UNIVERSITY OF TARTU
Institute of Computer Science
Computer Science Curriculum

Mohammad Mahdi Mohebbian

Real-Time Event Detection System for Mobile Data

Master's Thesis (30 ECTS)

Supervisor(s): Amnir Hadachi, PhD
Erki Saluveer, PhD

Tartu 2020

Real-Time Event Detection System for Mobile Data

Abstract:

Mobile data is one of the prior data sources which can be used for urban study analytics due to the amount of valuable information they contain, such as type of mobile data, time and most importantly data's coordinates. In order to keep the cell services stable for users all across the country it is crucial for authorities to be aware of unannounced gatherings which can cause traffic overload on cell towers in the area. In this thesis implementation of a enterprise system has been demonstrated for monitoring the behavior of the cell towers under the administration's authority. The core functionality of this system is detecting ongoing events in different areas on an hourly-basis schedule utilizing multiple statistical approaches for abnormality detection. The output of the event detection section of the system is an approximate estimation of the ongoing event's location on the map. Current design of the system is aiming to fullfill the downsides of similar approaches for event and crowd detection such as high processing expenses and non-comprehensive resources by using parallel servers, distributing the processing load while keeping the pipline clear for user's demands, and utilizing Call Detail Records (CDR) data as input resources which gives the advatages of containing the majority of mobile transactions and human behavior in the city.

Keywords:

Mobile Data, Event Detection, Crowd Detection

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

Mobiilse andmeside reaalajas sündmuste tuvastamise süsteem

Lühikokkuvõte:

Mobiilne andmeside on üks eelnevaid andmeallikaid, mida saab kasutada linnauuringute analüüsimisel, kuna selles sisalduvat väärtuslikku teavet on palju, näiteks mobiilse andmeside tüüp, aeg ja mis kõige tähtsam - andmete koordinaadid. Kärgiteenuste stabiilsuse tagamiseks kogu riigi kasutajate jaoks on ülioluline, et ametivõimud oleksid teadlikud etteteatamata kogunemistest, mis võivad põhjustada liikluse ülekoormamist piirkonna rakutornides. Selles lõputöös on demonstreeritud ettevõtte süsteemi rakendamist administratsiooni alluvuses olevate rakutornide käitumise jälgimiseks. Selle süsteemi põhifunktsioon on eri piirkondades toimuvate sündmuste tuvastamine tunnipõhise ajakava abil, kasutades kõrvalekalde tuvastamiseks mitut statistilist lähenemisviisi. Süsteemi sündmuste tuvastamise sektsiooni väljund on ligikaudne hinnang käimasoleva sündmuse asukohale kaardil. Süsteemi praeguse disaini eesmärk on täita sündmuste ja rahvahulga tuvastamise sarnaste lähenemisviiside, nagu näiteks suured töötlemiskulud ja mittetäielikud ressursid, varjuküljed, kasutades paralleelseid servereid, jaotades töötlemiskoormuse, hoides samal ajal torustiku kasutaja nõudmistest lahti ja kasutades kõne Detail Records (CDR) andmed sisendressurssidena, mis annab eelised suurema osa mobiiltehingute ja inimeste käitumise kohta linnas.

Võtmesõnad:

mobiilne andmeside, sotsiaalse kogunemise tuvastamine, rahvahulga tuvastamine

CERCS: P170 (Arvutiteadus, arvuline analüüs, süsteemid, juhtimine)

Acknowledgement

In the beginning I would like to express my appreciation to Positium company's CEO Erki Saluveer for proposing this topic and giving feedback with such a positive attitude and motivations, company's CTO Marko Peterson for providing me with the initial necessary data and technical support for deploying this system on company's servers and Positium staff in general for their warm hospitality and motivational spirit in many different stages of work.I would love to express my gratitude toward Dr Amnir Hadachi for always believing in me and with unbelievable positive energy always motivated me to try different methods and approaches when the previous ones failed.

I also like to thank my lovely wife Saina for always believing in me and encouraging me to do my best in the whole process and helping me to finish this document in time. I would like to express my love and appreciation for my parents who always put my education and academic life their priority, even though they are far from me right now they always want the best for me.

Table of Contents

1. Introduction

In this section of the thesis main objective, goal and the importance of this project will be demonstrated alongside a brief look at the next sections and what information you can expect in the following sections.

In this thesis you will be introduced to the concept of mobile positioning data and how this source of data can be used in urban study for a variety of projects and subjects. In this project implementation of a system for detection of social events and gathering using mobile positioning data will be demonstrated. Using this system, the authorized user can detect on going events throughout the city on hourly basis with an estimation of the location of the event. This project was created and tested based on synthetic data modeled after real world user's interaction and also real data. The goal is to have high accuracy on detection of the location and the time of the events in the data. The results from this project can be beneficial in numerous situations other than events, since not all of the abnormalities detected by this system indicate an event, and in most cases detection of these abnormalities can assist avoiding mobile connection bottlenecks when a certain area in the city gets a higher demand than anticipated.

In the upcoming sections you will be familiarized with the terms and notations used in this area of study. The theoretical information required will be provided consequently in order to cover the basic knowledge you might need for every aspect of this project. The document will follow with the inspection and analysis of input data for the system and description of implemented methods for the main task of event detection. The results of the tests on the system will be provided in the results and discussion segment for both synthetic data and real data, and the documentation will be concluded with acknowledgement and references involved in this particular project.

2. Terms and Notations

In order to make sure everyone understands the thesis content in the upcoming section, it is helpful to clarify the special terms and notations commonly used in the text. The description provided here is brief and concise since the complete overview of these terms and concepts will be mostly covered in the theoretical background section.

Cell: is the term used to describe a Cell site or Cellular base station which commonly are in the form of tall towers carrying antennas throughout the city.

MPD: Is the abbreviation of “Mobile Positioning Data”. In general, MPD data represents the data gathered from tracing of mobile data transactions which the location of the phone is determined.

CDR: is the abbreviation of “Call Detail Record”. Call detail record is one form of the data mentioned in the MPD section which we use to track a user's location. Mainly CDR data are the data we work with in this thesis.

Events: is the term used in order to describe social events and gatherings in the city. In this document anytime the word “event” is used, a social gathering concentrated in a certain area of the city is the intent of the word.

Traffic: is the term used in this document in order to describe the amount of CDR data associated with cells. If the cell would be involved with higher mobile communication transactions the traffic would be higher.

Transceivers: is a term used to describe a device in radio communications that is able to transmit but also receive information as well. This device is a combination of transmitter and receiver hence the name transceivers.

Big data: is a term used to describe a specific type of data resources that due to their massive volume it is not possible to process them all at once and they require particular techniques and technologies such as “Spark” or “Partial data processing” in order to be analyzed.

API: is the acronym for Application Programming Interface, which is a software intermediary that allows two applications to talk to each other. This phrase is mostly used in web servers where the user's demanded application is accessed through a particular URL call.

3. Theoretical Background

This section provides you all of the information necessary in order to understand this project, please keep in mind that concepts and materials covered in this section can be explored on a much deeper level but in order to avoid confusion information unrelated to this study won't be covered in this chapter.

3.1. Machine Learning

Machine learning is a subgroup of Artificial Intelligence which explicitly involves applications and algorithms that have the ability to learn, improve and optimize automatically without being particularly programmed for. The main resource of these types of applications are vast amounts of data, which the program uses in order to improve itself over time. Regardless of different methods of machine learning the essential procedure is the same. Machine learning models accept two main batches of input data called "Train data set" and "Test data set". Train set will be used in order for the model to learn the data patterns and improve itself. Test data will be used after the learning is done and the model is ready in order to assess and evaluate the final status of the machine learning model and its improvement. Each data set contains a number of "features" which are the columns of data used in the learning process and a "label" which is the answer that the model is meant to be predicting. Machine learning is an extremely wide topic with various different categories, therefore the focus will be on the following specific contents of machine learning:

- Classification: is a specific type of models designed to predict fixed outcomes [1]. In classification models, the labels are limited and indicating which category that row of data belongs to. As an example a model which uses physical body characteristics as features and predicts the gender is considered to be a classification task because simply the outcomes are limited to either "Male" or "Female".

- Regression: has the same basic principles as classification but the only difference is that the outcomes are not limited to a certain number of classes and categories. As an example a model which uses country, city and job specialty as features and predicts the salary of an employee is considered to be a regression task due to unlimited possibilities of outcomes.

- Neural Networks: is a complex of algorithms that aim to detect underlying relationships in a data set through a process that mimics the way the human brain operates [2]. Neural networks can adapt to changing input; so the network generates the best possible result without needing to redesign the output criteria.

3.2. Cells

Cells also known as Cellular base stations are the main stations supplied with antennas and electronic communication equipment, built commonly on top of radio towers. The main structure of a Cell mainly supports antenna, a set of transceivers, digital signal processor unit, a gps receiver for current time estimation, control electronics, primary electronic power sources and backup electronic power sources. The purpose of a cell tower is to aid the signal reception of cell phones and other wireless communication devices like telephone, television, and radio in a cellular network.

3.2.1. Cell Towers Structure

Cell towers are structures built on high masts in order to cover more areas in the surrounding. Since Cell towers can be provided by different companies at times different companies will use the same mount their antennas in order to save money. Typically, a cell tower consists of at least 3 antennas with 120 degrees' coverage area for each antenna, This arrangement can be different with different subdivision of degrees and different number of antennas. Each antenna receives the signals from its corresponding coverage area. The range of a cell tower is dependent on various of parameters such as:

- The height of the antenna
- The frequency of the signal in use
- The rated power of the transmitter
- Ambient weather condition in the cell tower area

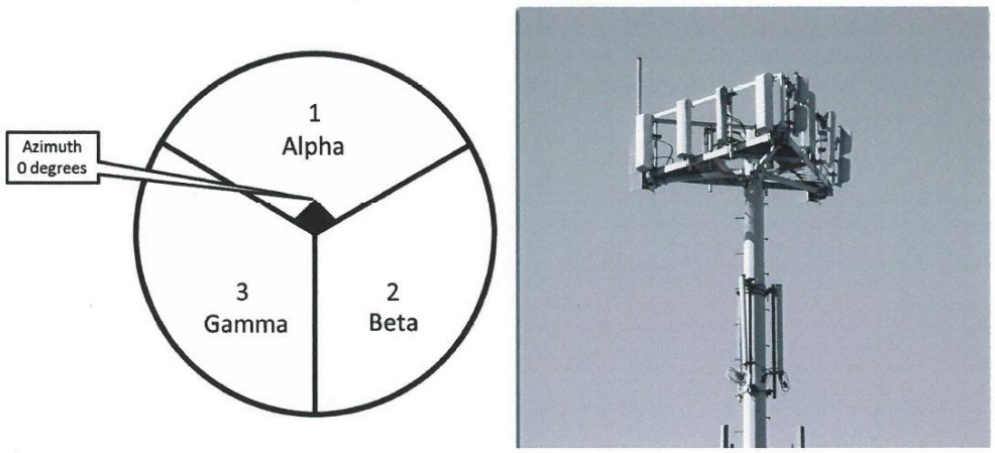


Figure 1. demonstration of antennas on cell towers [3].

Next to the elevated structure of the cell tower there is a base transceiver station also known as Base Station Subsystem(BSS). The signals received from the antennas are transferred to the base station for further processing and call redirection.

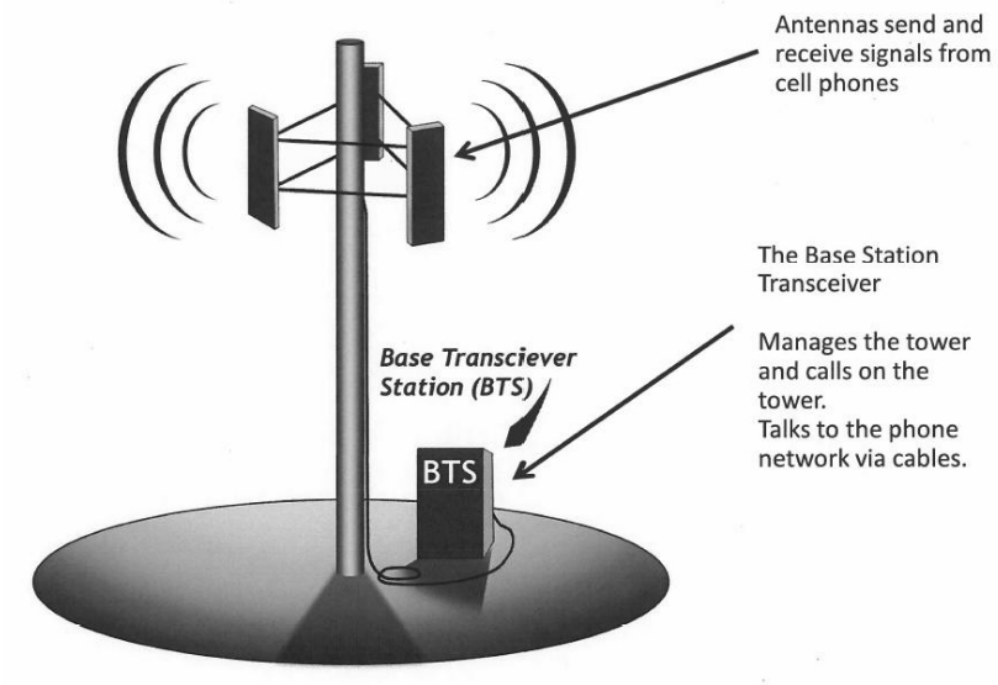


Figure 2. Schematic demonstration of transactions between cell tower and BTS [3].

3.2.2. Cell Towers working principle

Every time that you use your cell phone to make a call, your mobile phone emits electromagnetic radio waves aka radio frequency. After this action the nearest cellular tower will receive the RF(Radio Frequency). As explained earlier the antennas on the cell towers have the ability to both receive and transmit signals. The received signal from the mobile phone will be transmitted to a switching center which acts as a telephone exchange for mobile phones. This transmission allows the call to be connected either to another mobile phone or to be transferred to a network of telephones.

The placement of the cellular towers in the city creates a cellular network which provides the cell service coverage for mobile phone users. As mentioned regularly cell towers contain 3 antennas and coverage of each antenna will collide with antennas from other towers creating the coverage network as the figure below.

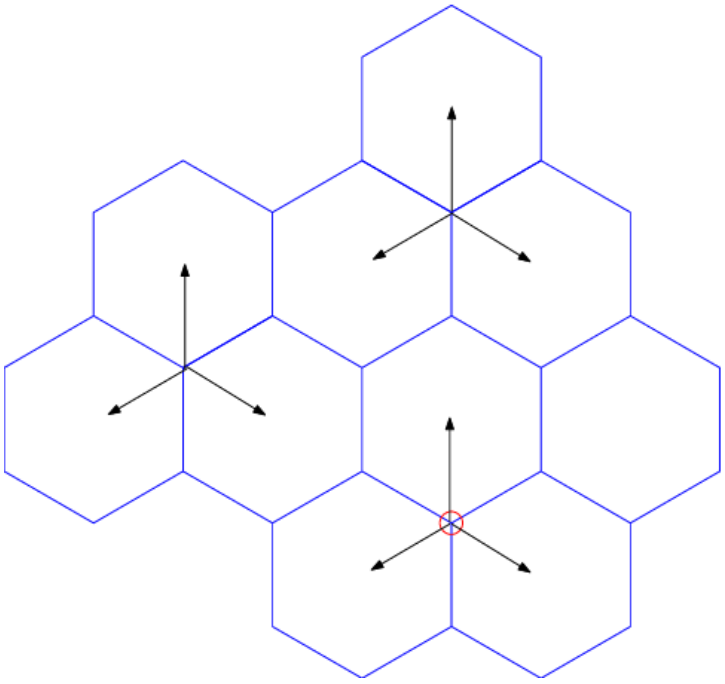


Figure 3. Collision of coverage area between multiple cell towers resulting a coverage network [4].

3.3. CDR (Call Detail Records):

As explained in the cell tower section the antennas are responsible to interact with the incoming signals so that the call would be redirected to its destination. The mobile phone users are known as “subscribers” to the network, and each subscriber has its own unique identifier known as SIM (subscriber Identity Module). The cell site can be administered by any Mobile Network Operator (MNO) companies (i.e. Telia and Tele2 in Estonia). The logs and metadata about the call transformation will be stored in the base station under supervision of the responsible network operator. This data is known as Call Detail Records(CDR) which get stored commonly as csv file formats and it contains information regarding the call signal and its destination with the following fields:

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

This information is collected by the network provider company. Companies with the field of business related to urban studies or research centers can access this data from through the network providers in the country. In order to respect the subscriber’s privacy, the “calling party” and “called party” fields will be modified and changed to an anonymous unique ID, which means every subscriber will be assigned a unique identifier.
In this project the focused fields are “calling party identifier”, “state time” and “CGI”. Other than the CDR records cells information will be requested as well. This additional information contains the CGI identifier for each cell, matched with the cell site’s coordinates. The combination of CDR data and the cells coordinates data, is the basic information that will be used in this project as raw input data. it is important to note that data collected from an antenna will be assigned to that antenna, meaning that each antenna has a unique ID identifier, so the CGIs are not exclusive to a cell tower but they are assigned to individual antennas.

3.3.1. CDR based analytics and research areas

According to GSMA real-time intelligence data, currently, there are 5.20 Billion mobile device owners in the world. This means that 67.11% of the world’s population has a mobile device.

Back in 2017, the number of people with mobile devices was only 53% and breached the 5 billion mark. Statista predicts that by 2023 this number of mobile device users will increase to 7.33 billion. We need to keep in mind that this number comes from all of the world's population with no specific filtering, so the remaining 32.89% also contain certain groups of people who are not allowed or can't access mobile phone services, such as, under age children and people in specific areas without cellular network coverage in rural countryside's.
This statistic makes CDR data a really informative and comprehensive source of information which can be used in various studies and many different fields such as urban studies, traffic behavior, routing optimization, tourism and population monitoring.
The studies in on CDR data can be distinguished into two different large groups:

- Human mobility studies
- Trajectory reconstruction and localization

The first group's focus is mainly on the mobile users behavior analytics by tracking their records from CDR data and the visited locations near cell sites. This area can be linked to many subdivisions such as tourist site visitation patterns, traffic optimization or crowd detection systems etc.
For example, Ratul Sikder et al in [5] implemented a method to distinguish the tourists in the CDR data gathered from the mobile operator datasets using analytical approaches on the large input datasets. Saluveer et al in [6] methodology is in the next stage considering they are not only identifying the tourists among the CDR data but their methodology extract statistical results from tourist's data behavior for different organizations. Honghui Dong et al used CDR data for urban traffic estimation purposes in [7] and identifying different zones and sections of the traffic by analyzing the CDR data from data operators. Another traffic aspect is considering optimization for public transport such as [8] where the implementation of a system for the estimation of the density of subway stations using CDR data in a recurrent neural network. Even projects related to complex social behavior prediction based on CDR data analysis on individual anonymous subscribers done by Casey Doyle et al in [9] proves that CDR data is a viable source of information in many different fields.
On the other hand, trajectory reconstruction studies are more focused filling the missing information in user's movements. Essentially user's movement can be tracked by aggregating the user's connection with different cells in the path, but some specific positions will remain unknown in the path and trajectory reconstruction aid in order to fill in the gaps and help with their sparsity of the data so the movement pattern and positions of the user would be seamless and whole. Toivo Vajakas et al implemented a method in [10] to reconstruct the movement path of CDR data subscribers by following the pattern from cell-to-cell handshake method. Similarly, in [11] movement route reconstruction is the goal but with the taking into account the patterns and similarities in subscriber's previous movement behaviors. The main issues and challenges with the data localization and path reconstruction has been demonstrated in [12] and Guangshuo Chen et al proposed methods for route reconstruction based on tensor factorization. Due to the

progression in the development of different methods new localization and path reconstruction techniques with higher accuracy are being developed such as [13] from Lind et al. with the extension follow up in the next year by Hadachi et al. [14]
This project will be categorized under the first group and the trajectory reconstruction is not going to be the focus. Essentially this thesis is considered to be more of a practical project and implementation rather than a research based project.

3.4. Concept of Event

From the perspective of MPD analytics and research, occasions that contain people gathering in a specific location for an extended amount of time is called an event such as a concert, social gathering, conferences, music festivals, etc. The concept of event detection is a subgroup of crowd detection which is an important concept in urban studies. The effect of events in MPD data will appear in CDR records. As explained earlier the main input data used in this project comes from the combination of CDR records and the cell site's coordinates, using these resources together we can detect the amount of CDR data registered in each cell also known as traffic and the occurrence of an event will result in abnormal peaks in the amount of traffic registered in the cell sites present in the event area.
The event detection concept can aid in order to help maintain the quality of cell services in the event area by detecting the event in time so if necessary portable cell sites could be transferred into the event area in case the event was not announced in advance.
The event detector system needs to be trained via both actual data and synthetic data, in each of these dataset there will be event test cases in order to test the system's functionality. The events in synthetic data are artificial and pre-arrange in the synthetic data sets, and the event in the real data is the Metallica concert which took place on July 18th at "Raadi Airfield" in Tartu, Estonia.

3.5. Problem Statement

The problem that we're addressing in this thesis is the detection of live ongoing events in the city with the approximate location of the event. In many cases the studies and methodologies based on MPD in human mobility related studies are based on historical records from previous years, but in this case the historical records are not going to suffice and analysis on the live data is required as well. The main goal is to use the combination of historic data and live data analytics in order to determine the most probable location for an event in a certain time and date based on CDR data traffic in a network of cell sites in the city. The key and primary source of information which the system will use in order to train for event detection is the cell tower's traffic quantity, ergo in the following chapters the term of "traffic" will be used a great deal. Considering the amount of calls and text messages exchanged between people in a short amount of time, it indicates that the CDR data collected from the cell sites are considered to be big data and the memory usage optimization related challenge is to stream the CDR big data into the system, in a

manner that the process of the newly added data won’t overload the computer’s memory or the processing unit. The methodological challenge is to implement an approach on how to use and manipulate the input data in order to find abnormalities in the cell towers traffic in different areas of the city.

4. Data and Methods

In this chapter, two main concepts of this project will be demonstrated precisely, input data and the system's methodologies. In the first section, the nature and types of input data will be covered in order for everyone to have a clear and comprehensible vision on both synthetic, and the real life raw data used as the input resource. In the second part, the approaches and methods designed in order to fulfill the final goal will be introduced, failed attempts along with the successful attempts.

Essentially for projects such as event detection where the user's location is a crucial aspect of the project, there are a variety of resources that can be used. The options range is too wide to cover all of the methods in this thesis, but some valid examples would be:

- Social media tracking: Using web scraping tools on well-known social media platforms, the trend locations in a city can be found, and the summarization of the retrieved data can lead to clues on crowded locations in the city due to the ongoing events. [15][16][17]
- GPS location tracking: Mobile applications can be a great interface for the people to voluntarily share their accurate location via GPS location tracker on the mobile devices. This method does not require any further data mining in order to track the exact location of the crowd, because the GPS location tracker delivers the accurate coordinates points of the user. [18]
- Image Processing from traffic camera feed: The traffic camera feeds constantly are streaming live feed of current traffic to traffic stations. By applying proper image processing processes on the incoming live camera feed an estimation of the present population of people in each certain part of the city will be obtained. The huge downside to this method is its cost and processing expenses, since processing on stream of images requires a high level of processing capabilities [19][20][21].
- Using CDR data: As explained earlier, a user's mobile transactions such as phone calls or short message services (sms) with cell towers leave a single record in the base station of the cell tower assigned to the corresponding user. By tracing the user's records among the cell towers in an area considering the time of the initialization of the record in different cell sites, we can estimate a route of the taken path by the user on a relatively good accuracy.

Each of the above approaches has their own advantages and disadvantages, which by demonstrating them makes it more conspicuous why this project requires a certain resource over the others. Social Media tracking is a relatively fast pace method to detect ongoing events in the city, because it cuts out the processing time needed in order to detect a crowd from coordinate points due to the concentration of the data, for example, the scraped data from social media will be convenient if the post indicates the location and the time by using indicators such as hashtags, location markers, etc., In that case a simple data filtering can lead to an estimation of the number of people present in the event and detecting the event itself. This method has its particular downsides as well. Many issues will arise due the nature of this type of resource, as an example, information from social media tracing for people's behavior will contain considerable amount of noise (false positives and true negatives) due to the fact that this data does not have a reliable

source and the content is provided by people themselves, and the lack of accuracy is the reason why social media tracing can't be the primary source for such project, but it can be used as confirmation data. The GPS location tracking applications are in an opposite situation compared to the previous method. GPS tracking data is one of the most accurate and reliable sources of information due to the fact that this resource is provided by devices in the form of coordinate points. This is a proof of reliability and accuracy. The primary complication with this particular method is the quantity of data. GPS location tracking requires specific applications with user privileges in order to allow devices to report GPS locations. Lots of security and privacy issues cause trust issues among people regarding use of such applications which causes quantity issues with this resource. Data quantity is an extremely important factor when it comes to event detection, because it can cause false results simply because we do not have access to the whole population's behavior present at the event, which can render an active event as a true negative situation. CDR data is a well-balanced source of information compared to the other sources mentioned above. CDR data is reliable since it is provided by automated systems (cell towers and mobile devices) without direct human contact, and also the quantity is not a problem for CDR data since the primary role of any mobile device is the establishment of a call regardless of device's brand or generation, which makes this data relatively comprehensive resource of data.

4.1. Historic Data

Historic data is one of the primary sources of data used in this project. This study is focused on the year 2019 (due to the Metallica's concert) and the historic data cover previous year, from January 2018 until the end of December 2018. The historic data from the previous year is the CDR data gathered from cells all over Estonia (8934 cell towers) in the form of .csv files with fixed features. Each record in the historic data indicates the occurrence of a single CDR record. The historic raw data contains the following features:

- Cell ID: The Identifier of the corresponding cell tower that the mobile device connected to in order to establish the CDR record.
- Pos_usr_ID: The anonymous identifier assigned to the subscriber who made the call.
- Pos_date: The exact date and time when the CDR instance was recorded.

Historic data's purpose is to predict the amount of traffic in each particular cell tower. But according to the sample CDR records from the historic data above, it is clear that dates and times are accurate to the seconds. The need for data reformatting shows itself at this stage, when we can realize that prediction can not be made for the exact time and dates stated in the historic data because the level of detail in the date and time creates many gaps in the coverage of dates and times, for example, CDR data for a particular cell tower does not cover every second of the dates and times, depending on the time and cell tower's location there are considerable duration of time during the day when the cell tower does not interact with any mobile devices simply because no calls are being made in the tower's area, these durations will be unfilled gaps in the historic data. The data reformatting goal is to transfer the historic CDR data into a comprehensive resource which covers the whole 2018 years without any gaps. In order to accomplish such a task, A series of analysis must be performed on the historic data in order to reveal any specific patterns to replace with the "pos_time" field.

Currently the datetime field also known as “pos_time” shows the exact time of a CDR record, the new changes will transform the data in a manner that the amount of CDR data also known as traffic in a particular cell tower aggregates a unit of measurement, such as a day, an hour, a week or a combination of units based on the patterns revealed from the analysis. The analysis goal is to determine the best and most suited units of measurements in order to get the best and most reliable traffics for historic data.

In the first step of analysis the historic data will be aggregated into daily based traffic, meaning that the current form of data that reports the exact time and date of a CDR record on a particular cell site, will be changed into a new dataset which aggregates the amount of traffic gained by the corresponding cell site in one whole day. The goal is to create a timeseries data from daily traffic of the cells. Depending on the cell location the amount of traffic per day can differ, but in order to have the perfect and comprehensive timeseries, a specific edge case needs to be covered, an edge case which can happen on multiple occasions depending on the cell tower’s location and can create gaps in the final time series, which is cell towers that have no activity in one or multiple days. The absence of CDR records in particular dates will create gaps in the time series and non-comprehensive time series are obsolete for this study. By taking into account this edge case and covering every day in the full scope of the CDR data’s date field, the time series will be without any gap and wholesome and would be ready for further analysis for data reformatting.

Tracking the basic behavior of the cells throughout consecutive months is the most convenient and general method to reveal any hidden pattern that can guid in reformatting the data.

The following figure is a selection of different cell tower traffics in five consecutive months. Individual colors demonstrate different months of data and the whole plot is daily traffic based. By analysing each month's transactions in individual for every cell, the behavior of the pattern becomes more visible. The cell’s activities in the duration of five months demonstrates that each month’s activity is slightly different than the next or the previous month. Essentially isolates the cell's activities during each month. Essentially this means that people’s behavior changes through our different months, and this fact is noticeable even by observing routine life styles of people, specific months mean different to others, as an example, June and July are more often to be used for vacations, this affects the daily traffic in city center area, where there are more companies and work related buildings, and also rural area, where is the usual destination for local visitors and vacations. The same concept can be used in the opposite situation, months where the workload of the companies, universities and schools are the maximum, which means less traffic in rural areas and vacation destinations.

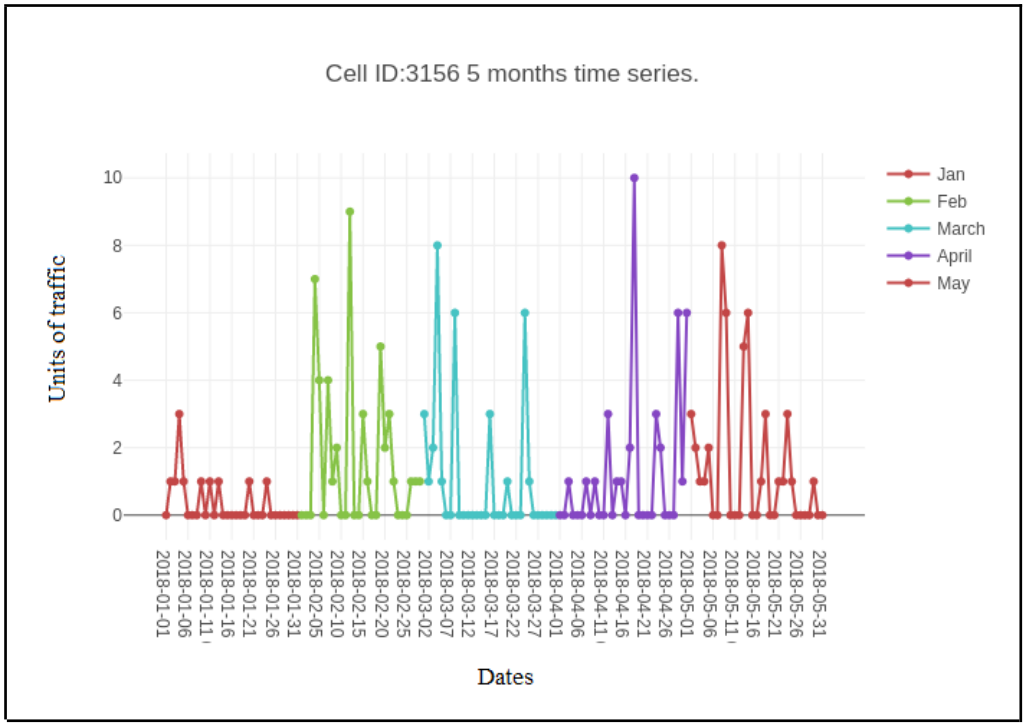


Figure 4. Continous time series of traffic volume of cell ID: 3156 for duration of 5 months .

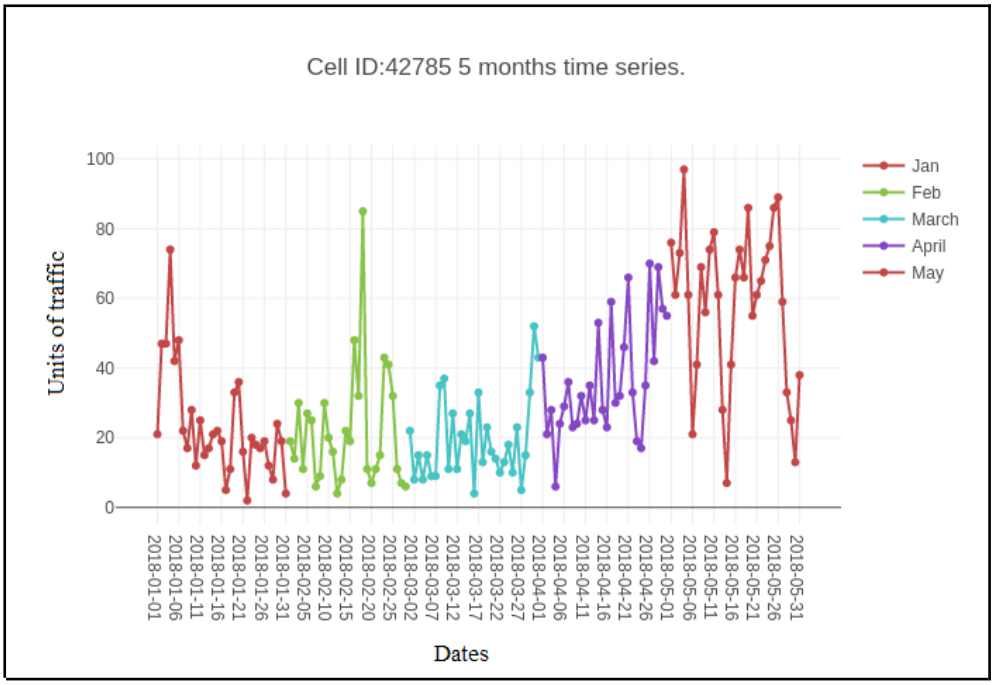


Figure 5. Continous time series of traffic volume of cell ID: 42785 for duration of 5 months .

The cell tower traffics presented above could also be presented as box plots but scatter line plots show the exact details needed for pattern recognition in the data. The comparison of the monthly interactions in each of the above examples demonstrates that, monthly activities in most of the time differ from their next or previous month's activity. The difference will increase more by progressing in the months as an example, in the cell tower with the ID of *589* by comparing the traffic behavior from February to May, the difference in the pattern is completely visible. This pattern won't be seen for all of the cells in all of the months but this won't be the reason to rule out this pattern for data reformatting. This analysis is case sensitive, meaning that finding specific patterns in data such as the monthly pattern explained above, will suffice to use the results in data reformatting.

The larger units of datetime measurements after "Month" will be "Seasons" and "Years". Due to the limitation on historic data which covers only the previous year of the test data, neither "Seasons" nor "Years" can be a sufficient parameter for data reformatting. The year parameter would be redundant since we only have one year of data and "Seasons" will be redundant but not for quantity reasons but due to redundancy in the concept. Essentially the concept behind using "seasonality" as a useful parameter for data reformatting is already being covered by the "month" parameter, so the next patterns should be uncovered in datetime units more detailed than "month".

Units of datetime measurements more detailed than month would be "days" of the month. The goal of this segregation is to find specific parameters that cell tower traffics would differ

depending on that parameter. So far "month" has been proven to be a suitable parameter, moving down to more detailed units there would be "days".

The following analysis targets the validity of "days" as a valuable parameter in the system. In this figure we can see the comparison of the amount of traffic gained on a daily basis in the same scope of the first fifteen days of five months, but grouped by the day's date numbers. In the following plot the same color bars demonstrate the same day number of the month, and since we are monitoring the five months of the year there would be five bar of the same color next to each other, and each group of bars represents day number, and since we are analyzing the first fifteen days of each month, we have fifteen different color grouping. This form of data representation creates the best analytical view on the data in a manner that we can determine whether the day number of each month is in harmony with each other. If any specific correlation can be found in the same day number of each month it would prove the day numbers to be the next fitting parameter for data reformatting.

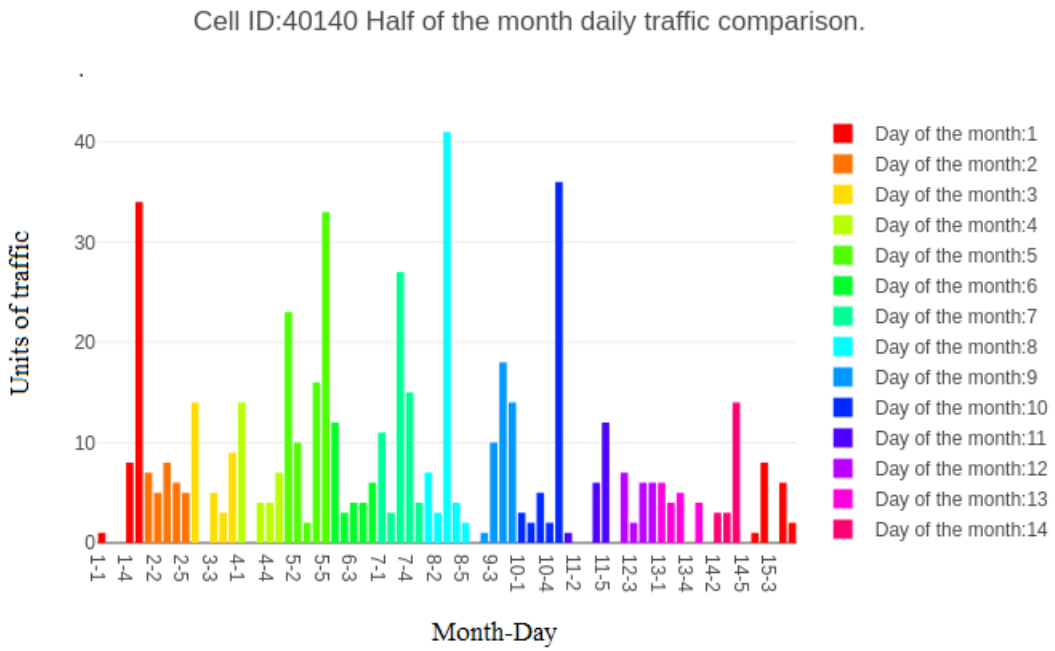


Figure 6. Comparison of daily behavior of cell ID:40140 in the same days during different months.

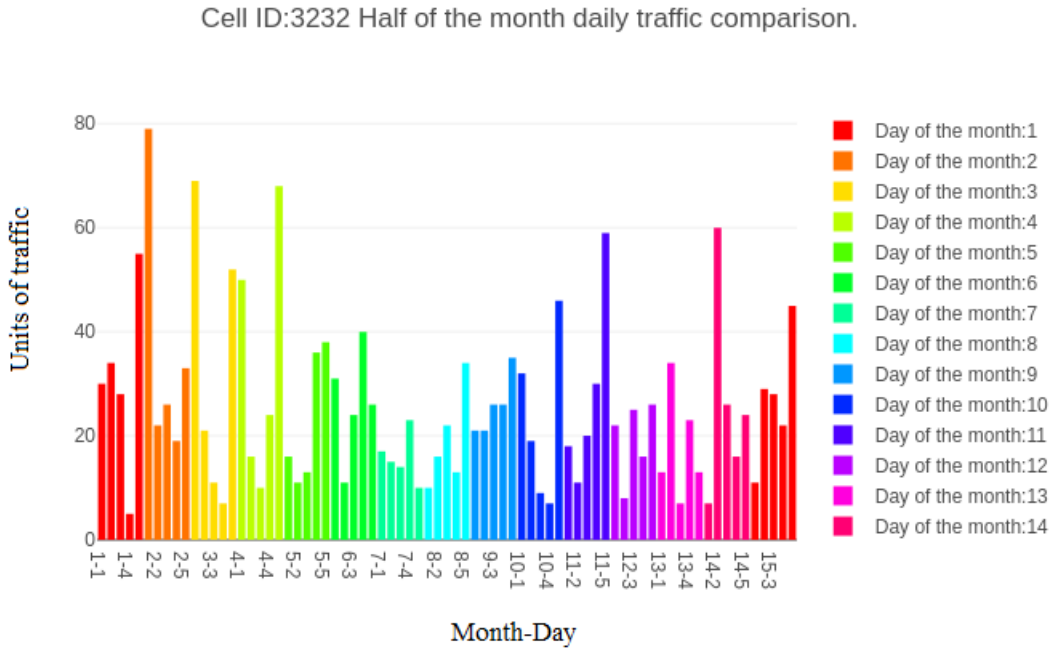


Figure 7. Comparison of daily behavior of cell ID:3232 in the same days during different months.

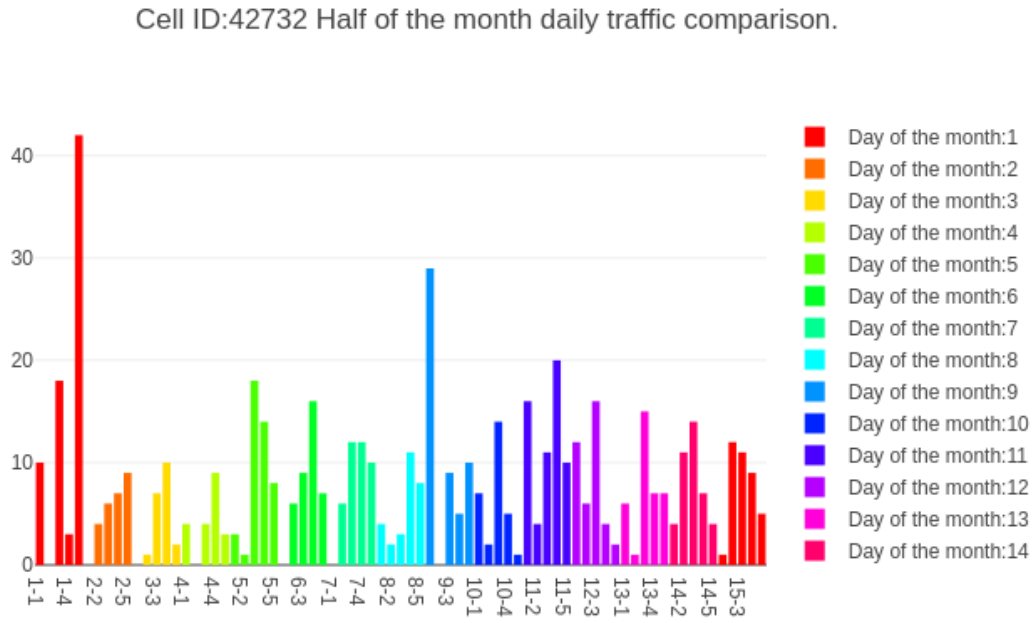


Figure 8. Comparison of daily behavior of cell ID:42732 in the same days during different months.

Three different cell's daily traffic has been shown above. In each plot the daily traffic comparison of a particular cell tower is demonstrated grouped by the number of the day. For example, let's consider the cell tower with the ID of 3232, the color codes demonstrated that the red color shows "Day of the month: 1", this means that the first five columns which have red color, indicate the first day of the first five months (1st of January, 1st of February, 1st of March, 1st of April and 1st of May). By analyzing the individual group's behavior in each of the plots, the conclusion of the validity of "day's number of the month" as a data reformatting parameter will form. As a more detailed study on the presented plots, the first plot's detailed statistical information is presented in the following table.

Table 1.Statistical analysis of the traffics on the same days but on different months for cell ID 40140.

Day Number of the Months (Jan,Feb,March,April,May)	**Average**	**Standard Deviation**
1st	**8.6**	**14.5876**
2nd	**6.2**	**1.3038**
3rd	**6.2**	**5.4497**
4th	**5.8**	**5.2153**
5th	**16.8**	**11.9037**
6th	**5.8**	**3.6331**
7th	**12**	**9.7467**
8th	**11.4**	**16.6523**
9th	**8.6**	**7.9246**
10th	**9.6**	**14.8087**
11th	**3.8**	**5.2153**
12th	**4.2**	**3.0331**
13th	**3.8**	**2.2803**
14th	**4**	**5.7879**
15th	**3.4**	**3.4351**

From the fifteen cases represented in the table, fourteen instances have standard deviation higher than two, which shows that the traffic of the cell tower on the same day of different months does not correlate with each other. The same conclusion can be made based on the presented plots of other cell towers.

At this point the investigation on the exact number of the day in the months is closed, but the days of the month can be addressed in a different manner. As explained before correlation between inner month's traffic data has been proven and the more detailed unit of date-time measurement was day, but the exact day of the date did not correlate between consecutive months, alternatively another form of a day in a month is "week days". Based on the presented analysis every 1st day of the month won't correlate with each other but the same analysis can be performed for days of the week. The last comparison clearly showed that the comparison should not be one-to-one and identical, so as an example comparing every Monday of each month will also have the non-correlating results as we already have. Instead the "types of the week" are the target. Basically every day of the week can be split into two major groups, "weekdays" and "weekends". This segregation has a better situation conceptually, meaning that even considering it outside of the scope of this study, weekdays of a month are much more correlated and identical rather than a particular day of the month compared to the same day in the next month. As long as the traffic behavior of the weekends and weekdays are distinct enough, weekday type can potentially be the next parameter of the data reformatting.

The following plot is a box plot designed to compare the amount of traffic gained by the cell towers on weekdays with the amount of traffic gained on the weekends of the month February. This analysis has been done solely for February due to the proven fact that month's traffic related behavior tends to change over time, therefore gathering multiple months of data together for analysis will mislead the final results.

The conclusion from observing the following plot will either lead to the conclusion that weekdays traffic and weekends traffic for each cell are distinct enough and at the same time concentrated enough in their own group, which means weekdays and weekends are fitted to be used as the next parameter in the data reformatting, or it disproof's this hypothesis. This plot is grouped by colors and each color represents a cell tower. Each group contains two boxes indicating the weekdays data (on the right) and the weekends data (On the left).

By observing the results from the plot, a common property can be seen among almost all of the cell's weekdays and weekends data. Every group of plots except for the group belonging to cell ID:42740 seems to be concentrated enough around the dataset's mean and the outliers are not out expectation, and also at the same time the concentration of the data between two boxes in the same group seems to be apart from each other indicating the different behavior between weekdays and weekends load of traffics in almost every cell tower presented in the plot. This majority in the results shows that the weekday types are proven to be suited as a parameter for data reformatting.

From each analysis there are lessons to be learnt that can be used in the next analysis, from the weekday types analysis the fact that looking into individual days in a month will end up with random results without any patterns that show they are related to each other, and also the reformatted data would be too sparse for event prediction, this is why grouping the days of the

month in only two major groups will provide better and more obvious patterns and will be helpful in data reformatting.

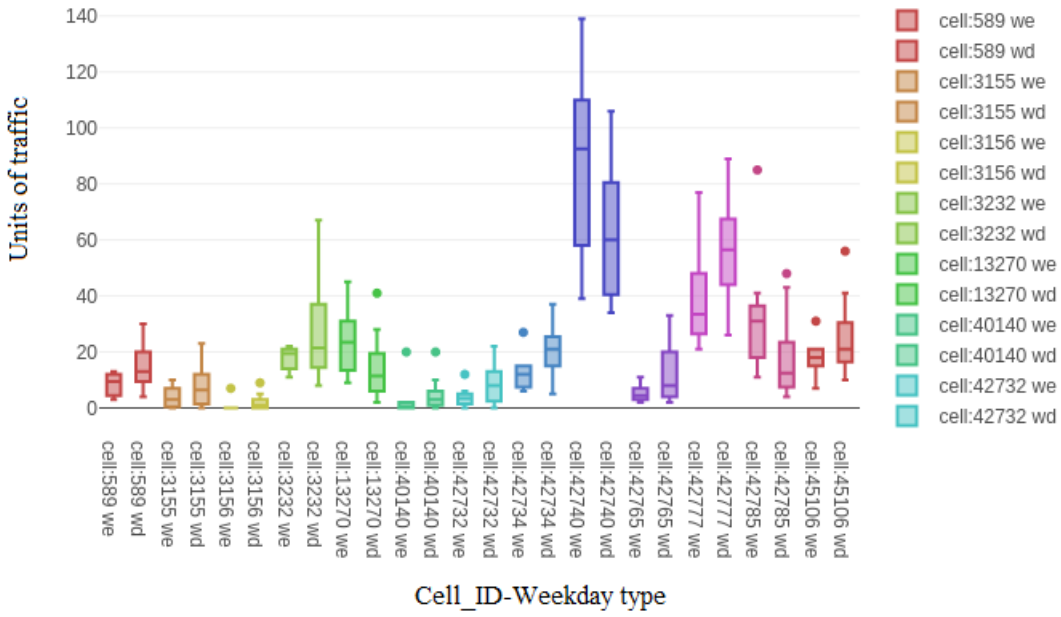


Figure 9. Box plot comparison between weekdays and weekends behavior among sample cell towers.

As it was done for the “days” section the same scenario can be done for the next parameter. The procedure is the same, the next parameter will be chosen from a more detailed unit of date-time measurement compared to “days”, which is “hours”. With the same logic and lessons learnt from the “days” section for the “hours” the individual 24 hours of the day won’t be used as a separate group due to the lack of trackable patterns in different days at the same hour and also depending on the cell tower’s location there will be many hours without any CDR data on a cell site which will end up with redundant zero traffic in the final reformatted data. The hours of the day must be grouped into a certain number of segments that would suit the final goal the best. If the number of segments are Too high it will bring back the redundancy issue again, at the same if the number of segments are time too low it will generalize the data in an insufficient manner that trackable patterns will vanish.

There is no scientific method to determine the best number of segments for the 24 hours of a day, at the end each number of segments will have different mean and different number of CDR data assigned to them. The goal for hours’ segmentation is to convert every day’s traffic load hourly based so that the predictions could be made based on hours of the day. This parameter needs to be as close to people’s behavior in the day as possible, and since no scientific analysis can help

this matter, therefore the conventional hour's segregation used by people will be used in this study. The 24 hours of a day are commonly divided in 4 main segments, "Mid-night" which is from 00:00 to 5:00, "Morning" which is from 6:00 to 11:00, "after-noon" which is from 12:00 to 17:00 and "night" which is from 18:00 to 23:00. The origin of this separation comes from the different sections of the day in daily people's activity which makes this decision perfect for the current study.

All of the parameters needed for the data reformatting is now provided. The potential parameters for data reformatting is narrowed down to the following parameters:

- Month number: the numeric representation of the CDR record's month.
- Week day type: Indicates whether the day of the CDR record is during a week or on the weekend.
- Hour segment: Determines which segment of the day the CDR record's time belongs to.

As an example the following table contains samples of raw CDR records and their reformatted form.

Table 2. Sample of raw CDR data conversion to reformatted CDR data.

Original Raw CDR records			**Reformatted CDR records**			
Cell ID	**Pos User ID**	**Date**	**Cell ID**	**Month Number**	**Day of the week**	**Hour segment**
589	**27146155**	**2018-01-01 00:26:30+02**	**589**	**1**	**WD**	**1**
589	**27146155**	**2018-01-01 01:01:14+02**	**589**	**1**	**WD**	**1**
589	**2178915**	**2018-01-03 01:10:12+02**	**589**	**1**	**WD**	**1**
589	**15599147**	**2018-01-04 01:17:29+02**	**589**	**1**	**WD**	**1**
589	**29086356**	**2018-01-04 14:19:45+02**	**589**	**1**	**WD**	**3**
589	**11786585**	**2018-01-06 15:43:23+02**	**589**	**1**	**WE**	**3**
589	**11786585**	**2018-01-06 15:44:46+02**	**589**	**1**	**WE**	**3**
589	**11786585**	**2018-01-09 20:59:20+02**	**589**	**1**	**WD**	**4**

4.2. Live Data

The main objective of this study is to detect ongoing events utilizing analytical approaches on CDR data. The CDR data provided for this project must be imported to the application in order for the analytical sections to process the data and determine the possibility of an event in the data. The major challenge before the implementation of analytical modules is the data injection. The live data contains the newly generated .csv files gathered from more than 8000 cells. Two of the most challenging properties of the live data are:

- Continuity: The cells are active as the application progresses and new CDR data are being stored every second. The continuity of this data requires a stream like method in order to inject the input data into the application.
- Size: Considering the number of cells and number of mobile network operator subscribers making phone calls, the size of the input CDR data gathered from all of the cells will become one of the main challenges.

By implementing the correct method for injecting the data into the application both of the major issues can be fixed at once. The CDR records can't be streamed into the application and the input data must be in the form chunks of information due to the fact that solo records represent a single CDR in a specific time and date, but in order to analyses the live data for potential events a portion of time must be processed together. The main issue at this stage is choosing the correct portion of time, because this portion of time affects multiple parameters related to the detection of events directly. The time portion size has a direct connection with the input size, as the portion of time increases the amount of data inputted to the application will increase as well which directly affects process time and uploading time. Aside from the input data size, response time is another important factor. One of the main goals of the detection of possible ongoing events in the area is to provide connection support for the event if needed, therefore if the time portion is too large the window for the response time would be limited and the system will fail to serve its purpose.

Based on the comparison above, a middle ground in the time units can be chosen as the size of the blocks of data injected into the system. The most convenient and easier to process with an acceptable response time is the unit of one hour. This means that the data will be uploaded to the system for further processing every hour. As demonstrated in the chart, data size for one hour is manageable for quick upload to the system and light enough for short processing time with quick access to the final results. The uploaded CDR data will be aggregated in each hour, which at the end determines the number of CDR records in each hour (also known as traffic) for every particular cell tower. For instance, if the inputted raw CDR file contained 23 records from 13:00 to 14:00 for cell tower with the ID of 538, the data will be grouped in hourly basis so at the end of the input unit process the historic data shows that cell tower with the ID of 538 had 23 units of traffic assigned to the time 13:00.

4.3. Machine Learning traffic prediction (Failed Approach)

The initial attempt for solving the event detection problem and reason behind its failure is demonstrated in this chapter.

In this approach the reformatted historic data will be used for predicting the volume of traffic per each cell tower for the live data which is taking place one year after the historic data. As explained in the previous section the amount of traffic gained from each cell tower in the live data will be aggregated in each hour and in the result each particular cell will have a specific amount of traffic assigned to every hour. Two main data resources (live data and historic data) will be used at once for the event prediction. Essentially for any machine learning approach the model will need a train set in order to learn the dataset and use the patterns to predict the outcomes to actual data. In this study the train set is the historic data in the reformatted form, which means all of the features provided in the reformatted form of the historic data will be used in the machine learning model for training. After the model's training is done regardless of what machine learning model will be used, in order for the model to be able to predict the amount of traffic in live data, the live data's format needs to match historic data. The nature of historic data and the live data are the same, the only difference is the time which the data was recorded, therefore the same procedure used in historic data to convert "date" field to the reformatted fields ("Month number", "Week day type", "hour segment") will be applied to live data as well.

Due to the size of historic data normal data analysis tools for python programming language such as Pandas and Numpy won't be as efficient, so more advanced and specific tools for big data manipulation such as PySpark will be used for the implementation of machine learning approach. For every machine learning model the training data set needs to be split into two major sections called "Train set" and "Test set". Each set of data contains features and a label, features are the columns of data used for model training and the label is the goal of the prediction. The model will use the train data set to learn and develop and use the test data set for evaluation and assessment of the model based on the predictions made for the test data set. The input features for the machine learning model are the main columns in the reformatted historic data (cell_ID, month_number, week_day, hour_segment) and the label will be "traffic" value. Different machine learning methods will have different outcomes depending on the data. The main methods for testing in this project are "Linear Regression" and "Neural Networks" which are two different categories. Linear regression is being used in the first stage due to the unlimited number of possible values for traffic value, and for further investigation on the underlying levels of the data endeavors for higher accuracy compared to linear regression Neural networks will be implemented.

The main features used in the linear regression are the main characteristics of the historic data ("cell_id", "month_number", "week_day", "hour_segment") and label is the amount of traffic. In order to create the two main data sets needed for the regression technique (Train data and Test data) the historic CDR data needs to be split into two datasets. The first 75% of data will become the train dataset and the remaining 25% will be used as test dataset. The reason behind this specified percentages between train dataset and test dataset is that train dataset will be used in the learning process of the machine learning model but the only purpose of test dataset is to evaluate the model, and the training and improvement of the model demands large amount of data. The amount of data used as train data set has a direct impact on the enhancement of the model and its quality since train data set essentially trains the model for the variety of different combinations among the features and how they affect the label. As explained prior to this section the regression model does not predict specific fixed values and the output value of the prediction has unlimited possibilities, therefore for the evaluation and the assessment of the accuracy an error margin will be used. As the assessment of the accuracy, the prediction on the test data will be

compared with the actual labels of the test data, but including the error margin, the situation will be different in a way that, if the predicted value and actual label's absolute difference was smaller than the error margin, the prediction will be considered correct.

Due to the fact that data is being processed under PySpark framework, the primary strategy is to use PySpark library for the initialization of the linear regression model, and for further improvements in the results performing analytical enhancement on the model and finally eventually transforming to Neural Networks.

The results of the first regression model's prediction changed the path for the event detection application development. The following table contains the results from a particular data frame which stored the results from comparing the actual traffic in the test data and the predicted traffic from the trained regression model.

Table 3. Accuracy results from the linear regression approach on reformatted CDR data.

Summary	**Absolute differences**
Count	**3251**
Mean	**72.815749**
Standard deviation	**41.06109**
Min	**17**
Max	**140**

The content of this table indicates that there are 3251 instances of the test cases in the test data, the average of the absolute differences between actual traffic and predicted traffic from the model including 10 units of traffic as error margin (meaning that the prediction is allowed to be 10 units of traffic off from the actual test data) is 72.815749. The minimum difference is 17 (if the error margin gets excluded it means the minimum difference was in fact 27) and the maximum difference is 140 units of traffic. The behavior of the historic data was demonstrated in the beginning of this section in a variety of charts. By taking the numbers and observations on the historic data the norm of the traffic data in different hours of the day can be estimated. By comparing the minimum difference indicated in the table above with the norm of traffic values in the historic data, it can be concluded that the results from the regression model are excessively inaccurate. This results proved that the linear regression at its basic core resulted in a prediction where the most accurate prediction was 27 traffic units apart from the actual traffic in the test data. This issue raised questions and doubt regarding the current method at the time of development and led to a temporary pause in the next stages of the development. As explained in the previous sections at the next stages the regression model would be enhanced and at the end converted to a neural network, but due to the initial results and unexpected inaccuracy in a vastly

large margin, a series of investigations and consultants began. The inaccuracy is expected in the beginning phase of many machine learning solutions and by changing the properties of the model the accuracy can be improved but in the current situation the inaccuracy was far outside of the foreseen inaccuracy scope. Transformation to neural networks won't be a solution for this problem because neural networks require large amounts of datasets in order to be able to function properly and currently the historical dataset has limited outcomes due to the data reformatting. The latest accuracy results and the consultations of data specialists in this area of study suggested that the current method won't be useful for this application and essentially traffic prediction based on previous year's activity is not a viable solution.

4.4. Statistical approach

Based on the previous observations on the machine learning method, the results indicated that historic reformatted data is limited and relatively pattern less based on cell tower's location. In this situation the best alternate path is statistical approaches which are well suited for limited data with hard to detect patterns. At the time of the development of this application incidents such as covid-19 outbreak revealed the fact that historic data cannot be used as a sole source of data for event detection and traffic prediction, because situations like outbreak quarantines affect people's life's routine directly and render the historic data gathered from that year obsolete.

The changes in the development will lead to a system based on statistical approach with direct involvement of both historic data and live hourly based uploaded data sets, meaning that in the core process of this application not only historic data but also live data sets will be used as well. The following graph will demonstrate a brief hierarchy of the units and in the system and their functionalities.

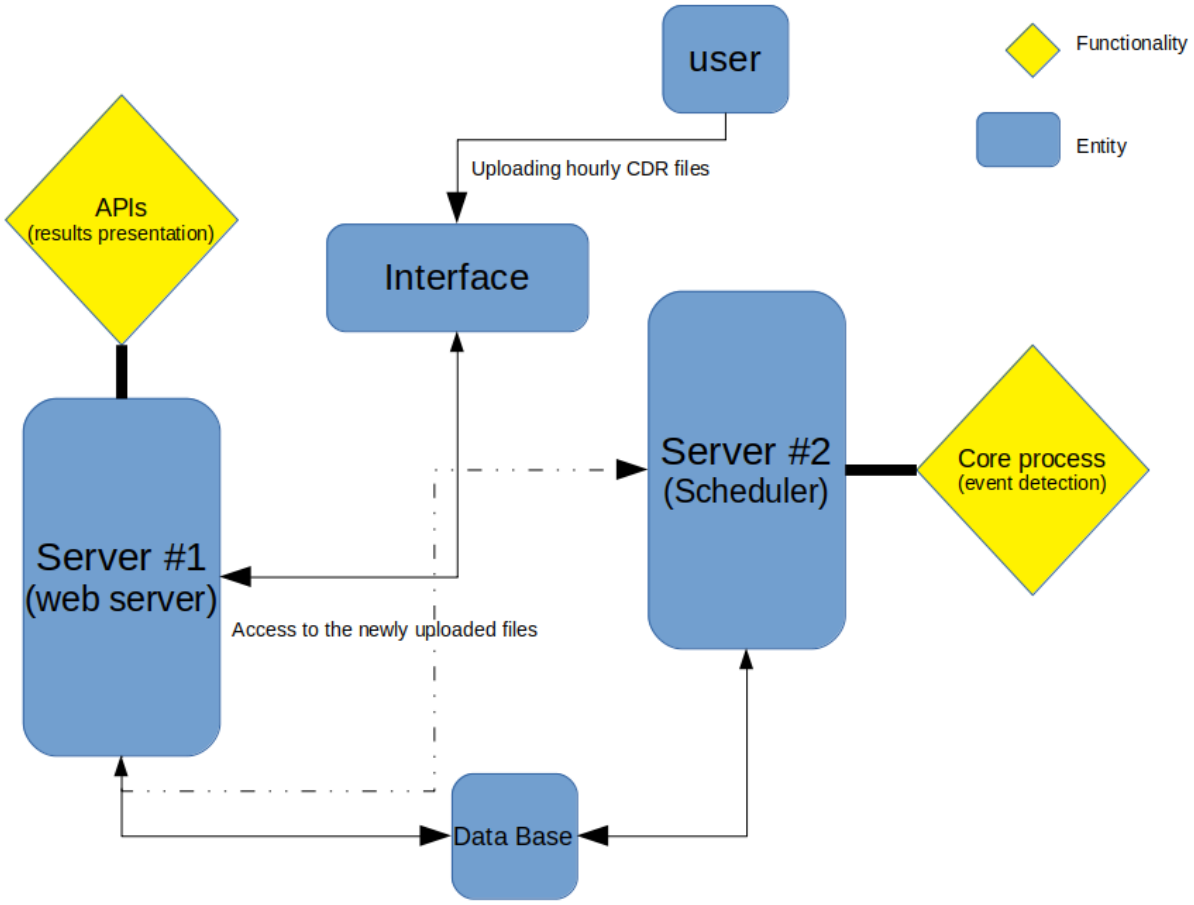


Figure 10. Schematic review of the main architecture for the system.

The major change in the flow of the development affected the primary sectors of the application such as the webserver. This application is aimed to be hosted as a django web server deployed at a local server located in the Positium company's office. The previous approach did not require access to live hourly data in the main process unit, due to the fact that historic data was already generated by the machine learning model beforehand. The system design demands different architecture considering that live data needs to be processed for event detection along with the historic data. Based on the represented graph the main two sectors of the application are the parallel servers running alongside with different responsibilities.

4.4.1. Scheduler Server

The core and primary sector where the event detection statistical algorithm is located is in the scheduler server. This unit is a server-like entity with a time based scheduler. This server checks the primary directory of the application on the server every 15 seconds for new CDR .csv files. When the new hourly file gets uploaded to the server it will be transferred to the primary

directory of this application and triggers the scheduler. The scheduler performs two main tasks sequentially called data preprocess and event detection procedure.

4.4.1.1. Data preprocess and rearrangement

As the initial steps in data preprocessing input data format needs to be converted to hourly based traffic counters. Input raw CDR data as mentioned before contains cell tower ID, records date and time and user ID which is unique for every subscriber to the network. The new data format integrates the data in hours of the date and time fields creating aggregated results for every hour, counting the amount of traffic in each hour for every cell tower, but prior to this step the input data as in the original state needs to be cleansed from possible noises since after the conversion to the new format the news and false data will be aggregated with the rest of the dataset and they will become indistinguishable. CDR data contains users calls or text messages records and the data is directly connected to people's activity which increases the chance of noise in the data, but the nature of this project implies the fact that noises regarding the volume of data is the point of interest in the preprocessing section. On certain occasions there are subscribers in the operating network who use this network connection for business or possibly advertisement purposes, and their CDR records will change the conclusion for certain cell towers. The cell tower's traffic will be affected by this situation and will trigger false events due to fabricated large peaks in the traffic. In order to address this problem users activities in the CDR data will be limited and CD data assigned to each user will be monitored in order to detect abnormal numbers of records for a particular user in the input data. Data assigned to the identified user will be removed from the dataset to normalize the dataset and increases the focus of the dataset on ordinary usage of the network.

This application is capable of handling multiple data uploads as well, for instance the user can also distribute the hourly based files into multiple files and upload them altogether to the server. The essential data form for event detection algorithm is time series and the constitutional property of a time series data is its continuity with No vacancies. In order to generate time series data from the raw CDR data, the gaps in the data needs to be filled in two situations, potential gaps in each file and potential gaps between files. As an example files "a" and "b" are uploaded simultaneously on the server for a particular cell tower and in this example the hourly files are being uploaded every two hours due to the light traffic on the cell towers. The CDR data in file "a" starts from 13:00 and continues until 15:00, and there are only two records in this file for 13:40 and 15:34. This means the data has one record at 13:00 hour and one record at 15:00 and in the time series generated from this file "13:00" and "15:00" will have one unit of traffic but since there is no mentioning of "14:00" this particular time needs to be filled with zero unit of traffic in the time series. The same scenario can happen between the files where the ending of the first file and beginning of the next file are not in consecutive hours. The gap filling procedure will conclude the data transformation and noise reduction and in results the input data is a fully formed and concise time series. The generated time series will be stored in the database for other units and the original files in the primary directory will be deleted automatically after the process.

Table 4. Comparison between the original input data and the aggregated, preprocessed input data. .

Raw input data			**Preprocessed time series**		
Cell ID	**User ID**	**Date**	**Cell ID**	**Date**	**Traffic**
590	**258079**	**2018-01-01 00:14:50+02**	**590**	**2018-01-01 00:00:00+02**	**1**
590	**258079**	**2018-01-02 14:18:50+02**	**590**	**2018-01-02 14:00:00+02**	**3**
590	**258079**	**2018-01-02 14:50:46+02**			
590	**258079**	**2018-01-02 14:51:06+02**			
590	**258079**	**2018-01-02 15:58:12+02**	**590**	**2018-01-03 08:00:00+02**	**2**
590	**258079**	**2018-01-02 15:58:12+02**			
590	**131988**	**2018-01-02 17:25:26+02**	**590**	**2018-01-03 17:00:00+02**	**2**
590	**131988**	**2018-01-02 17:25:32+02**			
590	**131988**	**2018-01-03 08:58:36+02**	**590**	**2018-01-03 08:00:00+02**	**1**

4.4.1.2. Event detection

Event detection unit is the second half of the scheduler server containing the main method for event detection using the time series generated in the previous section. As explained previously the machine learning approach did not succeed in the system and the historic data need to be used in a statistical method along with the live data. There are three main statistical methods for event detection, also named as “indicator” in this document which use live data and historical data, the conclusion of the three indicators decides whether an event has been detected or not [22]. The fundamental concepts behind each of the indicators are the same and the only

difference is the datasets. Indicators use the statistical concept of standard deviation in order to determine if the test value is in the ordinary range of traffic or an outlier. Each individual indicator gathers a dataset from a particular date and times related to time series data stored to the database in the previous section (will be known as “the test record”). The retrieved dataset contains multiple record samples from live data and historic data depending on the indicator, the statistical formula will be applied on the dataset in order for the results to be compared with the test record. The formula used in all of the indicators calculates the distance of the live test record’s traffic with the average amount of traffic in the retrieved query of data records, if the distance is higher than standard deviation, the test record is an outlier and determined to be an abnormality and as a result the corresponding indicator will flag.

$$|\mu - Y_i| \geq \sqrt{\frac{\sum_{n=1}^{N}\left(X_n - \mu\right)^2}{N}} \times sensitivity \qquad (1)$$

In the stated formula (1) **μ** is the mean of the retrieved data records, $\mathbf{Y_i}$ is the live data record instance, $\mathbf{X_n}$ is individual data from the retrieved data records, **N** is the number of retrieved data records and **sensitivity** is a constant.

The event detection application is designed to be applicable in various environments with different aspects and different characteristics, some cities have particular sub neighborhoods with vastly different population density, which leads to various different results from cell towers in the same city. The **sensitivity** constant is designed to make the formula and the results from the formula more dynamic and adaptable with the environment, so that the application can detect light abnormalities along with denser and distinct events. The minimum value for **sensitivity** is 1.1 where in city centers tend to detect many false positive cases of event detection due to their high cell tower traffics and the maximum value is 5.0 where minor abnormalities will be overlooked in order for the major events to be detected distinctively.

4.4.1.3. Indicator I (Historic data)

The first indicator uses the previously generated and reformatted historic data. As discussed previously the aggregation of the data based on the new converted columns (month_number, week_day, hour_segment) will create a series of numbers demonstrating the amount of traffics for the aggregated section, in the machine learning section the average amount of traffic has been

utilized but for the current method instead of solely using the average of the traffics the individual values will be used in the formula above in order to calculate the average and standard deviation. Essentially the queried data records explained in the formula for the “indicator I” are the corresponding aggregated data where the reformatted date and time columns match the current live data record. If the condition in the inequality formula is met, the “indicator 1” will conclude that the live data for this particular date and hour does not match with the historic data and it is abnormally higher which demonstrates potential events from the aspect of historic data.

Table 5. Demonstration of the input data for indicator 1.

Aggregated Historic data						**Live hourly aggregated data**		
Cell ID	**Month Number**	**Week Day**	**Hour Segment**	**Mean**	**Standard deviation**	**Cell ID**	**Date-Time**	**Traffic**
269	**1**	**w_d**	**3**	**6.8695**	**0.7034**	**269**	**2018-01-01-05 13:00:00**	**4**
269	**1**	**w_d**	**4**	**16.1449**	**1.6082**	**269**	**2018-01-01-06 20:00:00**	**19**
269	**1**	**w_e**	**1**	**3.1666**	**0.3450**	**269**	**2018-01-01-06 01:00:00**	**1**

4.4.1.4. Indicator II (live historic data)

Indicator number 2 uses the uploaded live data as a resource. By uploading the live CDR data for days, the system will have worth of days’ time series based data indicating the amount of traffic in the past few days. There are particular days with the same characteristics and communication flow through the cellular towers compared to current “day” and “time”, such as one day ago, last week at the same day and same time, two weeks ago at the same day and same time. The second indicator uses this concept and gathers data from these dates and times forming a data set of hourly traffics in different days but at the same hour. The average of traffic and standard deviation derived from these sets will be used in the formula in order to determine whether the current traffic is out of norm compared to the gathered data set or the cell tower’s traffic is reasonable. The following figure is a schematic demonstration of the chosen days to obtain the traffics from.

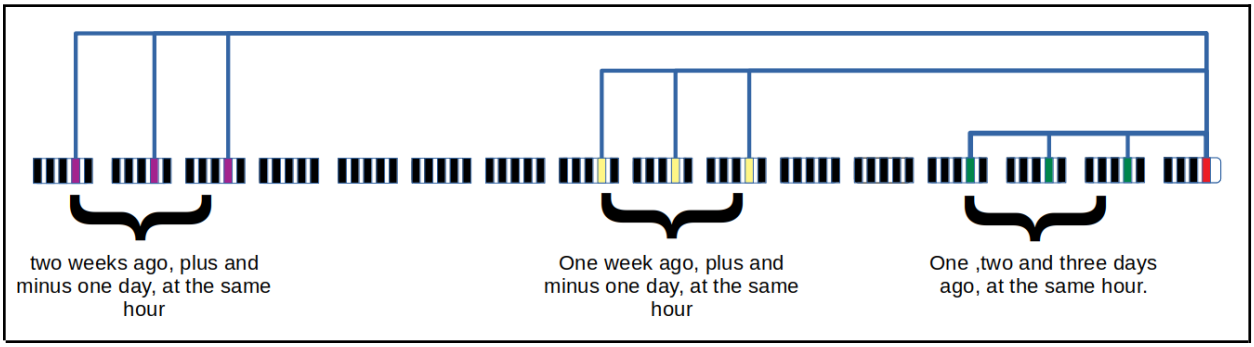


Figure 11. A schematic overview of the selected hours to obtain the traffic from in the live data for indicator 2.

In the represented schematic each black bar is a symbol for an hour of the day, each block represents one day, and the red bar indicates the current time of the latest uploaded CDR file and the bars with other colors demonstrate the hours of previous days chosen to be a part of the data set gathered for the second indicator. As demonstrated in the schema all of the chosen hours have the same daily characteristics of the current day such as, one day before, one week before, two weeks before. The behavior similarity is the main concept aimed in the second indicator in order to track down any abnormality in the current hour's traffic.

4.4.1.5. Indicator III (live time series data)

Same as the previous indicator, the source of the third indicator is the stream of uploaded live CDR data. In contrast with the second indicator, the third indicator does not use particular days with similar daily behavior at the same hour. The third indicator concentrates on the continuity of data to the current hour of the live data, in order to determine if the current hour's traffic is in the correct range with allowable deviation or the current hour's traffic is an outlier compared to the continuous stream of uploaded CDR data in the past 20 hours. The following schematic determines the chosen hour's the third indicator in order to gather the traffics from, and the comparison of this schema with the schematic provided for the second indicator exhibits the difference more precisely.

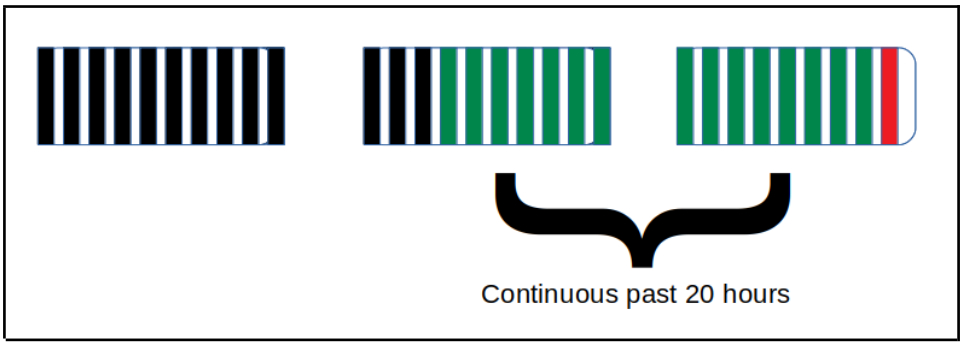


Figure 12. A schematic overview of the selected hours to obtain the traffic from in the live data for indicator 3.

In the presented schema same as the previous schema each block represents a day, each black bar is a symbol of an hour, the red bar represents the current hour of the live hourly CDR data and the green bars are the chosen hours in the third indicator to retrieve the aggregated traffics from. The individual derived data along with their average and standard deviation will be used in the formula to validate the current hour's aggregated traffic.

4.4.1.6. Indicators Combination

The combination of the flags from all of the three indicators will determine whether the event detection is correct or a false positive, for instance there might be situations where the first indicator flags the record as a potential event while the other two did not detect any event in the corresponding hour, the conclusion of this situation is that no event should be reported for the corresponding hour. Along with the three flags from the indicators there must be at least two positive flags in order for the current hour's traffic to be announced as an ongoing event. The logic behind the "two out of three" concept in the indicators results came from the differences in the nature of gathered data in each individual indicator. The times restricted in each indicator have familiar characteristics to the current time and date yet different from each other in that sense there are possibilities where a noise or unexplainable peak in the data compromises the results from one indicator, but the fundamental difference between the indicators guarantees that any possible destructive noise in one of the indicators won't affect others. The situation with the covid-19 outbreak proves the benefit of the "two out of three" rule. If this system is meant to be used in 2021 the historic data will be from 2020, but the historic data from January, February and March are all obsolete because people were working in quarantine and countries were at the state of emergency, so event detection system will use second and third indicators in the first three months of 2021.

Based on the results from the indicator combination process the exact **time and date** and the **coordinates of the cell** detected to be involved in an ongoing event will be stored in the database automatically. This procedure will trigger every time a new file is uploaded automatically and takes place in the second server (scheduler) because if this process was under user demand in the API section located on the first server (web server) the server load would block the server's

process and the waiting time will be unacceptable by the user. The current method of implementation will have no waiting time for the user in the event detection unit, and the latest provided events from the scheduler server will be available for the user instantly.

4.4.2. Interface

The interface is a simple and easy to use user interface designed for the user in order to perform number of tasks such as the following:

- **Upload panel**: this panel is used in order to upload hourly extracted csv CDR data on to the server.

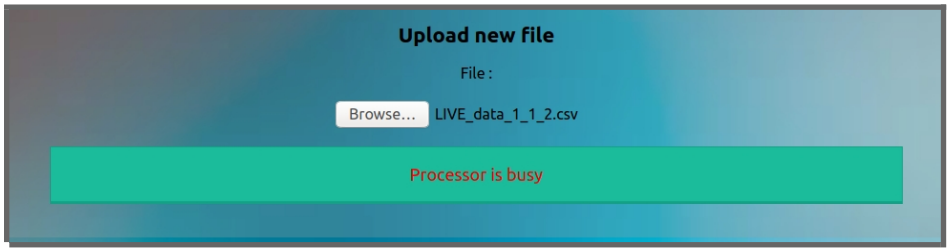


Figure 14. Upload panel for user interface.

- **Solo Cell tower status check**: this panel is used in order to represent visual and statistical information about a specific cell tower's traffic. The source of data represented and analyzed in this section is the results of the first section of the scheduler server.

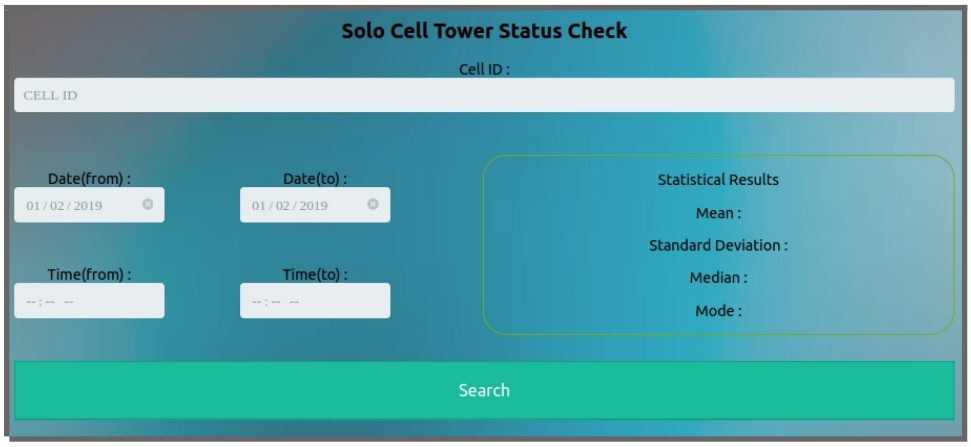


Figure 15. Cell tower traffic analyzis and visualization panel for user interface.

- **Event detection**: this panel is used in order to monitor particular date and time for potential events, and receive visual partitioning of the location on the map.

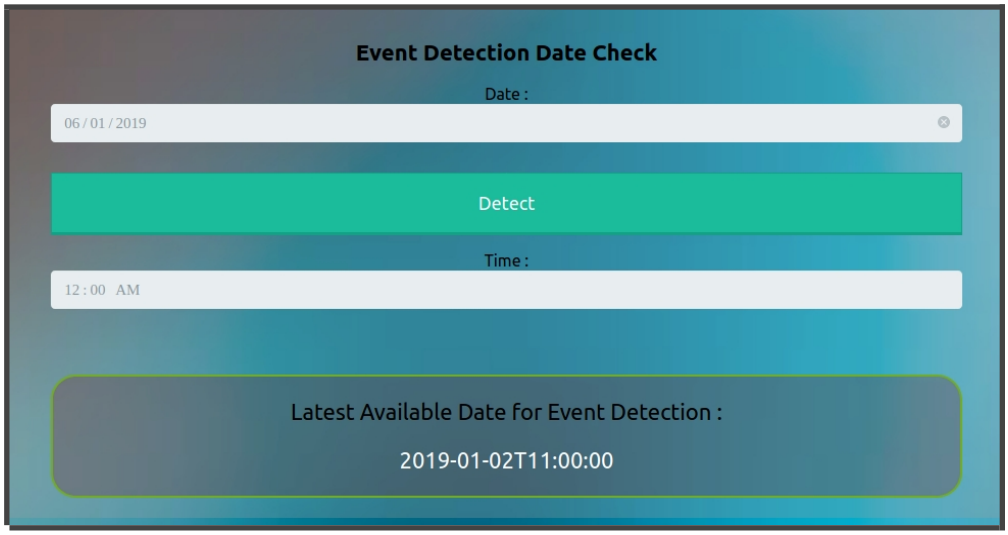


Figure 16. Event detection panel for user interface.

- **Malfunctioning cells report**: this panel requests a general check on the cell tower's activities and detects potential malfunctioning on cell towers that went offline after a specific date or cell towers which have particularly low traffic compared to the other cell towers in the neighboring geographical area in less than 2 kilometers radius.

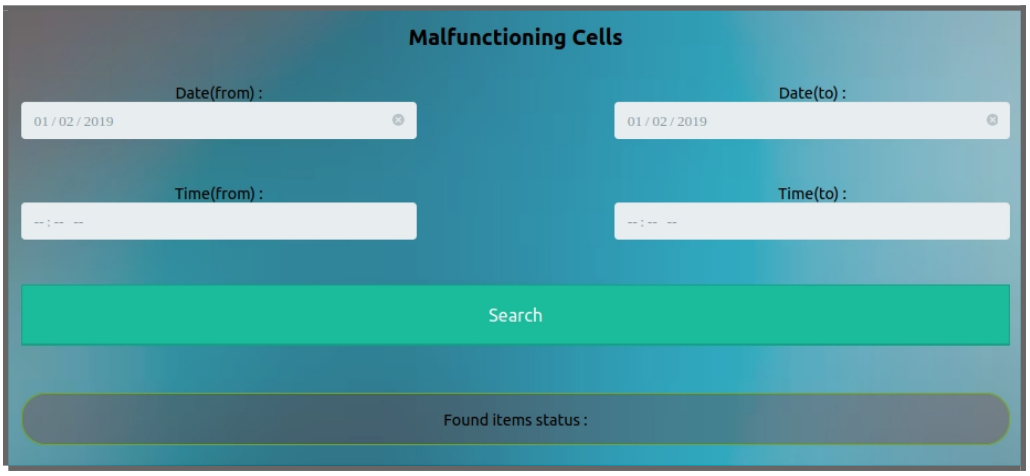


Figure 17. Malfunctioning cells panel for user interface.

- **Sensitivity settings**: this panel is used to change the sensitivity level of the event detection algorithm for a stated period of time and future CDR data.

Figure 18. Sensitivity settings panel for user interface.

4.4.3. Web Server

Each individual panel is connected to an API implemented on the web server, providing the information demanded by the user. In this section the focus will be more on “Event Detection Date Check” panel and “Solo Cell Tower Status Check” due to the fact that other panel’s functionalities are self-explanatory and demonstrated in the “Interface” section.

4.4.3.1. Solo Cell Tower Status Check

The API call designated for this panel accepts a range of date and time and a particular cell ID, in the response an overall statistical overview containing mean, standard deviation, mode and median will be represented along with a visual chart of the traffics in the indicated period of time and the location of the cell tower on the map. This panel is very useful in monitoring the suspicious cell towers reported in other panels such as “malfunctioning cell towers” or “event detection” panels.

Figure 19. Statistical results in the solo cell tower check panel.

Figure 20. Visual representation of the location of the cell tower on the map in the solo cell tower check panel.

4.4.3.2. Event Detection Date Check

This panel accepts particular date and time (detailed to hour) in order to determine if there are any ongoing events. As explained previously in the "scheduler server" section the results of the potential ongoing events in the city are stored in the database every time a new file is added to the system. The exact date and time and the coordinates of the cells detected in an ongoing event will be extracted from the results and filtered according to the input data in the panel. The responded data from the server include a list of cell tower coordinates reported to be included in an event in the stated time and date. This application can be applied in a wide range of cities or even country based, the issue that would rise with this information derived from the event check API is that the points can be far apart and without any previous knowledge about the city it would be particularly difficult to group certain points together as a distinct event area. Theoretically the reported cells close to each other in the same neighborhood should be grouped together as one event. Algorithms such as K-means clustering are designed for such purposes but there is a major downside with the K-means algorithm in the current system. The common procedure for K-means algorithm requires either the number of entities in each cluster or the total number of clusters that the entities are going to be distributed among, while in the current status of the results retrieved from the database the number of clusters (representing the number of events) and number of entities in each cluster (representing the number of cell towers in each event) are not clear and yet to be determined. Scree test [23][24] is a test for determining the number of factors to utilize in a factor analysis. There are situations in fields of studies where the number of factors included in an analytical process is not determined and final, and different number of factors involved, will have different effects on the final results. The results from the process needs to be balanced regarding the number of factors used in order to avoid overfitting or under fitting the process. In the current situation the number of clusters is the missing factor and the evaluation is based on the distance of cell towers included in a cluster from each other. The following pseudocode demonstrates the functionality and the procedure of Scree testing to find the optimal number of clusters for implementation of K-means algorithm. The numbers vary from 1 to the count of all the reported cells from the system from the event detection section.

Algorithm 1: cell tower coordinates clustering.

Input: event_cells_coordinates(x,y) coordinates of candidate event invovled cell towers

Results: Nested list of cell tower coordinates in clusters

```
for n = 1,...,length(event_cells_coordinates)
    SET model=KMeans(event_cells_coordinates,n)
    SET clusters=model.predict
    for cluster in clusters
        SET cluster_center = k_means_center_coordinate(cluster)
        for coordinate in cluster
            SET distance = Distance(coordinate,cluster_center)
            SET distances[i] = distance, i=0,...,length(cluster)
        SET avg_cluster_distance = Average(distances)
        SET batch_distance[j] = avg_cluster_distance, j=0,...,length(clusters)
        SET batch_cluster[j] = cluster, j=0,...,length(clusters)

    for k=0,...,length(batch_distance)
        IF batch_distance[k] < constant_maximum_cell_distance
            SET final_cluster = batch_cluster[k]
            BREAK

    if final_cluster != null
        BREAK
```

Based on the demonstrated procedure the K-means algorithm will be deployed in every iteration and in the response the reported cell towers will be grouped in a particular number of clusters. In each individual cluster batches, per each cluster included in the clusters batch the average distance of the cell towers from the center of cluster will be calculated. The average of all of the calculated average distances per each cluster in each of the cluster batches will also be calculated and represent each batch. The iteration will be in an order that the low number of batches with high cluster capacities will be processed first. The first batch that outputs an average of distance among all of its clusters smaller than the “constant maximum cell distance” will stop the process. The essence of this process is that as soon as a batch of clusters got identified where the average distance of the cells from their cluster center is relatively small the process will end and that batch of clusters will be announced as the correct number of clusters for the system. The “constant maximum cell distance” is the maximum value that the cell towers are allowed to be away from their cluster centroid, and this value is dynamic due to the fact that this concept is changeable in different environments.

The results of the previous section is a group of clusters each containing a certain number of cell towers representing a potential event. The group of cell towers representing an event will also be shown on the map as a closed polygon, in order for the user to estimate the approximate location of the event easier. The last algorithm used in the event detection API is a geometrical script to connect the cell towers in a cluster together so it would form a closed polygon. The following schematic and pseudocode demonstrates the methodology behind the algorithm and how the

implementation leads to creating a closed polygon. Per each cluster in the batch of clusters the centroid of the cluster will be calculated. By applying the centroid and each of the coordinates X and Y axis in “arctangent” formula the angles between the centroid and each of the cell tower coordinates in the cluster will be revealed, by sorting the set of coordinates by their corresponding angles a sorted list of coordinates will be formed. The final list will be sorted in a manner that by visualizing the coordinates using the sorted list a closed polygon will be formed.

Algorithm 2: Generation of a closed polygon from coordinates in each cluster

Input: A list of coordinates of the cell tower present in each cluster

Results: The list of the coordinates of the cell towers present in each cluster sorted to form a closed polygon in 2D geographical space

```
for coordinates in batch
    x_list.add(coordinates[0])
    y_list.add(coordinates[0])
SET centroid = x_list/length(batch), y_list/length(batch)
SET angles = arctan2(x_list-centroid[0],x_list-centroid[1])
SET unified_data = x_list, y_list, angles
final_sorted_coordinates = sort_data(unified_data,key=unified_data[2])
```

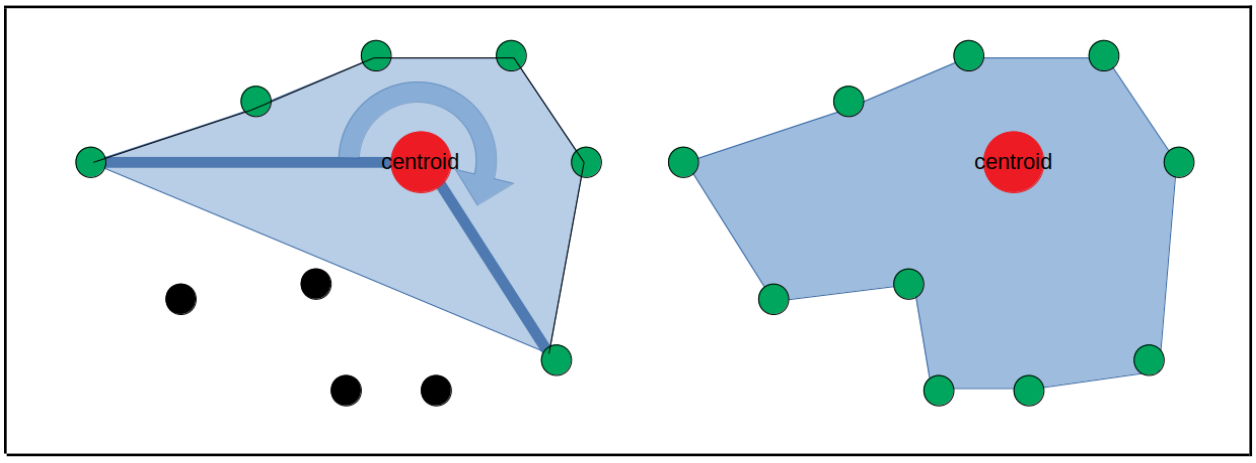


Figure 21. Visual representation of the closed polygon generator algorithm.

5. Experiments

In this section of the document the tests on the implemented method of event detection will be tested and the results of the tests will be reported individually per each test section.

5.1. Synthetic Data Generation

This system requires a large amount of data in order to function properly, and access to real world big data from operators is a difficult task due to privacy policies of the operator companies. This particular issue will be solved by generating synthetic data based on the behavior of the real world data. The behavior of the data has been reviewed in the earlier phases of this document, the information gathered from the real data behavior analysis will aid in the implementation of a dynamic and function script in order to generate synthetic data in vast volume similar to real world data.

The main characteristics detected from the real cell tower data are the effectiveness of the following parameters:

- **Hour segments**: Each segment of the day tends to have distinct behavior regarding the amount of traffic going through the cell towers.
- **Month**: In overall behavior of the cell tower traffic, the change of month and season is an important factor that changes the behavior of cell towers leisurely and the only method to address it is by analyzing monthly traffic data collectively.
- **Week day type**: The different behavior of the weekends calculated traffic compared to other days of the week has been shown in the previous sections.

All of the mentioned parameters will be used in order to finalize the synthetic data as close to real data as possible.

The following schema demonstrates the hierarchy of the processes in the synthetic data generator unit briefly. It is worth noting that for testing the system using the synthetic data the historic data and live data need to be generated from the synthetic data generator unit. The same process will be used for historic data and live data in the testing.

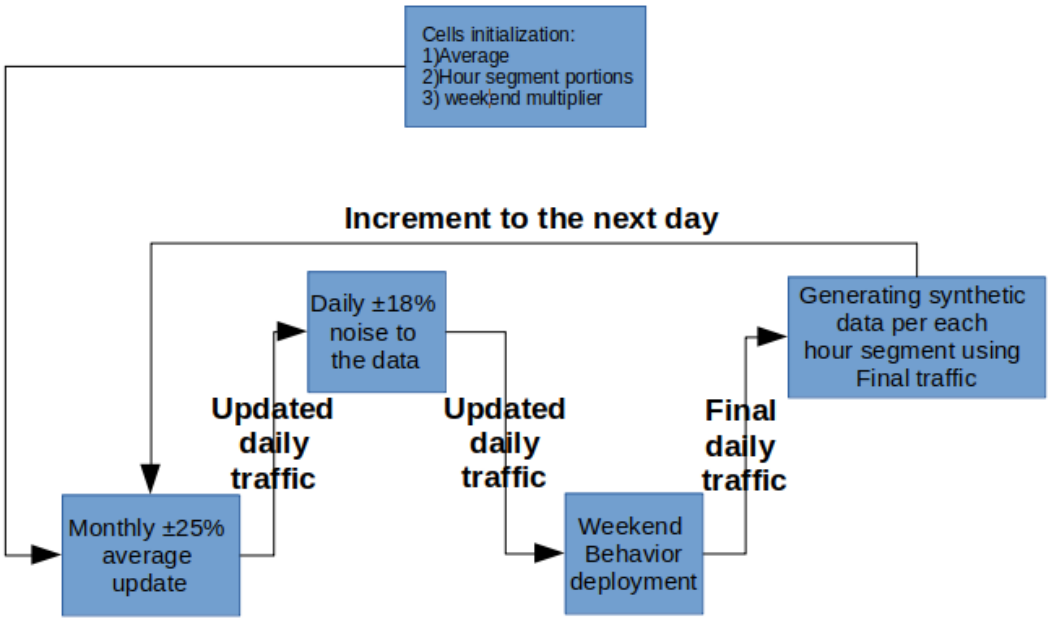


Figure 22. Procedural hierarchy of the synthetic data generator script.

The process begins by initializing synthetic cell tower instances with unique characteristics using the parameters above. Each cell tower will be assigned a daily average amount of traffic in the beginning of the process named as “Average. As explained in the previous sections there are 4 main hour segments of the day, for each cell tower 4 random numbers between 0 and 1 will be generated in a way that the sum of the numbers would be 1, the 4 numbers will be assigned to the individual cell tower as “Hour segment portion”. The behavior of the weekdays and the weekends will be distinct per each cell tower by distinguishing the weekends behavior from the rest of the week by adding a “weekend multiplier” parameter. This parameter can be either 1.6 (the cell tower has higher traffic on the weekends for this cell tower) or 0.7 (the cell tower has lower traffic on the weekends for this cell tower) and in some cases the parameter will remain as 1 (the cell tower’s behavior won’t change in the weekend for this cell tower). The following stages will be repeated per each day in the synthetic data. In the beginning section the timeline will be checked in the event of entering a new month the average characteristic will be updated a random portion between -25% to +25%. After every stage the amount of traffic for that particular date will be updated. In the following stage a daily noise will be added to the data by the amount of a random percentage between +18% and -18% of the total daily average traffic of the cell tower This value will be added to the daily traffic filtered from the previous stage, and if the percentage was negative the value will be deducted from the daily traffic. In the next stage if the current day in the iteration happened to be a weekend the percentage of the weekend stated for each cell tower in the initialization phase will be deployed for the daily traffic as well. After the last stage the daily traffic for the cell tower in the current day will be concluded and in the final section using the “hour segment portion” list of numbers indicated for each cell in the initialization phase the calculated final traffic for the current day will be distributed among the

hour's segments of the day. The following charts visualize two cell tower samples from the synthetic data for the duration of one week.

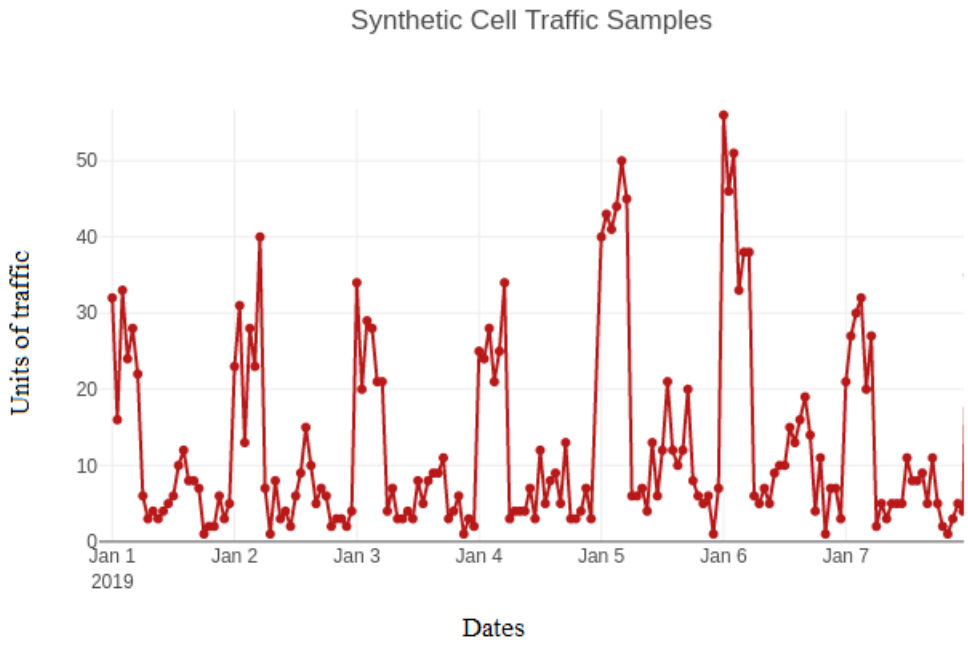


Figure 23. Sample of time series traffic obtained from synthetic data.

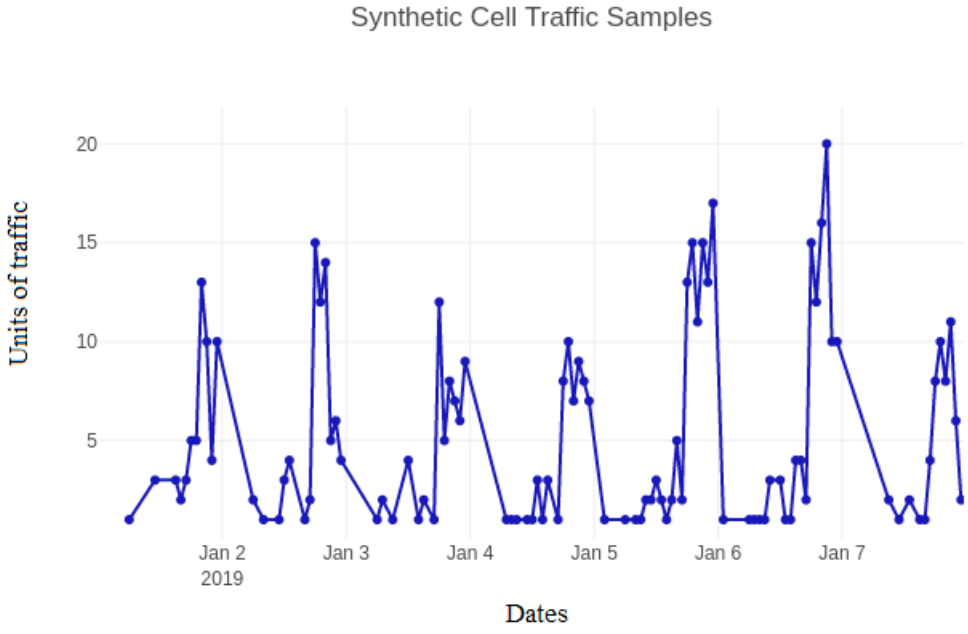


Figure 24. Sample of time series traffic obtained from synthetic data

5.2. System Assessment

Based on the results from the synthetic data and the comparison with the real data, the fact that synthetic data is more well-arranged and logical compared to the live data even though series of noise and traffic randomization is applied to the synthetic data.

This data is meant for testing the system for event detection and its other side functionalities, therefore specific abnormalities must be implanted inside the synthetically generated data regardless of the initialization phase. The implanted abnormality in the traffic will not be chosen by random between the cells, the particular cell towers participating in the events are known and the time of the events are Predetermined in order for the assessments to be easier. The following are a brief list of implanted abnormalities in the dataset and the corresponding cell towers involved in them:

1. January 2nd, at 9:00 including the following cells with the ID of “2035”, “589”, “40975”, “25142”.
2. January 2nd, at 10:00 including the following cells with the ID of “2035”, “589”, “40975”, “1341”, “28660”.
3. January 2nd, at 11:00 including the following cells with the ID of “1341”, “618”, “1998”.
4. January 2nd, at 16:00 including the following cells with the ID of “2436”, “45535”, “2326”, “4479”, “41053”, “6234”, “39993”, “2372”, “2483”, “6713” .
5. January 2nd, at 20:00 including the following cells with the ID of “6561”, “654”, “32320”, “4631”.
6. January 8th, at 9:00 including the following cells with the ID of ”9517”, ”10004”, ”25189”, ”269”, ”47899”, ”39353”, ”2016”,”13547”, ”1204”, ”41350” .
7. January 8th, at 15:00 including the following cells with the ID of “12718”, ”1523”, “40975”, “38226”, “6713”.
8. January 8th, at 16:00 including the following cells with the ID of “12718”, ”1523”, “40975”, “38226”, “6713”, “4112”.
9. January 14th, at 8:00 including the following cells with the ID of “2632”, “301”, “2068”.

As an example for the visualization purposes the following chart demonstrates the abnormality of the cell tower with the ID of “2035” involved in the artificial event on January 2nd, at 9:00. The amount of implanted peak in the data designed to be not too much out of the norm of the flow of traffic in the cell tower and yet not close to the average amount either.

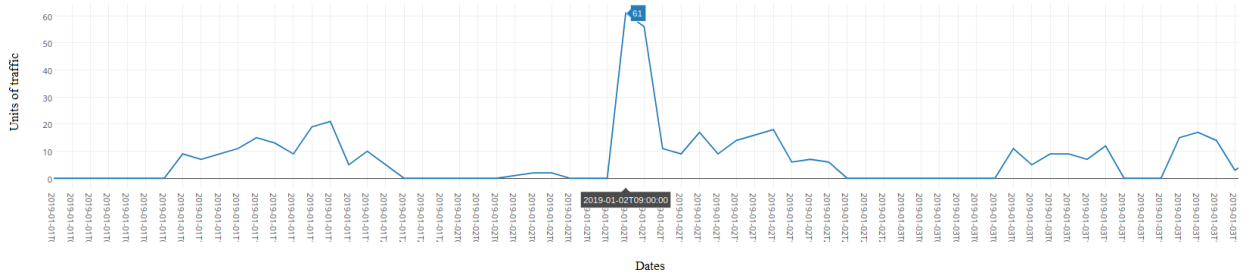


Figure 25. Visual representation of the implanted event in synthetic cell ID: 2035.

The malfunctioning cells report section will also be tested with the following implanted error cells:

- Cell ID: 2192 malfunctioning begins from January 2nd.
- Cell ID: 1293 malfunctioning begins from January 2nd.

The testing procedure is based on the dates of the mentioned implanted errors. The date of the events will be searched after the system is trained using the synthetic historic and live data and the results will be demonstrated on the map, since the error points have been chosen exclusively in the same area the clustering aims for a logical grouping of the cell towers involved in the events. The same procedure will be applied for actual live data as well after the synthetic data tests, but the real data have only one primary test case which is July 18th, 2019 Metallica's concert at Raadi airfield. The represented results on the map are distinguished in three different colors. Cell towers that have been flagged by only one indicator will be shown as light yellow, Cell towers that have been flagged by two indicators will be shown as light red, and cell towers flagged by all three indicators will be shown as bold red dots, but only bold red dots that have the potential to be involved in an event will be included in a polygon estimating the location of the event, the potential is evaluated by scanning for other bold red dots (3-indicator flagged cell tower) in the neighboring area of the current bold red dot.

6. Results and Discussion

In this section the results of the experiments will be demonstrated and discussed regarding the accuracy and system expectations for both synthetic and real data. The success of the system depends on the outcomes of the existing challenges in the datasets, and it also reveals the weaknesses of the system which will create opportunities for future works in order to optimize the system for better.

6.1. Synthetic data assessment

The assessment on the system using the synthetic data has two sections, the primary part of the assessment is dedicated to the event detection panel and the other part is dedicated to the malfunctioning cells reported, which is a side feature of the system. The dates of the implanted abnormalities are available and by testing the system in the mentioned dates and times the results will be shown on the map. The visual results from the map will be represented per each implanted abnormalities:

- January 2nd, at 9:00:

  The first group of test cases belong to the second of January, 2019, which means the second indicator will not be fully active for testing the data due to the lack of input data, only ⅓ of the second indicator is functional. In this test case there are 4 cell towers involved where all of them can be seen grouped together as one event.

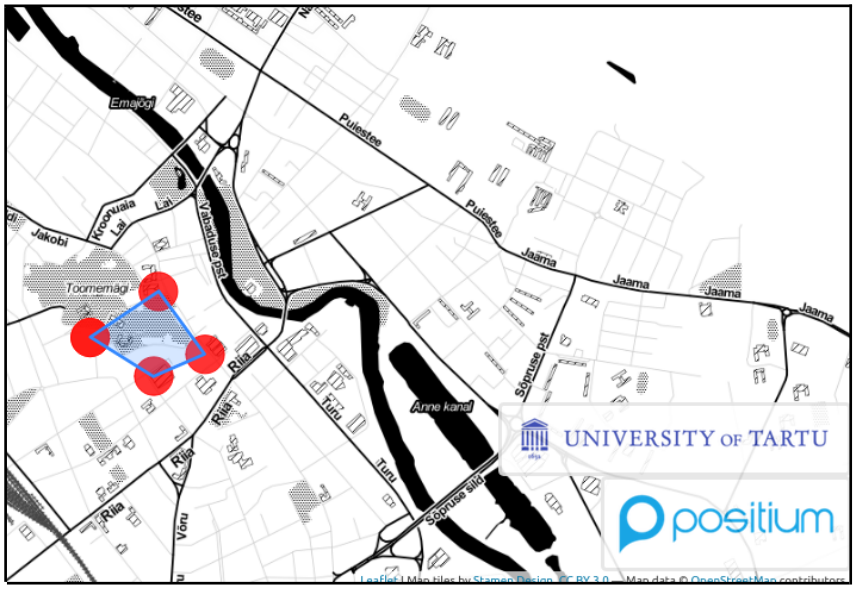


Figure 26. Results of the test case #1.

- January 2nd, at 10:00:

  In this test case only 4 of the candidate cell towers are forming an event on the map due to the distance of the other reported cells from the group, the distance of the other cells is higher than the minimum threshold distance, therefore those cells won't be included in the event. As explained previously a third party cell tower can be seen on the map which originally is not included in the test case cells but it got flagged for 2 indicators therefore it has been shown as a light red dot.

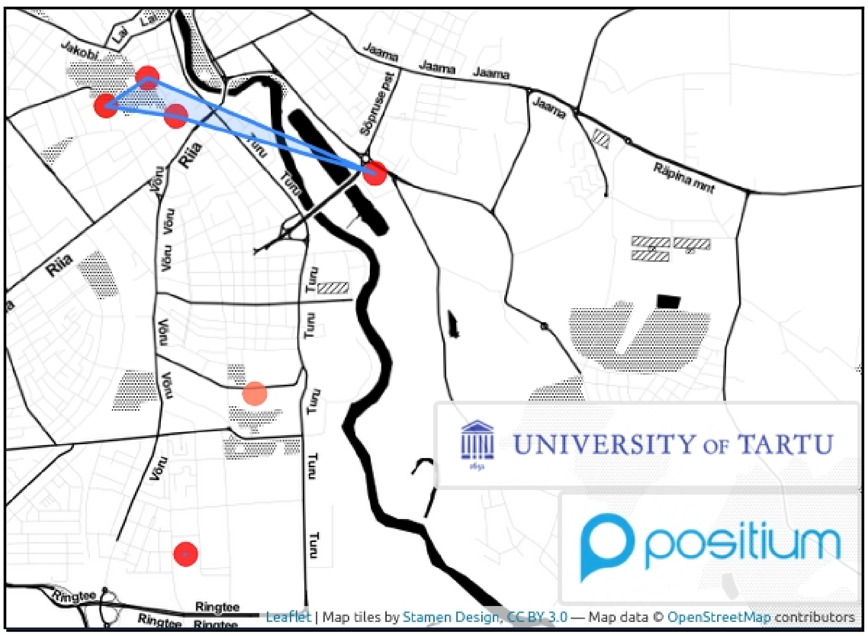


Figure 27. Results of the test case #2.

- January 2nd, at 11:00:

  All of the cell towers present in this test case are visible on the map, and since their distance from other candidate cells are less than the event threshold, all 3 of the cell towers will form a polygon on the map indicating the approximate location of the artificial event.

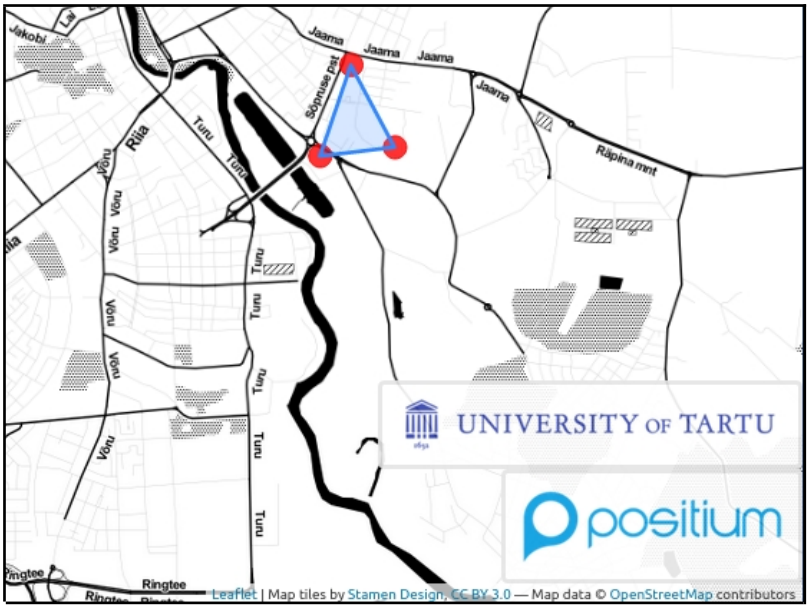


Figure 28. Results of the test case #3.

- January 2nd, at 16:00

  This test case take place in Narva(Left), Tallinn(Middle) and Tartu(Right) simultaneously, the cell towers in Narva and Tallinn are relatively close to one another, therefore they can form an event, on the other hand Tartu have only one cell with three flag situation and due to the large amount of distance from other candidates this cell won't be participating in any detected events.

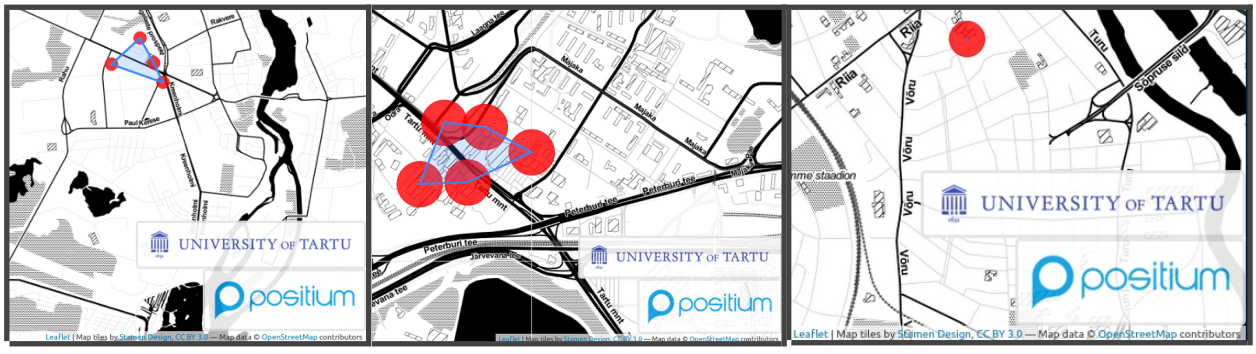


Figure 29. Results of the test case #4.

- January 2nd, at 20:00

  This test case is designed to have cells in the same city but relatively far away from each other making the distance higher than the threshold, therefore as the results demonstrate non of the cells form a group together indicating an event, and simply a peak in the traffic is causing the high traffic in the cells records.

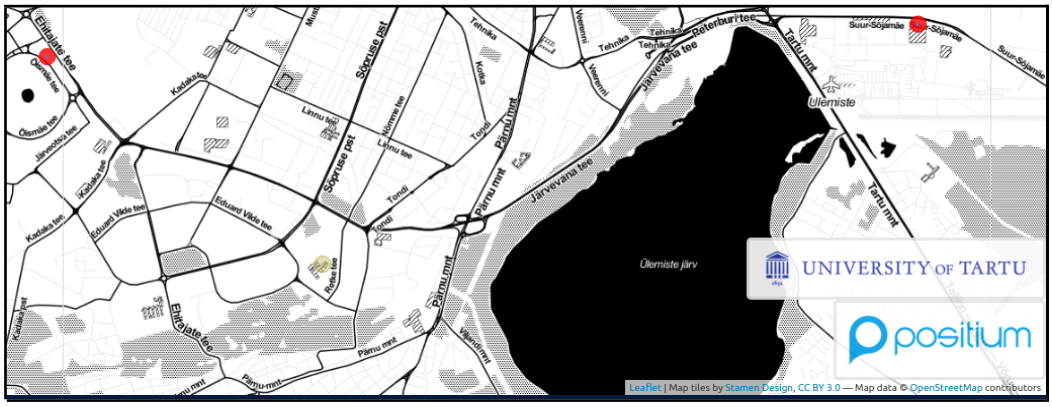


Figure 30. Results of the test case #5 (Tallinn section).

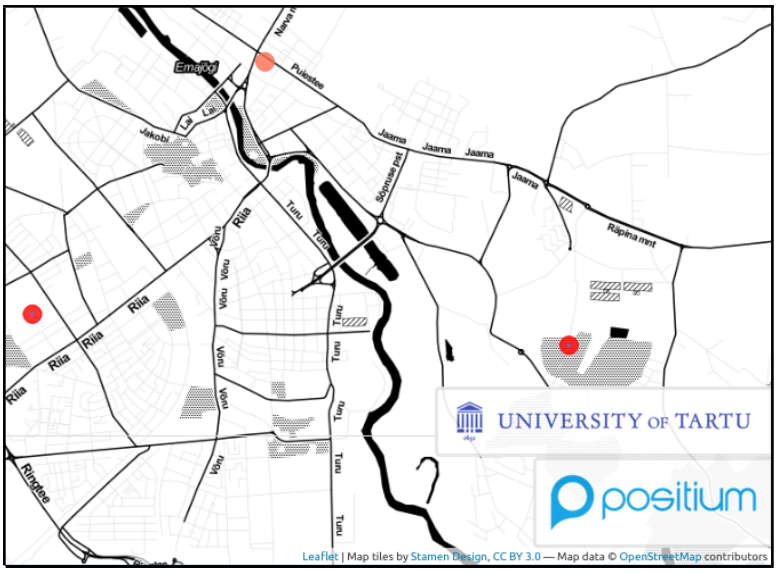


Figure 31. Results of the test case #5 (Tartu section).

- January 8th, at 9:00:

  The second group of test cases belong to January 8th, 2019 which gives the system more than a week of input data. Current data allow the second indicator to function with ⅔ of its subsection. This test case take place in Tartu having multiple 3-flagged cell towers in the same location, but as the result demonstrates on the map the clustering method does not mix up the cell towers together and does not report them all as one event, due to the correct threshold the cells in this test case have been detected as two primary events along with one cell candidate far away from both of the grouped clusters in the same city, proving the ability to distinguish the differences by monitoring their distance in the clustering process.

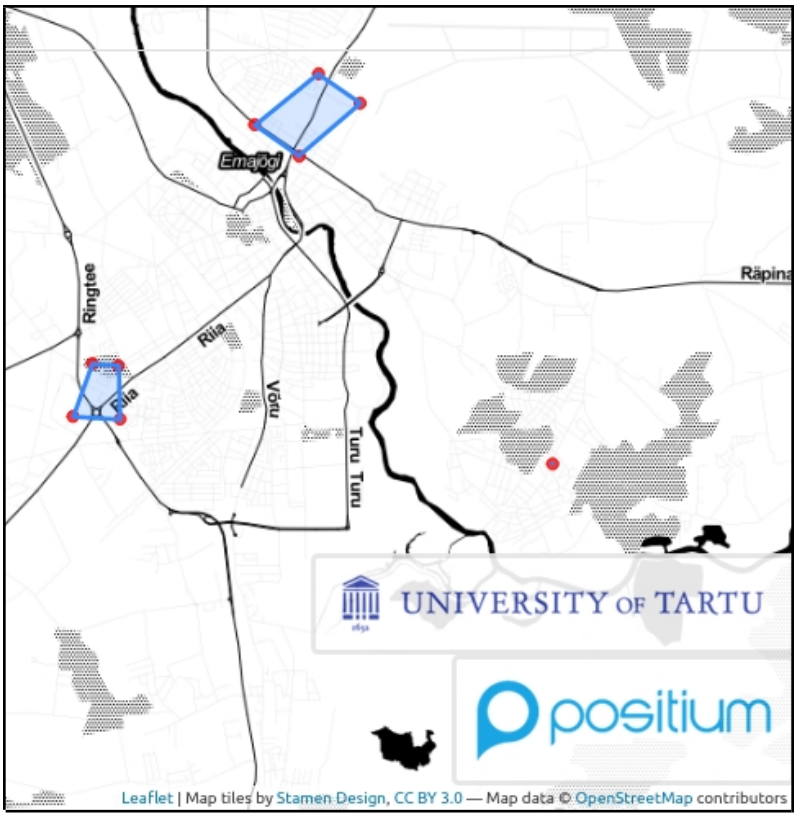


Figure 32. Results of the test case #6.

- January 8th, at 15:00 and 16:00:

  The cell towers for this test case demonstrate how an additional cell tower in the next hour can join the cluster of cells representing an event. The cells at 15:00 o'clock form an event and an additional cell tower will join the cell tower in the next hour at 16:00.

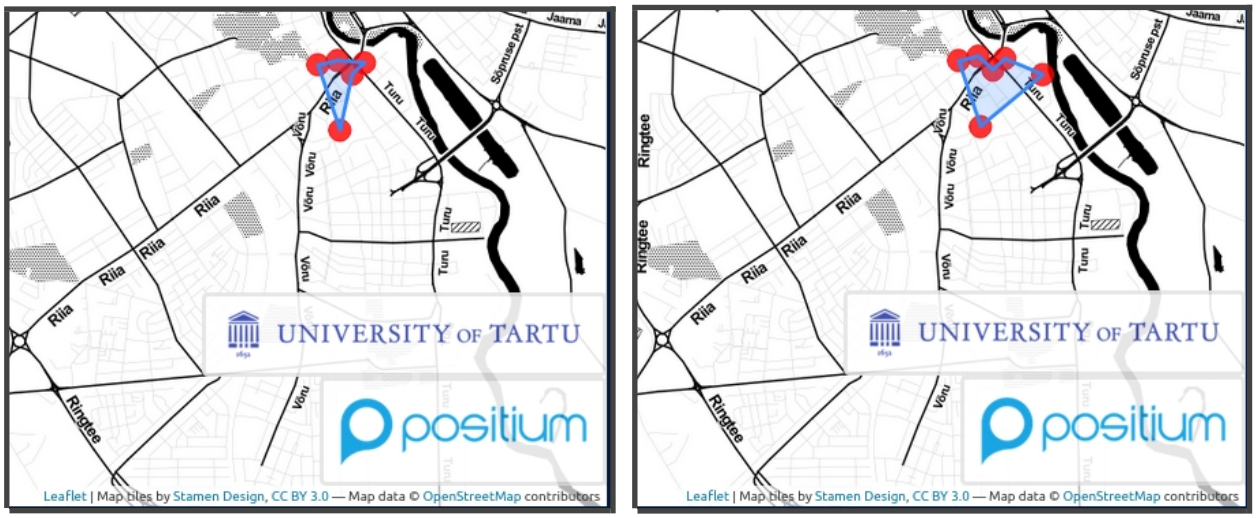


Figure 33. Results of the test case #7 and #8.

- January 14th, at 8:00:

  The 14th of January gives the system a full two weeks of live data, therefore the final subsection of the second indicator (two weeks prior check) will also be active and the system will be checking the input data with all of the indicator's subgroups fully active. As we can see in the demonstration of the result all of the cell towers in the test case have been detected and grouped together as an event just like previous examples.

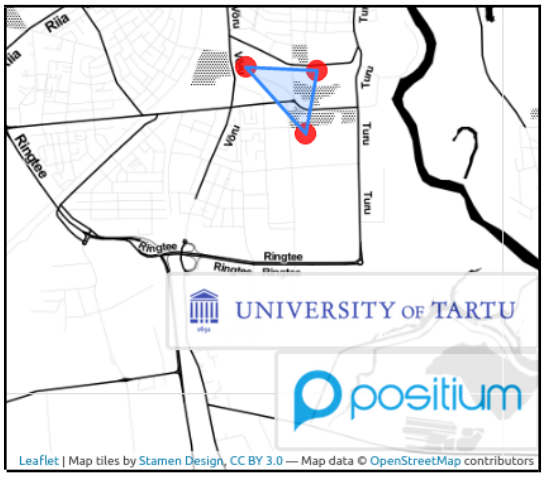


Figure 34. Results of the test case #9.

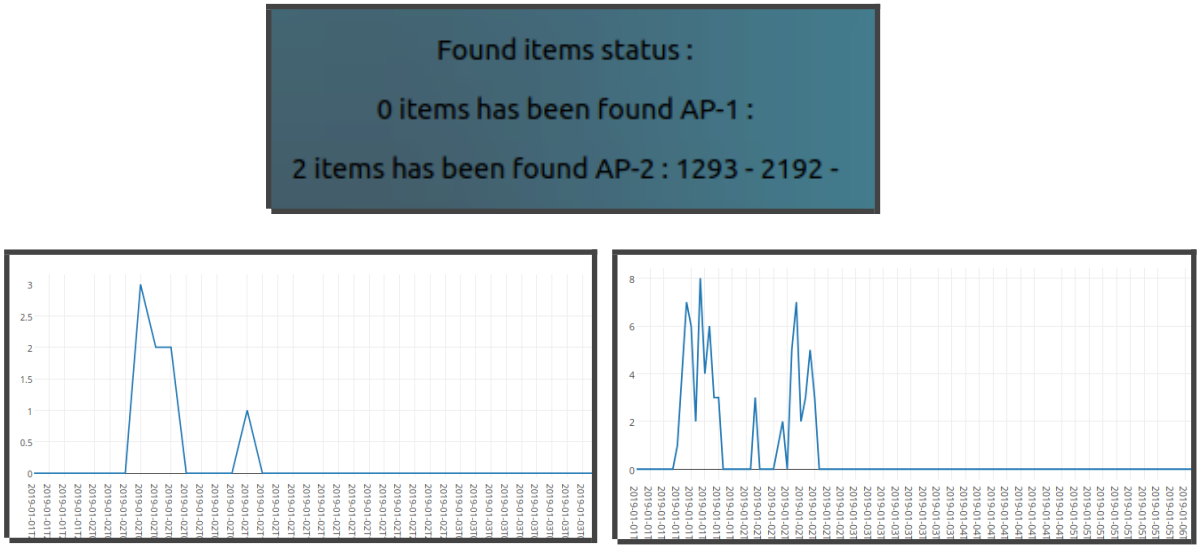


Figure 35. Results of the malfunctionf cells reporting section and the statistical proof of the cells malfunctioning.

The following cell towers had implanted malfunction as test cases for the malfunctioning section and the testing unit for the malfunctioning section reported them accordingly due to the inactivity of the cells from a certain date. By checking the reported cells status at the “Solo Cell Tower Traffic Check” panel of the system the inactivity of the cells will be revealed.

As Explored in the previous section despite the efforts to generate synthetic data based on features gained from the real data, the outcome and the results of the synthetic data is different than the real data. The synthetic data regardless of the injected noises in the data appear to have a harmony while real data can have unpredictable noises on really large amounts. So the system parameters will not be applicable for the test with the live data and needs local change depending on the city and the crowd situation. The system main architecture has been used in the synthetic data testing section and the different values given to the constant values and thresholds in the system formed the training pipeline, the architecture will be the same and the procedure will remain, but the thresholds and constant values need to change according to the cities and the situation of the area.

6.2. Real data assessment

The assessment of the system based on real data is much limited and shorter than the synthetic data due to limitations on the access to real CDR data, this was the initial motive to generate the synthetic data generator script. The only test case available for testing belongs to July 18th, located at Tartu. As mentioned previously in the generation of synthetic data, the fundamental process of synthetic data was based on the behavior of the real data thrived from analytical procedure performed on real data, therefore The limitation on the real data will affect the quality

of synthetic data. Based on early test results on the real data the constant statistical values caused most of the cell towers in Estonia to have at least 2 flags.

Figure 36. Initial results of the real-data tests without balancing the statistical constants.

This incident does not question the architecture of the solution; it only means the constant values used as thresholds in the statistical formula needs to be changed to match the current real data. The reason was that the behavior of real data did not correlate completely with the synthetic data so changing the constant statistical values returned the situation of the cells to normal. Prior to the event the increase in the flow of traffic in the cells located around Raadi airfield is visible on the map.

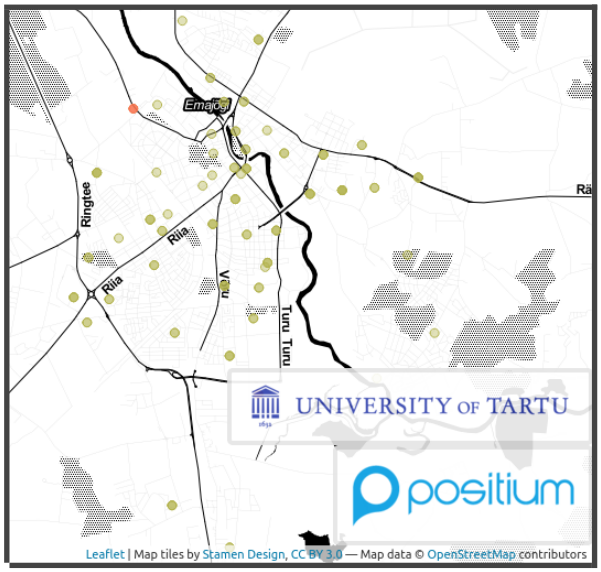


Figure 37. real-data results one day prior to the concert date.

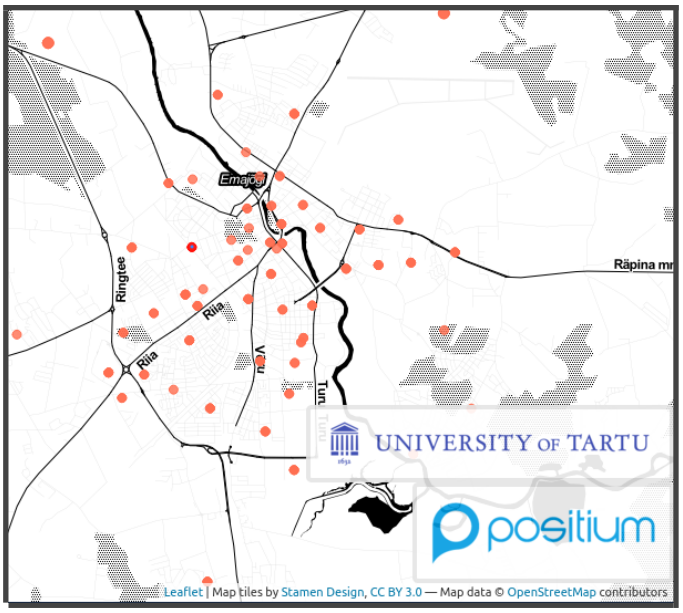
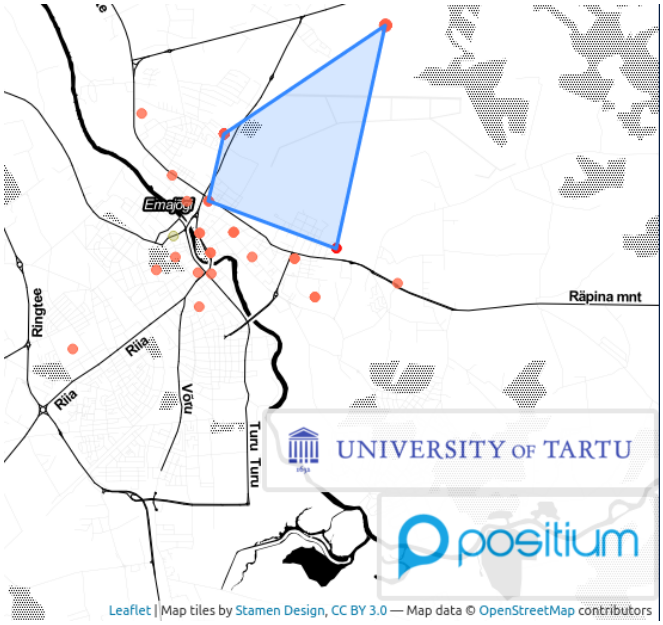


Figure 38. Real-data results hours before beginning of the concert.

As the flow of traffic among cell towers increases in Tartu but none of the cell towers have 3-flags to be involved in an event. After the beginning hours of the concert at 22:00 and 23:00 the cell towers surrounding the concert area reach the threshold levels and all flag with 3 indicators creating a closed polygon around the Raadi airfield.



igure 39. real-data results at one of the hourly snapshots at 23:00 during the concert.

The closed polygon is located directly on Raadi airfield, estimating the approximate location of the event correctly. The problem with the real data test was that the exact location of the event was revealed in snap shots of two hours not for the whole duration of the concert. After checking each individual cell tower at the concert area and reporting in the above picture, it appears that the flow of traffic does not remain high continuously and the only peaks in the cell towers at the location of the event was during the only two hours that the system detected the events. Investigation on the other involved cell towers had the same results. This result from real test data reveals the fact that CDR data will help reveal the location of the ongoing events but there are possibilities that the event location will be represented in hourly snapshots not for the full duration of the event due to the non-continuity nature of CDR data. Using CDR data in different time lines will have different results, currently the technology of 4G and 3G made this accessible and available to the public to have internet connection nearly everywhere, and most of the communications are being established by internet connection rather than traditional phone calls. During numerous types of events people are constantly using internet packet data to communicate but at some points of the event tend to make a phone call.

Ultimately the best method to have full access to an event's approximate location for the full duration of the event, is to convert the input data to a dataset with higher continuous rate. By converting the input data from CDR data to internet packet transaction between mobile devices and the cellular towers, the location of the event will be available for the full event’s duration. The architecture and the procedure of the system remains the same; the only difference will be the nature of input data in order to improve the hourly snapshots of event location to full access for the whole duration of the event.

6.3. Future Works

Systems for detection of crowd in a concentrated area have been studied, developed in a variety of methods with different requirements. The current method studied in this thesis used one the most common resources for achieving the goal. Changes, alternate methods and fundamental changes in different sections of the application can help improve the system for more accuracy and flexibility in order to deal with different residential areas dynamically. Currently the constant values affecting the core functionality are stored in the database and in order to change the values to adapt the system with a new dataset the values need to change in the database, adding a new section in the user interface panel will give user the ability to adapt the system with new datasets and possibly for testing purposes without needing access to database. Particular ground breaking changes require access to much more data. Essentially access to more real data will affect different sections of the application such as reliability and accuracy of the synthetic data, deeper and more tests on live data, quicker adapt with variety of real data. An important addition to the system in the condition of having access to more real data is adding new indicators to the system from the machine learning department. The main reason behind the failure of the machine

learning approach in this project was the lack of access to large volumes of real data, as mentioned in the earlier section of this document by providing more data the neural network approach will become an option and possibly can be a great candidate as the next indicator.

7. Summary

Predicting human behavior in many different fields of studies has always been challenging in many different levels. Studying human behavior always leads to a common goal among researchers which is optimization and better quality for people’s daily lives.

The urban studies are not an exception in this list, and there are a variety of fields of studies dedicated to people's behavior regarding their transportation in the city. One of the most important fields of study is crowd detection and social adhering detection. Essentially any means that indicate the presence of people can be used as a resource for this area of study. In the field of event or social gathering detection there are many different methods using different tools to gather essential resources for the same purpose such as using an accurate GPS tracking system in, performing image processing on traffic camera feeds in order to estimate the flow of people in each section of the city and etc. Compared to other approaches current approach taken in this project is more comprehensive than approaches with volunteered GPS candidates in the crowd and more low cost compared to the approaches requiring high level image processing units from public video feeds.

In this study the focus is on using the call details and people interactions with mobile networks as resources to trace for event detection. Anytime an operating system's subscriber initiates a phone call the mobile device will connect to the closest cellular tower in the area in order to connect the subscriber to a network of mobile devices and phones to initiate the call. Every subscriber is assigned with an anonymous identifier and in the process of phone call initiation the base station of the cellular tower will store a record of the ongoing call including subscriber ID, date and time of the call and identifier of the cellular tower itself. Gathering this information from all around the city or possibly the country generates a big data resource of call records from different cell towers. Another data represented by the operating systems are the meta-data about the cell towers including the coordinates of the towers, by joining the previous big data with current information a well suited resource for event detection will be obtained. By applying a procedure of data aggregation the amount traffic gone through each cell tower in a particular date and time will be concluded also known as time series data, which is the main resource used in the current study.

In this project a series of statistical methods will be utilized on a two parallel servers in order to monitor and process the traffic of the cell towers for possible abnormal peaks in the time series data. The peaks in the traffic demonstrate a large number of phone calls initiation through the corresponding cell towers. The statistical method uses three different approaches named “indicator” for studying the cell tower’s behavior due to the fact that the abnormality in the data must be proven from many different aspects. Using three different approaches and comparing the current behavior with previous year’s traffic and previously imported data in the previous days and hours the abnormal peak in the data will be proven. By identifying cell towers with the same situation in the same area and creating a geographical cluster of cell towers, an estimation on the location of the event or the social gathering causing the abnormality will be achieved. Based on the tests on the synthetic data and real data the results of this system depends on the amount of data the system has access to, the training of the system and tests for different possible

challenges in the input datasets will be done by generating synthetic data based on real data, therefore by accessing more data the accuracy of synthetic data will be higher and the system will be better trained. In overall this system will generate acceptable estimation on the data due to the fact that phone call initiation is still considered to be a common means of communication using a mobile device, the downside observed in this system is the lack of continuity in flow of traffic in an event related cell tower. Based on the results from the real data provided for the system the people present at the event won’t be making phone calls constantly during the event and at some point the majority of the auditions will be more focused on the gathering and the system will monitor for abnormality detection on hourly basis, therefore the estimation of the event location will be in snapshot hours not continuous hours. Utilizing a more stable and continuous source of information that matches more people's interactions with their phone in 2020 will result in much more continuous and accurate results such as internet packet transactions per each subscriber to the cell tower’s base station, since people tend to use internet on their phone far more often than initiating phone calls.

8. Conclusion

This thesis is aimed for the detection of a large group of people forming a gathering in a social event based on mobile positioning data. Among many other methods proposed for accomplishing such a task, the current method had a well-balanced situation between the cost efficiency of the system, being comprehensive and resource accessible. The implemented system is an enterprise for monitoring the cell towers traffic in general and the core functionality is the event detection mechanism. Based on the behavior of the mobile positioning data as input resource for this project a statistical method comprising three different approaches has been implemented for detecting the potential events. Based on the tests done on the system using both synthetic data and real data, the system was able to detect ongoing events on the synthetic data and pass all of the test cases while on the real data due to the extended length of the event and non-continuity of the call detail records used as input data the event got detected in snapshots of hours not for the whole event duration. In the future this matter can be addressed by replacing the call detail records with internet packet data transaction records the continuity of the data even during the event regardless of the length of the event will be provided, because considering the current technological capacities of mobile phones the internet usage is highly more frequent compared to cell phone calls for communication purposes.

9. References

[1]Kotsiantis, Sotiris B., I. Zaharakis, and P. Pintelas. "Supervised machine learning: A review of classification techniques." Emerging artificial intelligence applications in computer engineering 160.1 (2007): 3-24.

[2]Hansen, Lars Kai, and Peter Salamon. "Neural network ensembles." IEEE transactions on pattern analysis and machine intelligence 12.10 (1990): 993-1001.

[3]Catanzaro, Michael 2018. “electrical Department Cellular Program”, 8-9.

[4]https://www.researchgate.net/figure/Hexagonal-cells-in-a-GSM-system-created-by-placing-the-base-stations-at-the-intersection_fig4_4293285

[5]Sikder, Ratul, Md Jamal Uddin, and Sajal Halder. "An efficient approach of identifying tourist by call detail record analysis." 2016 International Workshop on Computational Intelligence (IWCI). IEEE, 2016.

[6]Saluveer, Erki, et al. "Methodological framework for producing national tourism statistics from mobile positioning data." Annals of Tourism Research 81 (2020): 102895.

[7]Dong, Honghui, et al. "Traffic zone division based on big data from mobile phone base stations." Transportation Research Part C: Emerging Technologies 58 (2015): 278-291.

[8]Liang, Victor C., et al. "Mercury: Metro density prediction with recurrent neural network on streaming CDR data." 2016 IEEE 32nd International Conference on Data Engineering (ICDE). IEEE, 2016.

[9]Doyle, Casey, et al. "Predicting complex user behavior from CDR based social networks." Information Sciences 500 (2019): 217-228.

[10]Vajakas, Toivo, Jaan Vajakas, and Rauni Lillemets. "Trajectory reconstruction from mobile positioning data using cell-to-cell travel time information." International Journal of Geographical Information Science 29.11 (2015): 1941-1954.

[11]Li, Mingxiao, et al. "Reconstruction of human movement trajectories from large-scale low-frequency mobile phone data." Computers, Environment and Urban Systems 77 (2019): 101346.

[12]Chen, Guangshuo, et al. "Individual trajectory reconstruction from mobile network data." (2018).

[13]Lind, Artjom, Amnir Hadachi, and Oleg Batrashev. "A new approach for mobile positioning using the CDR data of cellular networks." 2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS). IEEE, 2017.

[14]Hadachi, Amnir, and Artjom Lind. "Exploring a new model for mobile positioning based on CDR data of the cellular networks." arXiv preprint arXiv:1902.09399 (2019).

[15]Ranneries, Søren B., et al. "Wisdom of the local crowd: Detecting local events using social media data." Proceedings of the 8th ACM Conference on Web Science. 2016.

[16]Gu, Yiming, Zhen Sean Qian, and Feng Chen. "From Twitter to detector: Real-time traffic incident detection using social media data." Transportation research part C: emerging technologies 67 (2016): 321-342.

[17]Zhu, Zack, et al. "Human activity recognition using social media data." Proceedings of the 12th International Conference on Mobile and Ubiquitous Multimedia. 2013.

[18]Hu, Shaohan, et al. "Smartroad: Smartphone-based crowd sensing for traffic regulator detection and identification." ACM Transactions on Sensor Networks (TOSN) 11.4 (2015): 1-27.

[19]Davies, Anthony C., Jia Hong Yin, and Sergio A. Velastin. "Crowd monitoring using image processing." Electronics & Communication Engineering Journal 7.1 (1995): 37-47.

[20]Yin, Jia Hong, Sergio A. Velastin, and Anthony C. Davies. "Image processing techniques for crowd density estimation using a reference image." Asian Conference on Computer Vision. Springer, Berlin, Heidelberg, 1995.

[21]Reisman, Pini, et al. "Crowd detection in video sequences." IEEE Intelligent Vehicles Symposium, 2004. IEEE, 2004.

[22]Palshikar, Girish. "Simple algorithms for peak detection in time-series." Proc. 1st Int. Conf. Advanced Data Analysis, Business Analytics and Intelligence. Vol. 122. 2009.

[23]Cattell, Raymond B. "The scree test for the number of factors." Multivariate behavioral research 1.2 (1966): 245-276.

[24]D'agostino Sr, Ralph B., and Heidy K. Russell. "Scree test." Encyclopedia of Biostatistics 7 (2005).

9.1. License

Non-exclusive licence to reproduce thesis and make thesis public

I, Mohammad Mahdi Mohebbian

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,

Real-Time Event Detection System for Mobile Data

supervised by Amnir Hadachi, PhD and Erki Saluveer PhD.

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

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

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

Mohammad Mahdi Mohebbian

10/08/2020