatively high percentage of in-area calls (>80%) is also in all other top 5 most populous counties of Tartu, Pärnu and Lääne-Viru for the similar reason with Harju county. The least in-area centered county is Jõgeva county with only 63.98% of calls made staying inside the county boundaries. The reason for this is that Jõgeva has frequent connections with all 5 neighboring areas, especially with Tartu county which is a destination of 12.32% of calls made from Jõgeva county. For the same reason, a high percentage of calls from Põlva county and Rapla county have a destination in other counties. Põlva county has frequent connection with neighboring Tartu county (13.85%) due to many services and work locations being there for people of Põlva county. Rapla county has the most frequent connection with any other area than itself out of all Estonian areas: with Harju county (17.73%).

5.3 Call activity over time

This subchapter describes call activity by day and over hours. In regard to hours, weekdays and weekend days are also analysed both in merged form and separately.

5.3.1 Call activity by day

Figure 6 describes the call activity by days over the period of March 1st, 2015 to March 31st, 2015.

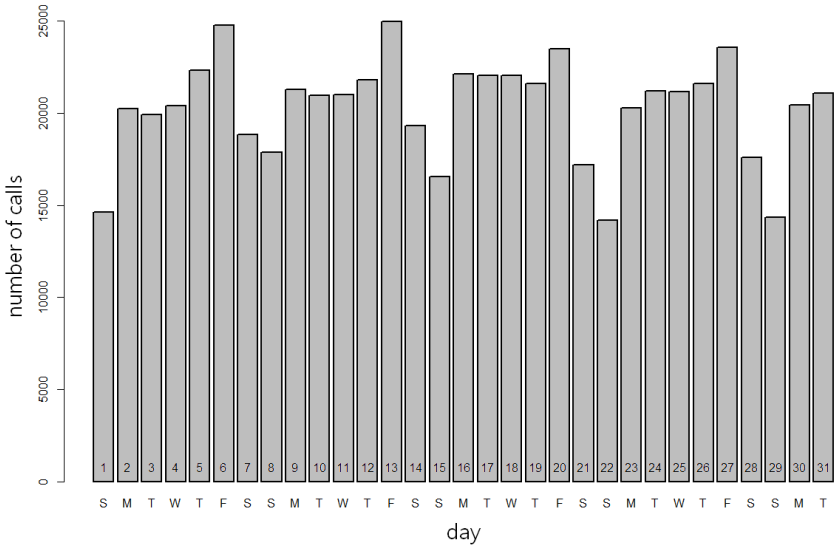

Figure 6. Call activity by day over the observable period

Friday is always the most active day of the week with an average of 24189.25 calls, being 19.27% above average activity. The reason for this is that for the majority of the people it is the end of the working period of Monday to Friday and therefore a favorable day to make appointments with people. From Monday to Thursday, there is generally a similar amount of calls made each day. Weekend days are always less active than weekdays due to the significantly smaller amount of work-related calls. Sunday was the least active day of the week with 15522.2 calls on average, which is 23.46% below average activity. The reason for this is that it has traditionally been a day of rest in most countries, including Estonia.

5.3.2 Call activity by hour

Figure 7 describes the average number of calls in a day by hour. For this figure, working days and weekend days are merged together in order to visualise average hourly activity over all days.

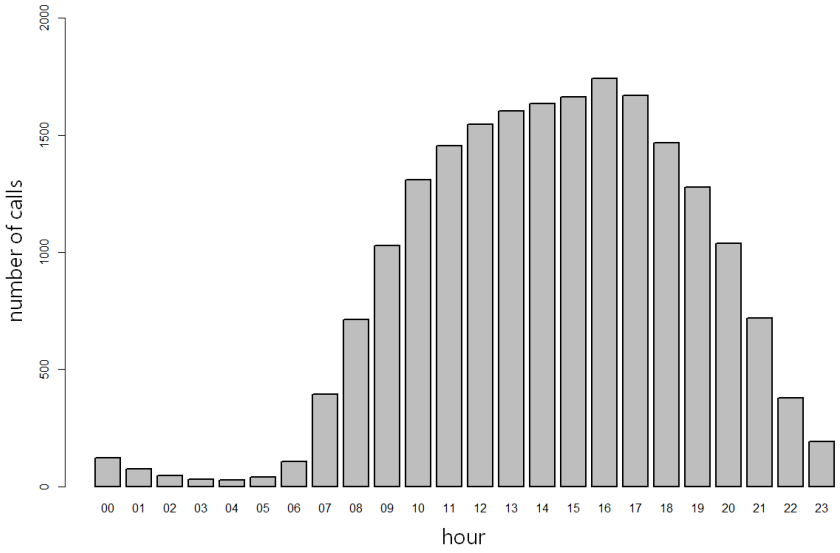

Figure 7. Call activity in a day by hour

Activity between the period of 04:00 and 16:59 is constantly increasing and activity between the period of 16:00 and 04:59 is constantly decreasing. The most active hour when taking all days into consideration is 16:00 to 16:59, which comes mostly from the period of Monday to Friday when work is commonly finished and people are available to call. The least active period is between 01:00 and 06:59 when on an average 53.94 calls were made in an hour compared to the daily hourly average of 845.05 calls per hour (15.67 times below average).

5.3.3 Call activity by hour during weekdays

Figure 8 describes the average number of calls in an hour when only weekdays (Monday to Friday) are considered.

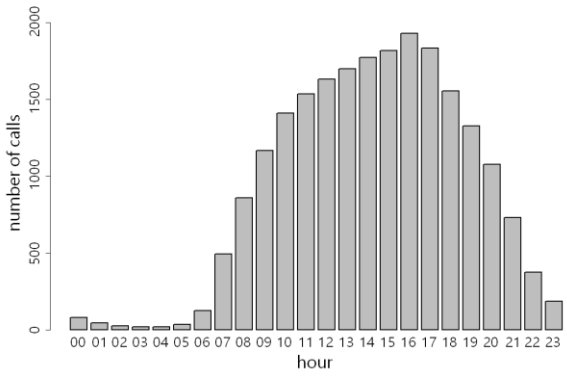

Figure 8. Call activity in a day by hour when only Weekdays are considered

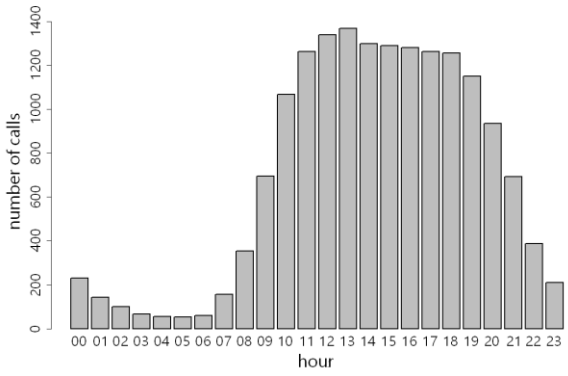

Figure 9. Call activity in a day by hour when only Weekend days are considered

Compared to Figure 7 that shows hourly activity over all days, the activity distribution in Figure 7 is similar due to 22 out of 31 observed days (70.97%) being weekdays. Weekdays are also more active than weekend days. The difference that was noticed is that the hour of 16:00 to 16:59 was even more active compared to other hours and the preceding period from 10:00 was more gradual. The activity increasing and decreasing periods were the same with 04:00 to 16:59 and 16:00 to 03:59 respectively. Another notable difference is fewer calls during the night period. While the merged data had 53.94 calls made on average during the period of 01:00 to 06:59, the average number of calls for the same period was 43.45 (21.53% difference), which is 0.20% of all calls during the day. The average number of calls in an hour was 905.69.

5.3.4 Call activity by hour during weekend days

Figure 9 describes the number of calls in an hour when only weekend days (Saturday and Sunday) are considered. The main differences compared to weekdays' hourly calls are more calls during the night, more even activity distribution between 10:00 and 19:59 and peak time of 12:00 to 13:59 instead of 16:00 to 17:59. The main reason for the distribution difference is that most of the people are available during the whole day and therefore it is possible to make appointments with other people earlier. The average number of calls during the period of 01:00

to 06:59 was 79.57, which is 0.48% of all calls during the days. This means that nightly activity is 2.38 times bigger during the weekend than during the working days. This can be explained with the fact that more people are making plans during the night due the next day being free of duties. Another difference compared to working days was activity increasing and decreasing periods. In contrast to 04:00 to 16:59, activity increasing periods during the weekend are 05:00 to 13:59 and 23:00 to 00:69. Activity decreasing periods are 00:00 to 05:59 and 13:00 to 23:59. The average number of calls in an hour was 696.81.

6 Effect of events on call activity

This chapter describes the impact of different natural events and different non-natural events on call activity.

6.1 Impact of natural events on call activity

This subchapter describes the impact of natural events such as weather, full moon and solar eclipse on call activity.

6.1.1 Impact of weather on call activity

Figure 10 shows daily call activity in comparison with the average temperature. The days are sorted in the order they appear in week. For average temperature, a table was created that displays temperature and weather conditions of four Estonian most populous cities (Tallinn, Tartu, Narva, Pärnu) at hours 00:00, 06:00, 12:00 and 18:00 over the observable period of March 1st, 2015 to March 31st, 2015. High-temperature value was used for each of the hours. The source of weather data is Time and Date AS in combination with Estonian Weather Service. The average temperature during the observed period was 4.02 degrees Celsius.

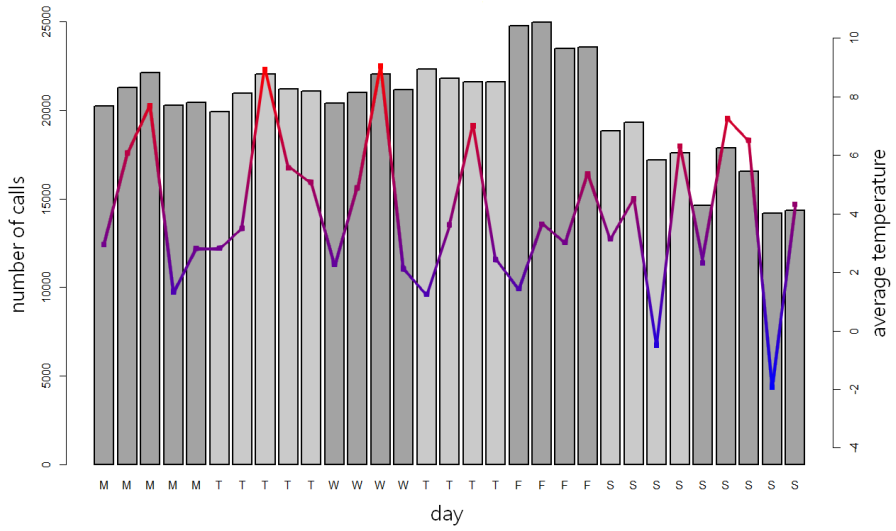

Figure 10. Call activity and temperature comparison

It can be seen from Figure 10 that the coldest day of the month (-1.94 degrees Celsius) was also the least active of the month. Although the day was Sunday (14188 calls compared to

15523 on average, 8.60% below average), which is the least active day of the week, it was the least active out of all Sundays. The preceding Saturday, which was the second coldest day of the whole month (-0.5 degrees Celsius), was also the least active Saturday (17198 calls compared to 18225.25 on an average, 5.64% below average) of the month.

It was also noted that the warmest day of the month (9.06 degrees Celsius) was the most active Wednesday of the month (22053 calls compared to 21152.75 on average, 4.36% above average). In general, the four most active days were also the warmest of that particular day of the week and the four least active days were the coldest of that particular day of the week.

6.1.2 Lunar effect on call activity

Figure 11 describes the number of calls for the observed period during the nights between 22:00 and 06:00. The period of 22:00 to 06:00 was chosen because it was the interval when the moon was visible and people are generally sleeping.

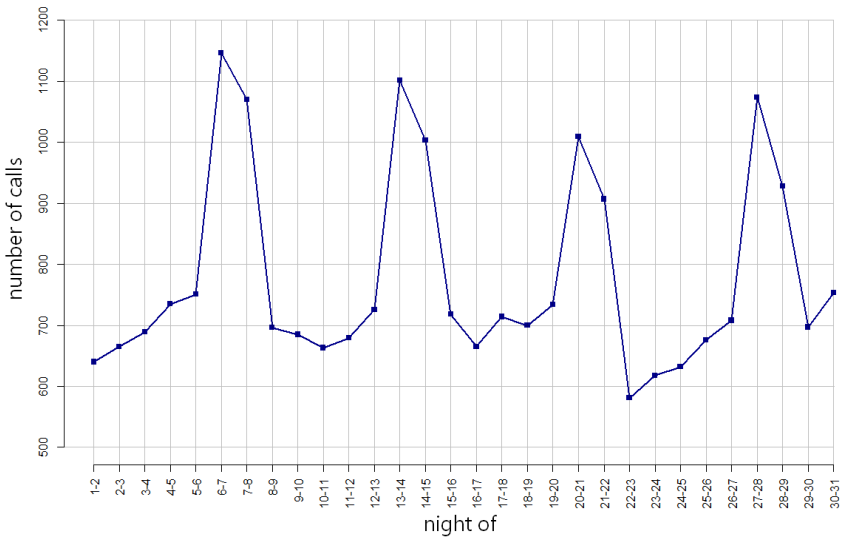

Figure 11. Lunar effect on call activity. Number of calls during the nights between 22:00 and 06:00

During the month, the full moon occurred in the nights of 5-6 and 6-7 [27]. For the night of 5-6, the full moon was in a form of Micromoon and also the furthest and smallest of the year [28]. The following night of 6-7 was a continuation of the full moon. Although the night of 4-5 was also considered to have the full moon in some areas of Earth [29], it was not at 100% visibility and still growing according to Estonian data. The night of 5-6 when the Micromoon

occurred, was between Thursday and Friday and 751 calls were registered. The average number of calls between Thursday and Friday night was 729.25, which means a 2.91% increase in activity. The following night of 6-7 when full moon also occurred, was the most active night of the month with 1146 calls registered in comparison to 1082.75 that was average between the nights of Friday and Saturday. This means a 5.84% increase in activity. It is notable that although the night of 4-5 had moon still in the growing phase, it was still the most active night of month between Wednesday and Thursday with a 5.38% increase in activity. Based on the activity displayed in Figure 11, there is a reason to believe that the full moon does make people more active during the night.

6.1.3 Impact of a solar eclipse on call activity

Figure 12 describes the number of calls during the period of 11:00 to 13:17 for each day, which is the time interval when solar eclipse in March 20th [30].

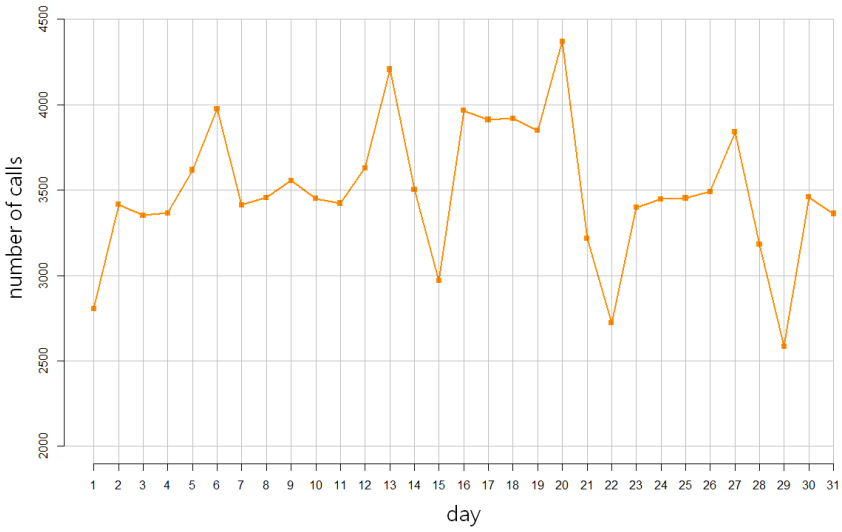

Figure 12. Number of calls in a day between 11:00 and 13:17

The day when the solar eclipse took place in that month was Friday. In Estonia, the solar eclipse was partial [30]. The average number of calls for that time interval was 3494 in a day and 4098 when only Fridays were considered. During the solar eclipse day, there were 4370 calls registered during the period between 11:00 and 13:17. Therefore, the activity increase during the solar eclipse was 6.64% compared to other Fridays. It is also notable that the preceding 4 days long period from Monday to Thursday before the solar eclipse was more active than usual. The

reason for this is that all four warmest days of the month were during that period: average temperature was 8.17 degrees Celsius compared to month's average of 4.02 degrees Celsius.

6.2 Impact of non-natural events on call activity

This subchapter describes the impact of non-natural events such as parliamentary elections, football match and infamous Friday the 13th on call activity.

6.2.1 Impact of parliamentary elections on call activity

To analyse the effects of parliamentary elections on call activity, two Estonian counties were chosen: the county with the lowest and the county with the highest percentage of voters during the election day of March 1st, 2015 [31]. Electronic votes, which are a part of elections result in Estonia, were excluded due them being cast during the time before the observable period. Other votes cast before the elections day were also excluded. In addition, Harju county and Tartu county were excluded for accuracy, because they include cities Tallinn and Tartu which form a separate voting district. The counties which activity was chosen for analysis are Ida-Viru county and Hiiu county. After excluding beforehand cast votes and excluded areas, Ida-Viru county had the biggest percentage of voters (83.27%). Hiiu county had the lowest percentage of voters with the same consideration (75.80%). Elections day of March 1st, which was Sunday, was compared with other Sunday with the most similar amount of calls (1.83% difference in activity): March 29th. Network comparison between elections Sunday and typical Sunday for Ida-Viru county can be seen in Figure 13. Network comparison for the same difference for Hiiu county can be seen in Figure 14. Towns of the counties are also displayed in Figures 13 and 14.

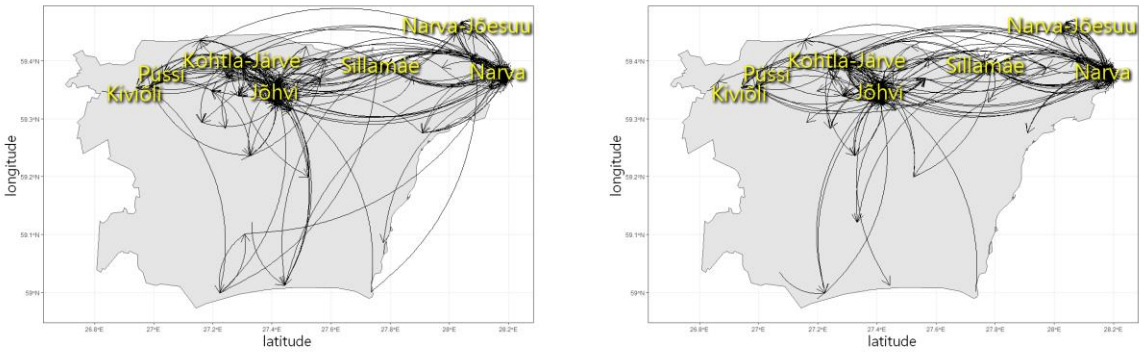

Figure 13. Call activity in Ida-Viru county during the 2015 parliamentary election voting period (Sunday, March 1 between 9:00 and 20:00) on the left side compared to typical activity in Ida-Viru county (Sunday, March 29 between 9:00 and 20:00) on the right side

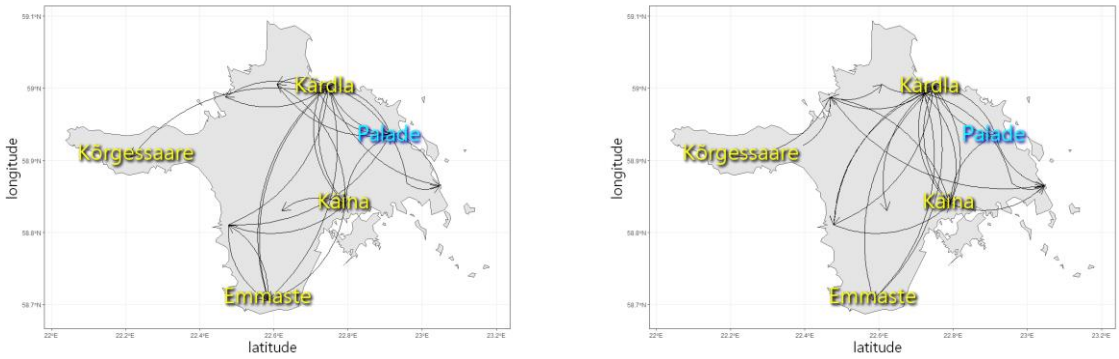

Figure 14. Call activity in Hiiu county during the 2015 parliamentary election voting period (Sunday, March 1 between 9:00 and 20:00) on the left side compared to typical activity in Hiiu county (Sunday, March 29 between 9:00 and 20:00) on the right side

In Hiiu county, the activity during the election day was 7.5% higher than during the other observed day. For Ida-Viru county, activity during the elections was just 0.9% higher compared to the other observed day. In the case of Hiiu county, the calls trajectory is similar in both days, with an exception of the Palade parish (population of <30, displayed in blue color in Figure 14) receiving more calls as one of the four polling stations of Hiiu county was there [32]. In the case of Ida-Viru county, there was bigger call activity around smaller parishes as there are 13 parishes there with polling station in addition to the main seven towns (shown in Figure 13).

6.2.2 Football match and call activity

On March 31st, there was an Estonia-Iceland football match [33] and Figure 15 describes the number of calls inside and around the A.Le.Coq arena where the match took place. There were

no other matches during the observed period in that location. A.Le.Coq arena is the arena where international matches featuring the Estonian national football team are played. The gray bar shows the total number of calls for each day. The blue bar shows how many of these calls were made when only the period of 17:00 to 22:00 is considered: the period from gates opening until 75 minutes after the match is finished during the match days. There were 5334 spectators during the match. The used dataset included 573 calls from the observed period around the arena giving an average of 18.48 calls per day. During the match day, there were 34 calls, 1.84 times more than the average. For the period of 17:00 to 22:00, the average number calls is 6.65. During the match time, there were 19 calls, which is 2.86 times above the average number of calls for that period. 42.11% of these calls were made before the match, 31.58% of calls were made after the match and the remaining 26.32% of calls were made either during the very beginning, very end of the match or during the break.

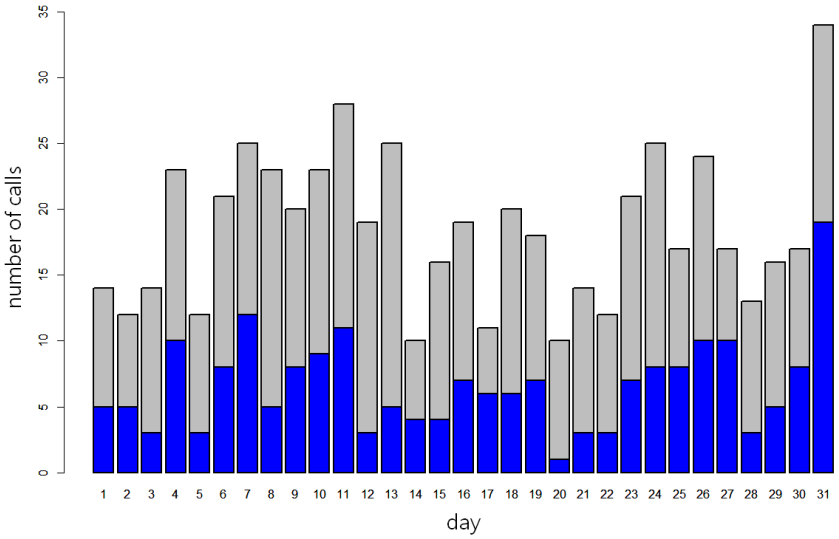

Figure 15. The number of calls around the main football arena in Estonia during March 2015. Blue color shows the number of calls during the typical match related period (17:00 to 22:00)

In conclusion, people were focused on the match and avoided making calls during the match time.

6.2.3 Impact of Friday the 13th on call activity

Infamous Friday the 13th is a day that occurs one to three times in a year [34] and is considered an unlucky day in many cultures, including the culture of Estonia. One of these Fridays occurred during the observed period. The number of calls made during that day was 24983, which was 3.28% above the average of all the Fridays, making it the most active day of March 2015. The number of calls made during the observed Fridays is displayed in Figure 16.

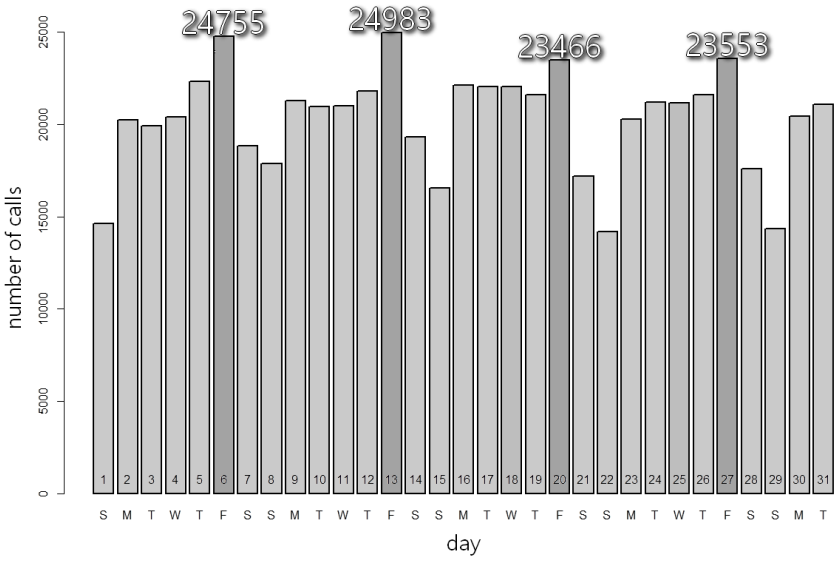

Figure 16. Number of calls by day with Fridays highlighted

There were no other notable events during that Friday, including the previously mentioned natural and social events. It also was not an effect of good weather that might have caused the activity to increase, as it was not the warmest Friday (3.64 degrees Celsius compared to the monthly average of 4.02, the warmest Friday was 5.38 degrees Celsius). Therefore, it can be assumed that Friday the 13th also affected the calling activity like all the previous events.

7 Summary

7.1 Conclusion

Call Data Records, which are collected by mobile operators to gather information for billing, are valuable data to predict human behavior and investigate social networks of different countries. Their value comes from the fact that they include IDs of calling parties, location coordinates and timestamp. This information can be combined with many other data such as statistics, surveys or even self-reported data in order to solve problems in a variety of fields like behavioral analysis, urban analysis, health or crime.

The first part of Call Data Records analysis in this thesis was a Descriptive Analysis. Firstly, after cleaning the data, 628 716 Call Data Records were used to understand the nature of Estonia's social network. Secondly, call activity was investigated over Estonian counties to find the connections between counties, population and popularity differences and most and least in-area centered counties. In the final part of this chapter, calling patterns were investigated by days and hours, with hours being investigated also by considering only weekdays and only weekend days.

Descriptive Analysis was followed by analysing the effects of events on call activity. The events were split into two categories. The first category was natural events, from which the impact of weather, full moon and solar eclipse on call activity were analysed. The second category was non-natural events. From non-natural events, it was analysed how parliamentary elections, a major football match and infamous Friday the 13^{th} affected people's activity. In the context of the parliamentary elections, call connections were also displayed on a map within the counties with proportionately the biggest and smallest amount of votes cast during the official elections day.

The results of this thesis indicate that human calling activity depends on the time period of calling and it also gets impacted by events happening around it. Among the results based on Estonian counties, it was also discovered that the largest counties are most in-area centered, meaning that a large amount of the calls beginning from the county have a destination within the same county.

7.2 Future perspectives

It is possible to combine Call Data Records with data of many events that were not covered in this thesis. In this thesis, the most notable events of the observed period (March 2015) were chosen. One of the possible future perspectives is to repeat the process with a larger amount of data covering more months in order to investigate how other events affect call activity. Furthermore, including more Call Data Records can enhance the accuracy of Descriptive Analysis results.

In this thesis, the analysis part was based on Call Data Records that were recorded in Estonia and therefore, discovered patterns are based mostly on Estonians behavior. It is possible to repeat the process with data from other countries in order to find differences between different sociocultural backgrounds.

References

[1] E. Letouzé, P. Vinck and L. Kammourieh, "The Law, Politics and Ethics of Cell Phone Data Analytics," Data-Pop Alliance, April 2015. [Online]. Available: http://datapopalliance.org/wp-content/uploads/2015/04/WPS_LawPoliticsEthicsCellPhoneDataAnalytics.pdf. [Accessed 13 January 2019].

[2] "Mobile cellular subscriptions 1960-2017," The World Bank, 2017. [Online]. Available: https://data.worldbank.org/indicator/IT.CEL.SETS. [Accessed 13 January 2019].

[3] "Elektroonilise side ülevaade - II kvartal 2017," Tehnilise Järelevalve Amet, 2017. [Online]. Available: https://www.tja.ee/sites/default/files/content-editors/Sideulevaated/elektroonilise_side_ulevaade_ii_kv_2017.pdf. [Accessed 13 January 2019].

[4] A. Ø. Sørensen, J. Bjelland, H. Bull-Berg, A. D. Lundmark, M. M. Akhtar and N. O. Olsson, "Use of mobile phone data for analysis of number of train travellers," 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2210970618300192. [Accessed 13 January 2019].

[5] "Regulation (EU) 2016/679 of the European Parliament and of the Council," 27 April 2016. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1528874672298&uri=CELEX%3A02016R0679-20160504. [Accessed 13 January 2019].

[6] Office of the Data Protection Ombudsman, “Pseudonymised and anonymised data,” [Online]. Available: https://tietosuoja.fi/en/pseudonymised-and-anonymised-data. [Accessed 13 jaanuar 2019].

[7] S. Phithakkitnukoon, T. Horanont, G. D. Lorenzo, R. Shibasaki and C. Ratti, “Activity-Aware Map: Identifying Human Daily Activity Pattern Using Mobile Phone Data,” 2010. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-14715-9_3. [Accessed 27 January 2019].

[8] R. Ahas and Ü. Mark, “Location based services—new challenges for planning and public administration?,” 2005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0016328704001521. [Accessed 26 January 2019].

[9] C. Ratti, D. Frenchman, R. M. Pulselli and S. Williams, “Mobile Landscapes: Using Location Data from Cell Phones for Urban Analysis,” 1 October 2006. [Online]. Available: https://journals.sagepub.com/doi/pdf/10.1068/b32047. [Accessed 26 January 2019].

[10] N. Eagle and D. Lazer, “Inferring Social Network Structure using Mobile Phone Data,” 2007. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.379.4719&rep=rep1&type=pdf. [Accessed 26 January 2019].

[11] J. L. Toolea, S. Colakb, B. Sturta, L. P. Alexander, A. Evsukoff and M. C. González, “The path most traveled: Travel demand estimation using big data resources,” 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0968090X15001631. [Accessed 29 January 2019].

[12] M. Kumar, M. Hanumanthappa and T. V. Suresh Kumar, “Crime investigation and criminal network analysis using archive call detail records,” 2017. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7951743. [Accessed 15 February 2019].

[13] E. S. Khan, H. Azmi, F. Ansar and S. Dhalvelkar, “Simple Implementation of Criminal Investigation using Call Data Records (CDRs) through Big Data Technology,” 2018. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8537389. [Accessed 27 January 2019].

[14] K. Farrahi, R. Emonet and M. Cebrian, “Predicting a Community’s Flu Dynamics with Mobile,” 2015. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01146198/document. [Accessed 28 January 2019].

[15] K. H. Jones, H. Daniels, S. Heys and D. V. Ford, “Public Views on Using Mobile Phone Call Detail Records in Health Research: Qualitative Study,” 2019. [Online]. Available: https://mhealth.jmir.org/2019/1/e11730/. [Accessed 28 January 2019].

[16] RStudio, “RStudio,” [Online]. Available: https://www.rstudio.com/products/rstudio/. [Accessed 14 Match 2019].

[17] Gephi, “Gephi. About,” [Online]. Available: https://gephi.org/about/. [Accessed 19 March 2019].

[18] R Core Team and contributors worldwide, “Documentation for package ‘base’ version 3.7.0,” [Online]. Available: https://stat.ethz.ch/R-manual/R-devel/library/base/html/00Index.html. [Accessed 8 May 2019].

[19] igraph, “Get started with R igraph,” [Online]. Available: https://igraph.org/r/. [Accessed 10 April 2019].

[20] R Documentation, “Creating igraph graphs from data frames,” [Online]. Available: http://cneurocvs.rmki.kfki.hu/igraph/doc/R/graph.data.frame.html. [Accessed 8 May 2019].

[21] R igraph manual pages, “Simple graphs,” [Online]. Available: https://igraph.org/r/doc/simplify.html. [Accessed 8 May 2019].

[22] R Documentation, “Read, Write, Format Excel 2007 and Excel 97/2000/XP/2003 Files,” [Online]. Available: https://www.rdocumentation.org/packages/xlsx. [Accessed 8 May 2019].

[23] R Documentation, “A Grammar of Data Manipulation,” [Online]. Available: https://www.rdocumentation.org/packages/dplyr. [Accessed 8 May 2019].

[24] CRAN, “Simple Features for R,” [Online]. Available: https://cran.r-project.org/web/packages/sf/vignettes/sf1.html. [Accessed 8 May 2019].

[25] Tidyverse, “Tidyverse packages,” [Online]. Available: https://www.tidyverse.org/packages/. [Accessed 8 May 2019].

[26] “RV022: RAHVASTIK, 1. JAANUAR --- Aasta, Sugu, Maakond ning Vanuserühm,” Statistikaamet, 2015. [Online]. [Accessed 7 April 2019].

[27] ilm.pri.ee, “Kuufaaside kalender. Märts 2015,” [Online]. Available: https://ilm.pri.ee/kuufaaside-kalender?month=3&year=2015. [Accessed 8 April 2019].

[28] ilm.ee, “Täna õhtul näeme mikrokuud,” [Online]. Available: https://ilm.ee/?513460. [Accessed 8 April 2019].

[29] Calendar-12.com, “Moon Phases March 2015,” [Online]. Available: https://www.calendar-12.com/moon_calendar/2015/march. [Accessed 8 April 2019].

[30] Astronoomia, “Päikesevarjutus 20. märtsil 2015,” [Online]. Available: http://www.astronoomia.ee/vaatleja/7456/paikesevarjutus-20-martsil-2015/. [Accessed 12 February 2019].

[31] Vabariigi Valimiskomisjon, “Hääletamisest osavõtu statistika (2015),” [Online]. Available: http://rk2015.vvk.ee/participation.html. [Accessed 18 April 2019].

[32] Vabariigi Valimiskomisjon, “Valimisjaoskonnad. Hiiumaa. Pühalepa vald,” [Online]. Available: http://info.rk2015.vvk.ee/valimisjaoskonnad/0039/0639. [Accessed 23 April 2019].

[33] Eesti Jalgpalli Liit, “Protocol 1152,” [Online]. Available: http://jalgpall.ee/koondis/1/mangud/protocol/1152. [Accessed 20 April 2019].

[34] “When is Friday the 13th?,” Time and Date AS, [Online]. Available: https://www.timeanddate.com/calendar/weekday-friday-13?ext=1. [Accessed 23 April 2019].

Appendix

I. Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Hendrik Hiir**,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,
"**The impact of external factors on Estonian mobile call activity"**,
supervised by Rajesh Sharma and Anto Aasa.

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

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

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

Hendrik Hiir

10/05/2019