UNIVERSITY OF TARTU

Institute of Computer Science
Computer Science Curriculum

Joosep Rõõmusaare

Probabilistic Location Estimate of Passive Mobile Positioning Events

Master's Thesis (30 ECTS)

Supervisor: Toivo Vajakas, MSc

Tartu 2016

Passiivsete mobiilipositsiooni sündmuste tõenäosuslik asukoha hinnang

Kokkuvõte: Uurijad, kes on püüdnud mõista inimeste liikumise mustreid, korjavad andmeid mobiilivõrkudelt. Mobiilid teevad sündmuse kirjeid iga kord, kui nendega helistatakse, saadetakse SMSi või kasutatakse Interneti. Sündmuste kirjed sisaldavad informatsiooni sellest, millisesse võrgu transiiversisse mobiiltelefon oli sel hetkel ühendatud. Võrgu ühe transiiveri leviala saab kasutada, et püüda positsioneerida telefoni geograafilist asukohta. Kasutades positsioneerimiseks transiiveri leviala, siis need hinnatavad asukohad pole punktid kaardil, vaid geograafilised alad, kus telefon võib olla kui ta on transiiveriga ühendatud.

Mobiilide ühendamine transiiveritega sõltub mitmest muutujast, mis tähendab, et mobiil ei ole alati ühendatud kõige tugevama signaaliga transiiveriga. See teeb mobiili asukoha hindamise keerulisemaks, sest transiiverite levialad võivad üksteisest üleulatuda.

Võrguplaan kirjeldab võrgus olevate transiiverite levialasid ning seda kasutatakse, et defineerida transiiverite levialasid.

Selles lõputöös hinnatakse mobiilisündmuste positsioneerimise kvaliteeti ruumilise jaotuse tihedusfunktsioonidega. Luuakse erinevad võrguplaani variandid ja erinevate võrguplaanide kvaliteeti hinnatakse Bayesi statistikaga ja kasutatakse reaalseid asukoha andmeid. Erinevate võrguplaanide kvaliteeti hinnatakse suurima tõepära meetodiga.

Võrreldi RSSI ja Voronoi põhjal tehtud võrguplaane ja nende modificatsioone ja leiti, et Voronoi võrguplaanide puhul paistis asukoha positsioneerimine paremini kui RSSI võrguplaanide puhul.

Lisaks uuriti, kuidas transiiverite levialade üleulatamisel arvestamine Bayesi meetodiga mõjutab asukoha positsioneerimise täpsust. Leiti, et Bayesi levialade üleulatamise meetod tegi halvemate võrguplaanide täpsust paremaks, aga paremate võrguplaanide täpsust halvemaks.

Võtmesõnad: võrguplaan, mobiilivõrk, visualiseerimine, mobiilne positsioneerimine

CERCS: P170 Arvutiteadus, arvutusmeetodid, süsteemid, juhtimine (automaatjuhtimisteooria)

Probabilistic Location Estimate of Passive Mobile Positioning Events

Abstract: Researchers, who are trying to understand human mobility patterns, collect data from cellular telephone networks. Mobiles are creating events every time they are used for calling, SMS, or the Internet. The events contain the information, in which network cell that mobile was at the moment of the event. Cell's coverage can be used for estimating the geographical location of the mobile. The estimated locations are not a point on the map, but the possible area, where the mobile may be when they are connected to that specific cell.

Mobiles connecting to cells are depending on multiple variables, meaning, that a mobile may not always connect to the cell with the strongest signal. That makes estimation of the mobile location more difficult, as the coverage areas may overlap with each other.

Cell plan is a description of cell coverage areas and there are multiple ways for defining cell coverage areas.

This thesis is about estimating mobile events positioning quality with spatial probability density functions. Different cell plan variants will be implemented and real ground truth location data is used to find the modification that maximizes the likelihood estimation.

RSSI-based and Voronoi-based cell plans and their modifications were compared. Results showed that Voronoi-based cell plans are better for location positioning than the RSSI-based cell plans.

Furthermore, Bayesian overlapping method was examined to see if applying it would improve location positioning accuracy. It was found that applying Bayesian overlapping methods improved the accuracy of the worse cell plans, but made accuracy worse for the better cell plans.

Keywords: cell plan, cellular network, visualizing, mobile positioning

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

Contents

List of Figures

List of Tables

List of Algorithms

List of Abbreviations

A-GPS	Assisted Global Positioning System
AML	Active Mobile Location
BSC	Base Station Controller
BSD	Best Serving Data
BTS	Base Transceiver Station
CDMA	Code Division Multiple Access
CDR	Call Detail Record
CGI	Cell Global Identity
CI	Cell Identity
CSV	Comma Separated Value
GPS	Global Positioning System
GSM	Global System for Mobiles
HSPA	High Speed Packet Access
HSPA+	Evolved High Speed Packet Access
IMEI	International Mobile Equipment Identity
JSON	JavaScript Object Notation
LAC	Location Area Code
LBS	Location Based Services
LTE	Long Term Evolution
MCC	Mobile Country Code
ME	Mobile Equipment
MLE	Maximum Likelihood Estimation
MNC	Mobile Network Code
MNO	Mobile Network Operators
MPS	Mobile Positioning System
MSC	Mobile Switching Centers
NMS	Network Management System
PDF	Probability Density Functions
PSTN	Public Switched Telephone Network

RSSI	Received Signal Strength Indicator
SIM	Subscriber Identity Module
SINR	Signal-to-Interference-and-Noise Ratio
SPDF	Spatial Probability Density Functions
TA	Timing Advance
UMTS	Universal Mobile Telephone System
VLR	Visitor Location Register
WCDMA	Wideband CDMA
WiFi	Wireless network
XML	Extensible Markup Language

Definitions of some basic terms.

Mobile network

Base Station Controller - Part of the mobile network. Controls BTSs in a single location area. Handles allocation of radio channels, frequency, power, signal measurements, and handovers within a single location area.

Base Transceiver Station - Mobile equipment access point to the network. Handles speech encoding, encryption, multiplexing and modulation/demodulation of the radio signals. Each BTS has an assigned CGI.

Cell - The geographical area covered by a BTS.

Cell plan - Mapping of each cell in the mobile network so each cell has an approximated geographic area where the ME can connect to that cell. In practice, it is usually defined as a polygon for each cell. In this work we use the same term for the set of continuous SPDF estimates for each cell in the mobile network.

Handover - Changing ongoing call or data session from one BTS to another.

Mobile Equipment - Physical phone or device what is used for connecting to the cellular network. Each ME is uniquely identified by the IMEI number.

Network Management System - Part of the mobile network. Handles administrative and central procedures. Holds data storages warehouses and network handling servers.

Mobile Positioning Data - Data collected with mobile positioning about the mobile device location and movement.

Mobile Switching Center - Part of the mobile network. Controls BSCs. Handles call and messaging routing, updating different registries within MSC and the central system, and handovers between different location areas.

Statistics

Maximum Likelihood Estimation - MLE is a method of estimating population characteristics from a sample by choosing the values of the parameters that will maximize the probability of getting the particular sample actually obtained from the population.

Probability Density Function - The PDF of a continuous random variable is a function that can be integrated to obtain the probability that the random variable takes a value in a given interval. Formally, the PDF $f_x(x)$ of a continuous random variable X is the derivative of the $f_x(x) = \frac{dF_x(x)}{dx}$.

1 Introduction

1.1 Motivation

Researchers are trying to understand human mobility patterns. Insight to human mobility could help with different issues. In urban planning, understanding how people come and go helps to determine how to build infrastructure, how to reduce traffic congestions, or how to reduce pollution. Knowing people home and work locations, helps to plan commuting routes between work and home. In medicine, it could help to understand how some disease spreads. Human movement data is needed to analyze this kind of problems. Collecting that data traditionally is costly. Doing movement surveys and direct observations need time, money and usually, result in small sample sizes [Becker et al., 2013].

A simpler way to collect data is to use the Mobile Network Operators (MNO) collected positioning data. Most people in the world use mobile phones and MNO collects events made when mobiles are used for SMS, calling or the Internet. These events metadata includes the cell tower where the mobile device was connected at the time of the event. This data is collected for other purposes but the location info can be used for positioning the mobile devices [Tiru, 2014].

The cell tower is at one point on the map, but that does not mean that the mobile device is there. Location positioning accuracy could be 100m to 1km in urban areas and in rural areas up to 30 km [Saluveer et al., 2012]. To describe the possible location area for the mobile device, that is connected to some cell, cell plans are used to visualizing cell effective area coverage. MNOs usually have details following: each cell tower location on the map; which direction each cell on the tower is facing; other cell describing attributes [Tiru, 2014]. What they do not always have is effective area coverage information. For that different algorithms are used for generating that area.

1.2 Structure of the thesis

Chapter 1 has the introduction. It gives an overview of the topic, describing cellular network and mobile positioning data. After that, there is described what has been previously done in this topic research and what this thesis will add to that. Chapter 2 describes the data and different methods what are used in this thesis. Chapter 3 shows the results of the thesis and Chapter 4 contains the conclusions and the discussion about the results.

1.3 Background

1.3.1 Description of the field

Mobile technologies. To call another person, you need a mobile phone and the network connection. So the first thing you need is Mobile Equipment (ME), that is your physical phone or device what is used for connecting to the cellular network. Each ME is uniquely identified in the network by the International Mobile Equipment Identity (IMEI) number. In Europe, everyday phone users do not encounter this problem, but ME must be able to operate on a cellular network. There are two basic technologies in mobile phones: Code Division Multiple Access (CDMA) and Global System for Mobiles (GSM) [COAI, 2016].

Segan [Segan, 2015] describes what technology you need when you want to buy a new phone in the United States. In Europe, GSM radio system is used. Problem is that with GSM carriers verify users with Subscriber Identity Module (SIM) card and CDMA carriers use network-based whitelists. Phones with only GSM support can not operate in CDMA networks and vice-versa.

Besides that when you change the phone you just have to take out the SIM in GSM phone and put it into another phone. In CDMA you need carrier's permission to change the phone.

The main difference is in the core radio signaling technology. CDMA uses spread spectrum signal encoding, which transmits data of each subscriber over the entire frequency spectrum of the antenna all the time, while GSM assigns a time slot for each subscriber's data where no other subscriber can have access [Tiru, 2014]. In CDMA "code division" each subscriber unique key is assigned which is used to decode combined signal into its individual calls. In GSM "time division" subscriber pieces the call back together from assigned time slots.

Code division method need more processing power, but is more powerful and flexible technology and when GSM was upgraded to 3G, it started to use CDMA technology, but it was called Wideband CDMA (WCDMA) or Universal Mobile Telephone System (UMTS). WCDMA uses wider channels than CDMA (like the name suggests) but has more data capacity. To increase data transfer speeds, GSM 3G was upgraded to High Speed Packet Access (HSPA), which is known as 3.5G. Later HSPA was upgraded to Evolved High Speed Packet Access (HSPA+)(3.75G).

Both systems will be soon upgraded to a common global standard named Long Term Evolution (LTE). LTE is globally accepted 4G wireless standard. Report [OpenSignal, 2016] shows that currently there are 148 countries with LTE networks and 10 countries are scheduled to go online with LTE networks. First LTE network was deployed in Estonia 2010 by Telia (then named EMT), and all three main MNOs had LTE support from the beginning of 2013.

Besides technological platforms CDMA, GSM or LTE, there are also different standards and protocols that must be supported by the mobile devices. This includes network "generations" 2G, 3G, 4G and frequency bands that they work in. 2G was first digital network and had added a link to the Internet. 3G increased the communication speed and added various value-added services like video calling, live streaming, mobile internet access, etc. 4G (also known as LTE) enabled ultra-broadband Internet access, gaming services, and high definition television [Tondare et al., 2014].

In Estonia, 2Gs is supported on bands GSM 900, GSM 1800, 3G on UMTS 900, UMTS 2100, 4G on LTE 800, LTE 1800, LTE 2600 [GSMArena, 2016]. Numbers behind the technologies are showing in what frequency band they are working (i.e. 900 means that it works in a frequency band that is near 900MHz).

Public switched telephone network Mobile equipment connects to the network by the Base Transceiver Station (BTS). It is the antenna that is on top of the radio tower. It carries out radio communications between the network and the mobile equipment. It handles speech encoding, encryption, multiplexing and modulation/demodulation of the radio signals [Tiru, 2014]. BTS has an assigned CGI value. CGI is made of four identification values: Mobile Country Code (MCC), Mobile Network Code (MNC), Location Area Code (LAC) and Cell Identity (CI). MCC shows in what country cell is location, MNC to which network operator it belongs to and LAC shows in which location area it belongs. CI shows cell id, depending on MNOs it may be unique over all network or unique in LAC [shareTechnote, 2016].

There are two types of BTSs: directional and omnidirectional. BTSs have usually the directed antennas and one BTS covers 120-180 degree sector of an area and a tower with 3 BTSs will cover all 360 degrees. Depending on use cases one tower could also hold only one or two BTS with different sector degrees and for redundancy one tower could also be serviced by several overlapping BTSs. Omnidirectional antennae transmit in all directions and are usually alone on their tower [Kwan et al., 2012].

Base Station Controller (BSC) controls a number of BTSs in a proximate geographical area, that is BTS in a single location area. BSC handles allocation of radio channels, frequency, power and signal measurements. If ME changes BTS in a single location area, then BSC is responsible

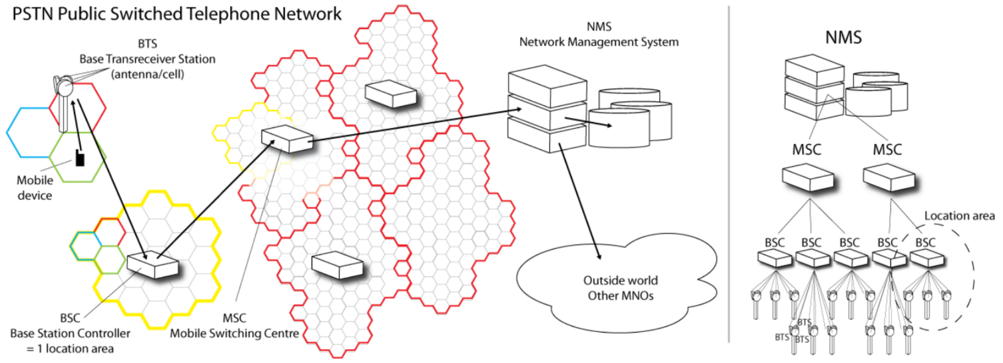


Figure 1: The Illustration of PSTN and hierarchy of its network components [Tiru, 2014].

for handover procedures. Handover is changing ongoing call or data session from one BTS to another.

BSCs are controlled by Mobile Switching Centers (MSC). MSCs are one or many, depending on the size of the MNO. MSC handles the call and messaging routing, updating different registries within MSC and the central system and handover between different location areas and MSCs. MSC also contains Visitor Location Register (VLR). VLR is the registry for holding the information about the location area and the BTS to which ME are connected.

MSC reports to the Network Management System (NMS). All the administrative and central procedures are done in NMS. Data storage warehouses and network handling servers are also there. This network with connections to the other MNOs is called Public Switched Telephone Network (PSTN) and is illustrated in Figure 1.

MNOs could share some of the network equipment and network modules between themselves (e.g. it is not uncommon to share the antenna between MNOs). There are even virtual MNOs that do not own any network infrastructure but rent it from other MNOs [Tiru, 2014].

Cell plan. Each BTS has a Cell Identity (CI), like was mentioned in the last subsection. Cell is a geographical coverage area, where ME could connect to that cell BTS. Radio tower where the BTS is connected is called the cell site. The cellular network is made of BTSs and their cells, hence the name. Each cell can cover a limited number of ME within the cell coverage area. Capacity is limited by available bandwidth and operational requirements. To give the best user experience, MNO has to add or remove BTSs and size cells after the need to handle expected traffic demands [COAI, 2016].

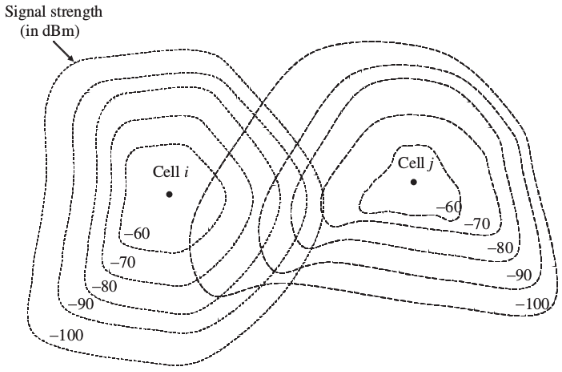


Figure 2: Cell RSSI decreases with distance increase from the cell site [Agrawal and Zeng, 2010].

BTS with omnidirectional antennae tend to have cell with a hexagonal pattern, directional antenna is ofter diamond-shaped [Kwan et al., 2012]. But in real world shapes are more distorted by atmospheric conditions, topographical contours or architecture. ME connects to a BTS over radio signal, so probability for ME to connect to a specific BTS is depending on Received Signal Strength Indicator (RSSI). Signal strength is measured in dBm, further the ME from BTS is, the lower the dBm value. RSSI change with cell's distortions and overlapping is illustrated in Figure 2 [Agrawal and Zeng, 2010].

When ME is close to the edge of a cell, where RSSI is low, then handover occurs and ME is serviced by another BTS. When ME is near the edges of multiple cells, then ping-pong effect may occur, meaning ME will be switched from one BTS to another multiple times in a row. For ME user everything works fine, but by BTS positioning, it will show that ME location starts jumping fast from one cell to another [Vajakas et al., 2015].

Cell size depends on its frequency and type. Higher frequency cell's strength will decrease faster than with lower frequency. Cells are categorized by their size from largest to smallest: macro, micro, pico, femto.

Macro cells main purpose are to cover large areas. Radius may be up to several kilometers. Generally most rural and suburban areas are covered by macro cells.

Micro cells are mostly used in urban areas with a radius of hundreds of meters. They are deployed to places with higher traffic densities, e.g. large shopping malls.

Pico cells are small cells with radius up to tens of meters. Used in places with very high traffic volume, e.g. train stations or office buildings. BTS of pico cells tend to be omnidirectional.

Femto cells are smallest cells that are used in residential indoors where the mobile operator network may not be provided by the operator. Instead, it may be connected to the digital subscriber line (i.e. digital data is transmitted over telephone lines) .

These categories are complemented with two additional types: umbrella and metro cells. Umbrella cells are the combination of larger cells with several smaller cells, where a larger cell is used for providing coverage and mobility between smaller cells. Metro cells are very similar to micro cells, but metro cells integrate all elements required for the BTS operation in one compact device [Penttinen, 2015].

Cell plan is the mapping of each cell in the mobile network so each cell has an approximated geographical area where the ME can connect

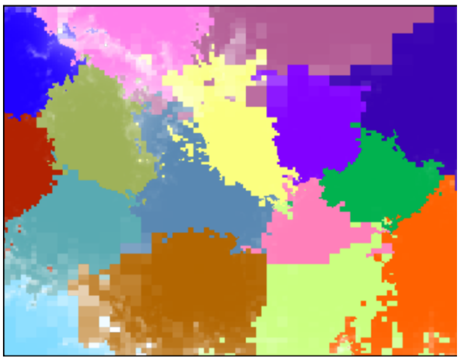


Figure 3: Best serving data where each cell is in the different color [Calabrese, 2011].

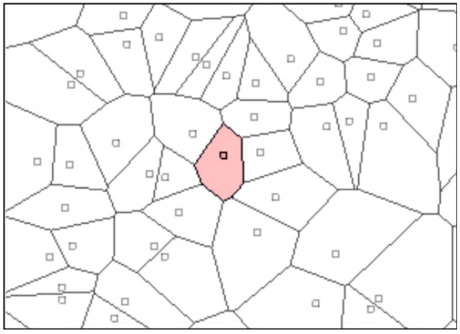


Figure 4: Voronoi tessellation where the rectancle points are cell sites [Calabrese, 2011].

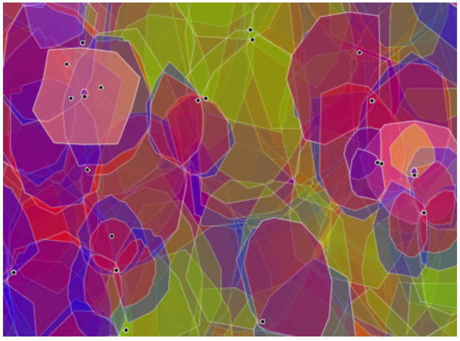


Figure 5: Cell plan with cells are colored by the frequency. Dots are the cell sites.

to that cell [Vajakas et al., 2015].

CDR gives a cell ID, usually in a form of CGI) where the ME was connected on when some event happened. That ID gives the cell site geographical location. When BTS is omnidirectional, then site's location could be good enough for ME location, because it is in the center of its coverage area. When BTS has a directional antennae, then BTS is at the corner of the coverage area and mostly ME is somewhere else when it is connected to that BTS. More precise positioning would be using the center of the coverage area geometry. There are two methods to get the cell geometry. One is to measure or calculate RSSI measures for any geographical location. MNOs have special programs for generating best serving data (BSD). Best serving data is cell map that takes into account cell sector and propagation models and is made on simulated coverage. Each location in a grid is associated with the BST which covers that location best as shown in Figure 3 [Calabrese et al., 2014]. This method assumes that when ME is in some location where some BTS has the best serving (RSSI value is better than other cells), then ME would connect to that cell.

Another way to create the cell geometry is a more simple way without the need for radio measurements. Free-space propagation is assumed with that all BTSs have the same equivalent radiated power. With these assumptions, Voronoi tessellation technique could be used on the cell site locations. Voronoi tessellation partitions the space into cells based on the distance between each point and the closest BST. Set of points that are closer to the one BST than others is included to that cell [Csáji et al., 2013]. Resulting graph is shown in Figure 4.

Cell plans in real world are not so clear as in Figure 3 and Figure 4. As mentioned before, cells can have different frequencies and sizes. They can also overlap to produce better coverage. Cells with different frequencies do not form cellular-like network, because it is not unusual that on the same tower two directional BTSs with different frequencies or technologies are directed in the same direction. When taking into account all the frequencies and technologies that are used in building the mobile network, picture would be more as seen in Figure 5, where each technology with its working frequency (e.g. GSM900, UMTS2100) are shown using different color for each different technology.

Mobile Positioning Data. Collecting people location information by surveys and observation is too resource-consuming work, a cheaper method for data with bigger sample size would be mobile data. That is considered most promising data source for measuring the mobility of people [Tiru, 2014]. Most people in the world has a mobile phone. There are 5 billion subscribers to mobile networks. Subscriptions devices itself has reached 7.4 billion. Which means that there are more devices connected to the mobile network than there are people in the world. And that number is growing around 3 percent per year. That number is so large due to inactive subscriptions, multiple device ownership and subscriptions for a different type of calls [Ericsson, 2016].

Mobile devices can be used for collecting their location data with mobile positioning. The mobile positioning means that it is possible to locate devices in time and space with certain accuracy using technology (e.g. mobile network infrastructure or Global Positioning System (GPS)) to get the location of the mobile device. That data about the location and movement of mobile devices is called mobile positioning data.

There are three methods for collecting data.

1. GPS and Assisted GPS (A-GPS)

2. Wireless network (WiFi) location databases

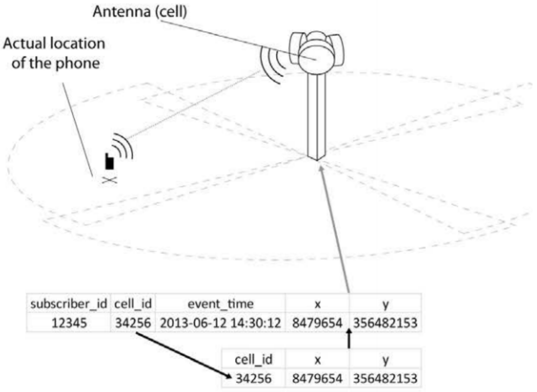
3. Network antenna-based location database

Assisted GPS is used to speed up first GPS location calculation. Usually using WiFi and network antenna-based location databases to pinpoint device location before getting a more precise reading with GPS [Waadt et al., 2010]. WiFi and network antenna-based location databases use information about where the given network antenna or WiFi transmitter geographically is located to predict mobile device location. WiFi and network antenna-based locations are used when there is not possible to use GPS location. As for GPS to work, there must be line-of-sight to at least three GPS satellite. So GPS do not work indoors and in the "urban canyons" created by buildings on either side of a street [Mountain and Raper, 2001].

Network antenna-based location database system with common name Mobile Positioning System (MPS) is used by MNOs to pinpoint the location of the mobile phones using network infrastructure. There are different methods to locate a mobile device [Tiru, 2014].

- Cell Global Identity (CGI)
- Trilateration of antenna to device and back
- Angle of received signal
- Timing Advance (TA - radio signal arrival time from antenna to device and back)

A mobile device knows to which cell it is connected and what is that cell's CGI. From MNO side they can see to which cell the mobile device is connected and what is the cell's identification. From that MNO can look what is that cell's geographical location and know that the mobile device must be somewhere near the cell's location as seen in Figure 6. Trilateration is three sites to pinpoint the location of a mobile device. The angle of received signal uses an array of smart antennas that help to determine the angle of the incoming radio signal. Then it is possible to triangulate known signal angles from at least two base stations. Timing Advance takes into account time needed for the signal to reach mobile device and back to cell station [Ratti et al., 2006].



subscriber_id	cell_id	event_time	x	y
12345	34256	2013-06-12 14:30:12	8479654	356482153

cell_id	x	y
34256	8479654	356482153

Figure 6: Phone location positioning using cell coordinates[Ahas et al., 2014].

Most commonly CGI+TA+angle of received signal is used, where cell's location is first found after cell's identifier and after that TA is used for calculating how far the mobile device is from the cell tower and direction of the device is determined after the angle on received signal [Tiru, 2014].

Mobile positioning data can be collected with two different methods: active positioning and passive positioning.

Active positioning. Active positioning needs active request and usually requires authorization from device owner. When the positioning request is made then device owner has to give the consent for the request. Most people have seen location request pop-ups on applications and map-related websites or user had to approve location request when installing the application . There are also requests made without user consent. User authorization is not needed if location request is done in the case of emergency. That usually means using MPS, but from July 2016 Estonia and the United Kingdom have implemented the Active Mobile Location (AML) system. AML is emergency call-based location solution in Android phones. When smartphone recognizes that an emergency call is made, then phone activates its location services and sends that info to the emergency services automatically before turning the location services off again [EENA, 2016].

Most usual active positioning is installed applications ability to use the location of the mobile device and positioning work real-time or near-real time. Location Active positioning has beneficiaries from different user groups [Ratti et al., 2006].

Active positioning applications can be divided by their target group of users.

- Services for individual users
  - Navigation aids
  - Geographically distributed yellow pages
  - Educational services
- Services for group of users
  - Distributed chats and friend tracking
  - Location-based gaming
  - Traffic services
  - Digital tapestries
  - Coordinated actions
- Services for third parties
  - Public safety and security
  - Family security
  - Emergency relief
  - Business safety and efficiency
  - Commercial and information services
  - Location sensitive billing
  - Urban Systems mapping

There are different applications that are used where location positioning is the main intention, but there are other applications where location data is still collected for value-added services or statistics. Social media applications (e.g. Twitter and Facebook) do not always need location data, but the data is still collected [Tiru, 2014].

There are advantages and disadvantages to active positioning compared to passive positioning. Just like with in person, on-line or telephone surveys, active participation could be requested and sampling techniques could be used. Data collection is easy because participants only need to install positioning application to their phone or approve the periodical network-based location requests. The disadvantage is the need to recruit the respondents, resulting small sample size [Tiru, 2014].

Passive positioning. The difference in a passive and active positioning is, that passive positioning does not need to send a request to a mobile device, to position it. Instead of actively making position request, data is already collected to some database by the MNOs or applications. This data usually is not collected with the purpose to collect the location. Applications and MNOs are logging mobile devices activities and these logs can have data that could be linked to a location [Tiru, 2014]. This data is used mostly internally for business and marketing purposes (e.g. billing clients, providing statistics, marketing). Location data is used seldom [Ahas et al., 2011].

Passive mobile positioning data is mostly referred to the data from MNOs. MNOs are collecting a lot of business purposed data, among them are Call Detail Record (CDR). One CDR contains a lot of technical information about the recorded activity, but the important fields are subscriber identification, time of the event and identification of the antenna, where the mobile device was connected at the time of the activity. CDRs include in and out calls, SMS and MMS messages activities. If mobile devices use also Internet traffic, then these activities are also included in CDR. MNOs can choose what data they want to store and because of that data about mobile devices locations and how precise could location detection be may differ. The size of the data may also vary, meaning that MNOs may store only 2-3 CDRs or even 200-300 CDRs per mobile device per day.

People carry their mobile devices always with them, so CDR data gives an overview of human movements based where they are using their phones. MNOs have also extra info about their subscribers, there may be also access to additional data about the socio-demographic profile (e.g. gender, age) for each mobile device owner.

The main advantages compared to active positioning is the cost-effectiveness of collecting a large sample of data for all the phone users. With passive data collecting, there is no burden on the respondents because data is collected automatically [Ahas et al., 2011]. Access to that data is more difficult. That is because of privacy. People do not like that other people get access to that sensitive data like their location history. At the same time, MNOs do not like to share their data on business related reasons. If MNO would give their CDR data for research, then it would still not contain the full sample of the whole country. One country can have more than one MNO providing their services (e.g. Estonia have three main MNO companies: Telia, Elisa, Tele2). If selecting only one company then sample contains only their clients and mobile devices that uses their cellular network [Tiru, 2014]. Another problem is the CDR data itself. Even if the data size is very large, data accuracy itself is not as good as with GPS collected data. In practice, most CDRs do not have enough information to use more precise positioning methods like using TA or angle of signal arrival because they need more than one base station, but only the base station that is carrying mobile communications are recorded in CDRs [Zang et al., 2010].

Data privacy. A big problem with Location Based Services (LBS) is that people do not like that their movements can be tracked. To encounter that problem there are two solutions. One is asking each person's permission to track them for research purposes. But this means small sample size because you have to interact with each one. A second way is to anonymize the data. That means there is not possible to connect the data with the person whose data it is. Ratti et al. [Ratti et al., 2006] shows the European Union directives, that ensures people privacy rights. There are very precise cases when service providers could work with people's privacy data and when they can give it to third parties. With the European Union directives they can only do that when they have the consent of the service user and the user is informed of the purpose and duration of the process. Another way is to analyze and to give that data to the third party to analyze it, is to anonymize it. Most researchers working with MNOs mobile positioning data are using anonymized and aggregated data.

Datasets are anonymized to bypass the problem of privacy, so researchers can not track data back to the user to whom that data belongs. Researches like [Ahas et al., 2007, Ratti et al., 2006, Ahas et al., 2011] have to use it to work with the data without the breach of privacy. Ahas et al. [Ahas et al., 2011] and Tiru [Tiru, 2014] bring out the problem, that anonymizing could lose the different aspects of the data, that could be useful to research. Wicker [Wicker, 2012] brings out that anonymizing must be done correctly because with location traces could be de-anonymized through correlation with publicly available databases.

Besides the people's data privacy, another big problem is MNOs business secrets. They do not want to share their cell tower locations and information to others. In this thesis users and cell towers identifications and are changed because of privacy issues. Cell towers real geographical locations are also hidden.

1.3.2 Related work

Passive mobile positioning with CDRs has been used for some time.

Articles [Mountain and Raper, 2001] and [Ahas et al., 2007] used it to research seasonality of foreign tourists space consumption in Estonia. Article [Ratti et al., 2006] used it to analyze urban activities in Milan, Italy. Article [Saluveer and Ahas, 2014] describes how passive mobile positioning can be used to find peoples home and work location. But it also mentions that there are accuracy problems and discusses what kind of extra information should be added to MNOs CDR data, to get the more precise location prediction (i.e.. take into account that ME location probability is higher in the places where there are roads, point of interests or buildings.)

Article [Vajakas et al., 2015] used passive mobile positioning with CDR data to reconstruct people movement trajectories.

Using Voronoi with clustering close BTSs together was used as a cell plan in [Csáji et al., 2013]. They used Voronoi-based cell plan with only omnidirectional cells. They took into account that ME may be in the other cell area not in the one, it is connected to. They proposed to use BSD information instead of simple Voronoi. They also used Maximum Likelihood Estimation (MLE) for estimating the position of frequent locations.

Article [Zang et al., 2010] described the methodology applying Bayesian methods for defining the Spatial Probability Density Functions (SPDF) for mobile positioning. Estimation was based on Signal-to-Interference-and-Noise Ratio (SINR) calculations. Bayesian probability estimate was provided for the situations where CGI was the only information to know and no other cells had good enough SINR. The method was tested on the subset of emergency call data, that was limited to situations where only one cell was within the reach of the BTS. They found that the difference between the location measured by the GPS and the MLE provided by that method was improved by 20%, compared to the baseline method. No direct Probability Density Functions (PDF) tests were performed and they did not consider that the probability to connect to a specific cell will be reduced where other cells are reachable.

Unpublished paper [Vajakas and Vajakas, 2014] created SPDF estimations with Bayesian rule method. Formulas were created that took into account that probability to connect to a concrete cell will be reduced in locations where other cells are reachable, but the paper was lacking in real world tests.

1.4 Contributions of this work

This thesis concentrates on algorithms on how to position mobile devices from CDR information. Different cell plans creation methods are described and tested with mobile device GPS location and connected cell CGI data.

Passive mobile positioning SPDF estimations are used for comparing quality of the cell plans. This thesis continues the work [Vajakas and Vajakas, 2014] which started the research on SPDF estimations with Bayesian rule. Important part of the article is added to the appendix C. SPDF estimations methods with and without Bayesian rule are used to verify if the cell plan overlapping is important.

The theoretical foundation of SPDF and application of Bayesian rule were devised by Toivo Vajakas (original ideas) and Jaan Vajakas (ensured mathematical correctness).

The author applied and adapted the theoretical base to concrete technological stack in Reach-U: devised conversion algorithms for concrete cell plan inputs, implemented the algorithms for cell plan data processing and evaluation (except the code described in appendix C), analyzed the results.

Paper [Vajakas and Rõõmusaare, 2016][1] will be published based on the theory of SPDF estimations with Bayesian rule and the results of this thesis. Paper that was sent to the review is included in the Appendix D.

[1] Jaan Vajakas was not included as an author, because as a Google Inc employee, he did not have time to get permission from Google Inc.

2 Methods

2.1 Description of the data

2.1.1 User data

There were five test users with the identifications 1,2,3,5 and 6. Data collection time period was almost 8 months: 01.04.2015 - 18.11.2015. User 4 only collected the YouSense data for the first month, and because of that this users data was discarded.

YouSense data. GPS and connected cell information were collected from our test users with android application YouSense, an active positioning application created by the company OÜ Positium LBS.

Example collected data is seen in Table 1. The timestamp is in Unix epoch milliseconds. the application collected GPS locations every 1 second if ME was moving faster than 3 m/s and every 16 seconds if it was moving slower than that. The application turns GPS tracking off when ME moves less than 20 meters in 4 minutes or GPS fix could not be obtained for 60 seconds. When ME starts moving again, GPS tracking will be turn on. There may be caps in the collecting when the ME do not have its GPS turned on.

Figure 7 shows how many GPS events each user collected in the whole period and how many events each month.

userid	cgi	lon	lat	timestamp	accuracy	altitude	bearing	speed
3	248-xx-xx-xxxxxx	26.68574	58.37747	1428003372860	136.0	92.0	NaN	0.0
3	248-xx-xx-xxxxxx	26.6848	58.37752	1428003424837	114.0	43.0	NaN	0.0
3	248-xx-xx-xxxxxx	26.68529	58.3776	1428045720961	147.0	77.0	NaN	NaN
3	248-xx-xx-xxxxxx	26.68582	58.3772	1428045942031	67.0	28.0	124.5	0.75

Table 1: YouSense GPS data example.

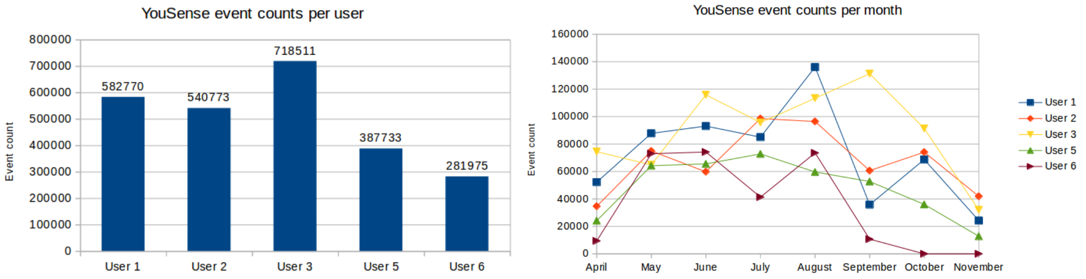


Figure 7: Users' YouSense event counts per user and per user per month.

CDR data. With our test users permission, we also collected their CDR data from MNO for the same period as GPS data. CGI values are used for positioning from CDR data because MNOs do not always have better CDR data for passive mobile positioning than just a CGI value. The example of the data is seen in Table 2. the timestamp is in Unix epoch seconds. CDR contains events with different types, this is given in the event column. They could be call, SMS, location update, etc. This thesis does not take event type into account for positioning.

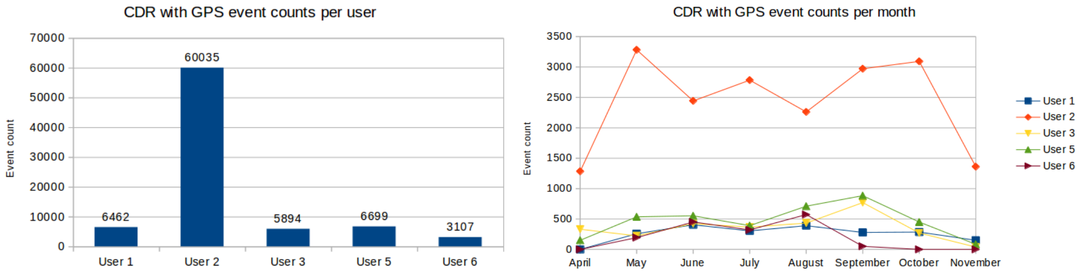


Figure 8: Users' CDR event counts per user and per user per month.

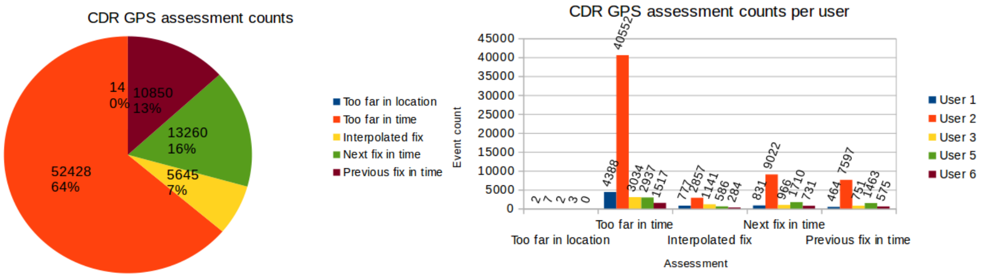


Figure 9: CDR GPS counts overview per assessments and per user per assessments.

Figure 8 shows how many CDR events each user made in the whole period and how many events each month.

userid	timestamp	MCC	MNC	LAC	CI	Event type
3	1427856959	248	xx	xx	xxxxxx	85
3	1427900159	248	xx	xx	xxxxxx	85
3	1427980327	248	xx	xx	xxxxxx	18
3	1427980332	248	xx	xx	xxxxxx	85

Table 2: CDR data example.

Combining CDR events with GPS data. A program was created to merge CDR and GPS events. It reads in CDR events and uses GPS data to find the ME's real geographical location. For each CDR event, GPS event fix will be created based on the GPS events before and after the CDR event time. GPS location selection process is described with Algorithm 1. Each selected GPS event fix has an assessment for how precise that GPS could be. Overview of the result is given in Figure 9.

GPS fixes points locations are showed in the Appendix A. Figure A.1 shows location points over Estonia and Figure A.2 shows close up in Tartu, where the majority of events were made. There are big lines from yellow dots with assessment "Too far in time" that show how unrealistic these measurements are. CDR events with GPS fixes with assessment "Too far in time" and "Too far in location" were discarded. That means from all the events there are 37370 useful CDR events with GPS fixes.

Algorithm 1 CDR and GPS events merging algorithm.

```
For each user:

• Remove CDR events that are outside of the GPS events time frame, because extrapolation is not supported.

• For each remaining CDR event :

    – Use binary search over time-sorted GPS events to find the GPS event with the smallest time that is
      larger or equal to the CDR event time.
    – Found GPS event will be known as the next fix.
    – The GPS fix before the next fix is known as the previous fix.
        ∗ If the next fix is the first recorded GPS event, then it will be used as the previous fix and the
          next GPS event after that will be the next fix.
    – Calculate interpolated GPS fix.
        ∗ The interpolated GPS fix values are a weighted average of the previous and the next fix with
          the time of the CDR event.
            · Each GPS fix weight is time difference from the GPS fix time to the CDR event time.
    – Select assessment for the interpolated GPS fix.
        ∗ If the previous and the next fix are farther than 15 minutes from the CDR event time, then
          assessment is “Too far in time”.
        ∗ If the previous and the next fix are both closer than 10 seconds from the CDR event, then:
            · if both fixes are farther from each other than 100m then “Too far in location” is chosen.
            · if both fixes are closer than 100m, then “interpolated fix” is chosen.
        ∗ if the previous and the next fix are closer than 15 minutes from the CDR event, then:
            · if the previous fix is closer than the next fix, then “previous fix” is chosen.
            · if the next fix is closer than the previous fix, then “next fix” is chosen.
    – Select GPS fix based on the assessment.
        ∗ With assessments “interpolated fix”, “Too far in location” and “Too far in time”, interpolated
          GPS fix is chosen.
        ∗ With assessments “previous fix”, previous GPS fix is chosen.
        ∗ With assessments “next fix”, next GPS fix is chosen.
```

2.2 Cell plan creation techniques

Cell plan can be generated using multiple different approaches.

What method to use depends on the cell plan's purpose. If it would be used for analyzing network coverages, then cell plan that includes cell's fringe areas could be better. If it would be used for showing the area where the ME more likely was when it was connected to that cell, then smaller cell size would give a better location estimation accuracy. This thesis uses cell plans for positioning ME locations.

Cell plan generation methods in this thesis and the implementation of them are created for application Demograft. Demograft is a tool for MNOs that uses passive location data of their customer base to help them analyze their network. It also helps them selecting subscribers for advertising campaigns. Demograft was developed by Reach-U [Reach-U, 2016] and STACC [STACC, 2016].

When the final intention of a cell plan is known, then generation depends on the available cell data. Simplest data would be cells identifications and their geographical locations. This will give the most inaccurate cell plan because as there is not any information about the direction of the cells. So all the cells could be generated as omnidirectional and the cells on the same site would have the exact same cell geometry. If data includes direction of the antenna, then directional cells could be generated and cells on the same site could each have the cell geometry in the correct direction. If frequency or technology is included, then each different frequency and technology could have cell plans on separate layers. More cell attributes help to generate more precise cell shapes. This information is usually given in a Comma Separated Value (CSV) or an Extensible Markup Language (XML) file.

When MNOs have the BSD files or RSSI data and they are prepared to share them, then this data could be used for generating more real-world-like cell plans. To generate cell plans from different inputs Cell Plan Calculator was created for Demograft application. This chapter describes the methods used by Cell Plan Calculator to create cell plans from different input data.

2.2.1 CSV file input

The basic information for each cell is usually given by the MNOs with simple CSV or XML file. Most CSV or XML files contain cells geographical points(cell site coordinates) with descriptive metadata. As cell plans are made of geographical coverage areas, then we have to generate areas from given cell location point and its metadata. Minimum requirement information for each cell are location coordinates and a cell's CGI, but to get realistic cell shapes, then direction and frequency is also needed. Basic cell table is shown in Table 3.

Longitude and latitude coordinates for the cell are needed for the site location, without it, there is not possible to know where the cell should be located on the map.

Direction shows in where the cell is pointed. Missing direction value should show that given cell is omnidirectional cell and does not have direction (in the Table 3 shown as “null”) . If the direction is missing for whole data, then we can only assume that every cell in the given data is an omnidirectional cell.

The frequency for each cell is important because cells in different frequency band should be separated from each other. If frequency info is missing, then we have to assume that every cell is in the same frequency. Sometimes instead of frequency column, each frequency is given in a separate file.

CGI	latitude	longitude	direction	frequency
248-00-01-121	59.568150	24.301561	0	LTE1800
248-00-01-122	59.568150	24.301561	120	LTE1800
248-00-01-123	59.568150	24.301561	240	LTE1800
248-00-01-131	59.175251	24.850207	null	UMTS2100

Table 3: Cells basic CSV example.

These requirements are minimal but may not be sufficient for a more precise cell plan. There may be more information given to each cell. If a cell radius is also given, it can be used for not letting the cell spread top far. With a cell type, we can tell how big cell area should be and even could help for deciding if a cell is omnidirectional or not. Normally, the direction field should be enough for knowing if a cell is directional or not, but the practice has shown that MNOs cell data is not always perfect and having more information about cell's attributes, helps to make better decisions for creating cell area for each cell.

Using the basic information, there are three ways to generate cell plan. Each method will create cell plans with different geometry shape.

Voronoi cell plan. The Voronoi method is used for generating cell plan that does not have overlaps between cells in the same frequency. This creates good areas for mobile positioning but does not take into account that in reality cells will overlap. This can be fixed by enlarging each cell after generating Voronoi cells.

This method is a fast way to divide whole coverage between each cell, but may create some irregular shapes, Figure 10 shows big cell number 3836 that is narrow in the west-east direction, but the cell direction is almost to the north.

Voronoi cell plan creation algorithm is described in Algorithm 2.

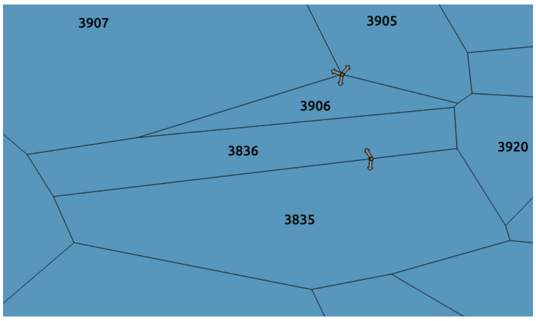


Figure 10: Voronoi cell plan with gray arrows as cell's direction and brown dots as cell sites. A questionable shape of the cell 3836 is depicted.

Algorithm 2 Voronoi cell creation algorithm.

```
For each BTS technology and frequency do following:
  • Separate omnidirectional BTS from directional BTS.
  • Create omnidirectional BTS cell hexagon geometry based on BTS cell's radius.
  • Cluster directional BTS that are close to each other together.
      – This is used because MNO does not always site information for BTS.
  • Cluster center point will be cell site for each BTS in that cluster.
  • Each BTS location will be moved a little in the direction of that BTS cell direction.
      – This will force Voronoi diagram builder to build cells for each BTS in the correct direction (one cell
        site will have correct sectors).
  • Add moved BTS locations to Voronoi diagram builder.
  • Build Voronoi diagram
  • For each BTS:
      – Find geometry in Voronoi diagram that has BTS moved location inside it.
          * That geometry will be that BTS cell.
          * Cut cell geometry with that hexagon with BTS maximum radius.
      – Set BTS site location back to its original location.
```

2.2.2 CSV and best serving data files input

MNOs provide sometimes so-called "best serving data" in raster format. Each frequency cells are in separate layers. BSD example was shown in Figure 3. For working with each cell geometry, these files must be first converted from raster to vector format. This is done with ogr2ogr tool [ogr2ogr, 2016]. As with Voronoi method, there are no overlaps between cells in the same frequency. Again cell enlargement modification could be added for each cell after creating the cell plan.

CSV files are still used for extra cell plan info, but the geographical areas are taken from the BSD raster files.

2.2.3 CSV and RSSI files input

As with BSD data, CSV files are used for all the extra cell attributes, but geographical data is taken from the RSSI data. RSSI data is the biggest in data volume. Each cell in the network has very precise geometries where each geometry has RSSI what value that cell should have in that geographical location. Figure 11 shows one cell RSSI values on the map. MNO holds this data in databases and it needs a little bit more preprocessing before Cell Plan Calculator could do anything with it. First signal strengths period is selected, usually, period -80...-83dBm is selected. Per each cell all the geometries that have the RSSI value in selected period would be downloaded from the database and the convex hull algorithm will be applied to these geometries. The resulting geometry is polygon where all the selected geometries are inside of it. Shapefile is created from these polygons and that is used like in the BSD method.

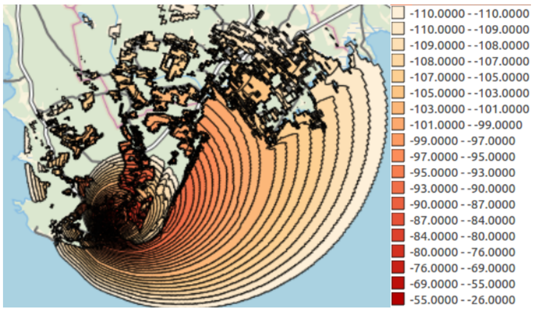


Figure 11: RSSI cell with its dBm values example near water.

Each cell area was selected according to its RSSI value. Unlike BSD, the cells derived from RSSI data can overlap in the frequency band. This method could leave some uncovered "holes" in the whole cell plan coverage - there may be areas where each cell RSSI value is lower than selected period, this could be improved with selecting RSSI period with lower decibels.

2.2.4 Cell Plan Calculator

Cell plan Calculator takes in a configuration file in JavaScript Object Notation (JSON) format. Everything necessary for its work is described there, but extra flags can be added to change calculator behavior. There are flags for changing the output file location for easier automation scripts. MNOs are changing their network day-to-day to adjust to subscribers needs and because of that cell plans need to be up to date with real world situation.

Cell plan generator was built to work in five steps. Generation could skip some steps if they are not needed for some specific MNO.

Parsing CSV files. First step is parsing CSV(or XML) input files. This step will read in one or more CSV files. Each file can be described separately to the calculator in the configuration file. As MNOs configurations and cell data files may vary then configuration file is made to be very flexible. A configuration file example is in Appendix B. The configuration file is used for reading CSV files in and then converting them to a specific form that all the next step can use.

Parsing rules. Rules are used for configuring two things: cell's radius and its boolean value with true if a cell is omnidirectional and false if it is directional. It is used when input file does not have a radius or omnidirectional attribute info. Rules are described in a separate file, which location is given in configuration file. An example of the rules file is given in Table 4.

```
radius|isOmni|expression
50|0|cell_type == "MICRO"
10000|1|frequency == "GSM900"
6000|1|frequency == "GSM1800"
10000|1|true
```

Table 4: Cell Plan Calculator input rules file example.

Rules file contains three columns, separated by a pipe character. The first column has a resulting radius. Radius is used for limiting cell areas from reaching too far from the cell site. It is used when generating omnidirectional cells: a radius is used for polygons radius and if Voronoi cell is very large, then it is also trimmed with radius. Radius depends on different things, for example, cells with different frequencies have different reaches. The second column has resulting omni cell boolean value. If the boolean value is true, then Cell Plan Calculator will generate that cell as omnidirectional. Otherwise, it will be counted as a directional cell. The last column holds the expression for given rule.In expressions, conditionals are described for the cells. if they match with cell's attribute, then selected radius and isOmni parameter values will be chosen to that cell.

The tool will start testing expressions from the first one and if expression fails, then will be moving to the next one. If successful validation has been found, then that radius and omni cell values will be used. e.g if the first expression is successful, then it would not try the next ones. The last one in the list should be the default value if other expressions have failed.

Expressions are written in MVEL [MVEL, 2016] and its comparison and logical operators can be used for describing rules. All cell attribute names, which are not removed from input data with configuration, can be used as variables in all three fields. Variables are case sensitive.

Radius, omnidirectional boolean value, and their conditions are selected with MNO representatives because it is unique with each MNO.

Generating first geometry using Voronoi generation method. After parsing rules, geometry is generated for each cell. Geometry shape style is given in properties.

Replacing first geometry if needed, using BSD or RSSI methods. After first geometry generation, cell's geometries that are given in BSD or RSSI input files will be overwritten. If both are given, then first geometries will be overwritten with BSD geometries and after that with RSSI geometries.

Add additional modifications to the generated cells. There are multiple modifications built for the modifying whole cell plan.

- Cut cell - Will cut the whole cell plan with given geometry. Leaves in only the cells that intersect with input geometry. Good for cases where you want to work with only one part of the cell plan

- Enlarge cell - Will enlarge each cell in a cell plan. Enlargement takes in buffer ratio parameter. The buffer will be added to the cell with the radius calculated with this formula: buffer ratio to the square root of the cell area.

- Stretch cell - Each cell in a cell plan will be stretched in cell's direction and perpendicular direction by direction multiplier and perpendicular direction multiplier.

- Convex Hull cell -Takes Convex Hull over cell and cell site for each cell in cell plan. When using RSSI generated geometries, sometimes cell geometry may not contain cell site itself. Convex Hull will add cell site to cell geometry.

- Ellipse cell - Replaces each cell's geometry with ellipse for each cell in cell plan. Ellipse is an approximation of the original cell geometry.

2.3 Cell plan quality estimation with maximum likelihood (MLE)

MLE principal can be used to compare the quality of different PDF estimates. Higher likelihood value indicates that the PDF is more precise estimate to the real probability distribution what generated the measured data.

Likelihood is defined as:

$$L = \mathrm{P}(measurements|model). \tag{1}$$

If we consider simplifying assumption that measurements are independent, then we can apply simplified formula:

$$L = \mathrm{P}(measurements|model) = \prod_i \mathrm{P}(x_i|\theta), \tag{2}$$

where

- θ is a vector of parameters;
- x_i are individual measurements.

In practice it is often more convenient to work with the logarithm of the likelihood function, called log-likelihood:

$$\log\, L = \log\, \mathrm{P}(measurements|model) = \log \prod_i \mathrm{P}(x_i|\theta) = \sum_i \log\, \mathrm{P}(x_i|\theta). \tag{3}$$

For applying log-likelihood to cell plan and mobile measurements, formula would be like this:

$$\log \mathcal{L} = \log \mathrm{P}(\mathrm{measurements}|\mathrm{model}) = \log \prod_i \mathrm{P}(x_i|C_i) = \sum_i \log \mathrm{P}(x_i|C_i). \tag{4}$$

where:

- x_i, $i = 1...m$ are mobile events actual locations;
- C_i, $i = 1...m$ are cells that event x_i was connected to;
- models are the SDPF cell plans created with methods in appendix C .

The implementation for changing the cell plan polygons into an SPDF was taken from the paper [Vajakas and Vajakas, 2014] - applied Gaussian blur with kernel proportional to square root of the cell polygon area. SPDF estimation and its implementation is described in appendix C because it is necessary part of the thesis and the presentation is not published.

The reason to use continuous SPDF is instead of practically used polygons is that MLE needs a function that does not equal to zero at any location, or else likelihood value will be zero due to single outlier in data. Small probability was assigned to outliers so that each ME still has an probability to be connected to the cell even when it is not near its area of coverage.

Based on the SDPF implementation in the C.1, conditional probability was calculated to visualize Bayes' rule influence to each cell. For each pixel, following probabilities were calculated:

$$\mathrm{P}(C_i|x) = \frac{\mathrm{P}(x|C_i)}{\mathrm{P}(x|C_k)}, \tag{5}$$

where

- C_i and C_k is the cell connected from location x;
- x is the ME location.

Formula 5 shows cell connectivity probability for each pixel.

In the Formula 5 we assume that probability to connect to the network is 100%, i.e:

$$\sum_i \mathrm{P}(C_i|x) = 1. \tag{6}$$

Formula 6 means that if ME is at the location x then it must be connected to a cell.

3 Results

The test area was divided into pixels with the length of 630 meters. The size was selected so it would be large enough that many pixels would contain multiple experimental GPS data location points. Raster $\mathrm{P}'(C|x)$ were calculated for all the cell plan variants. Cell plan variants were:

- rssi_default - cell plan based on RSSI data, extra modifications were not made.
- rssi_A1AP2 - rssi_default cell plan with “stretch cell” modification. Cell’s width is stretched twice.
- rssi_A2AP1 - rssi_default cell plan with “stretch cell” modification. The cell is stretched twice along the direction of the cell.
- rssi_A3AP1 - rssi_default cell plan with “stretch cell” modification. The cell is stretched along the direction of the cell three times.
- rssi_convexhull - rssi_default cell plan with “Convex Hull cell” modification.
- rssi_ellipse - rssi_default cell plan with “ellipse cell” modification.
- voronoi_default - cell plan created with Voronoi method, extra modifications were not made.
- voronoi_A1AP2 - Voronoi cell plan with “stretch cell” modification. Cell’s width is stretched twice.
- voronoi_A2AP1 - Voronoi cell plan with “stretch cell” modification. The cell is stretched twice along the direction of the cell.
- voronoi_A3AP1 - Voronoi cell plan with “stretch cell” modification. The cell is stretched along the direction of the cell three times.
- voronoi_convexhull – Voronoi cell plan with “Convex Hull cell” modification.

Formula 4 was used for calculating the likelihood of the positioning data with each SPDF raster variant. MLE results are shown in Figure 12. Horizontal axis shows different users and vertical axis shows average $\mathrm{P}(C|x)$ for CDR data of the given user. Results produced with the same processing parameters are connected with a line for easier comparison of methods.

Figure 13 shows the SPDF of all the cells along a sequential straight line of pixels. Each pixel has stacked probabilities of the cells. Upper chart is generated with applying Bayesian overlapping cell model, lower chart data is generated without considering overlapping effects.

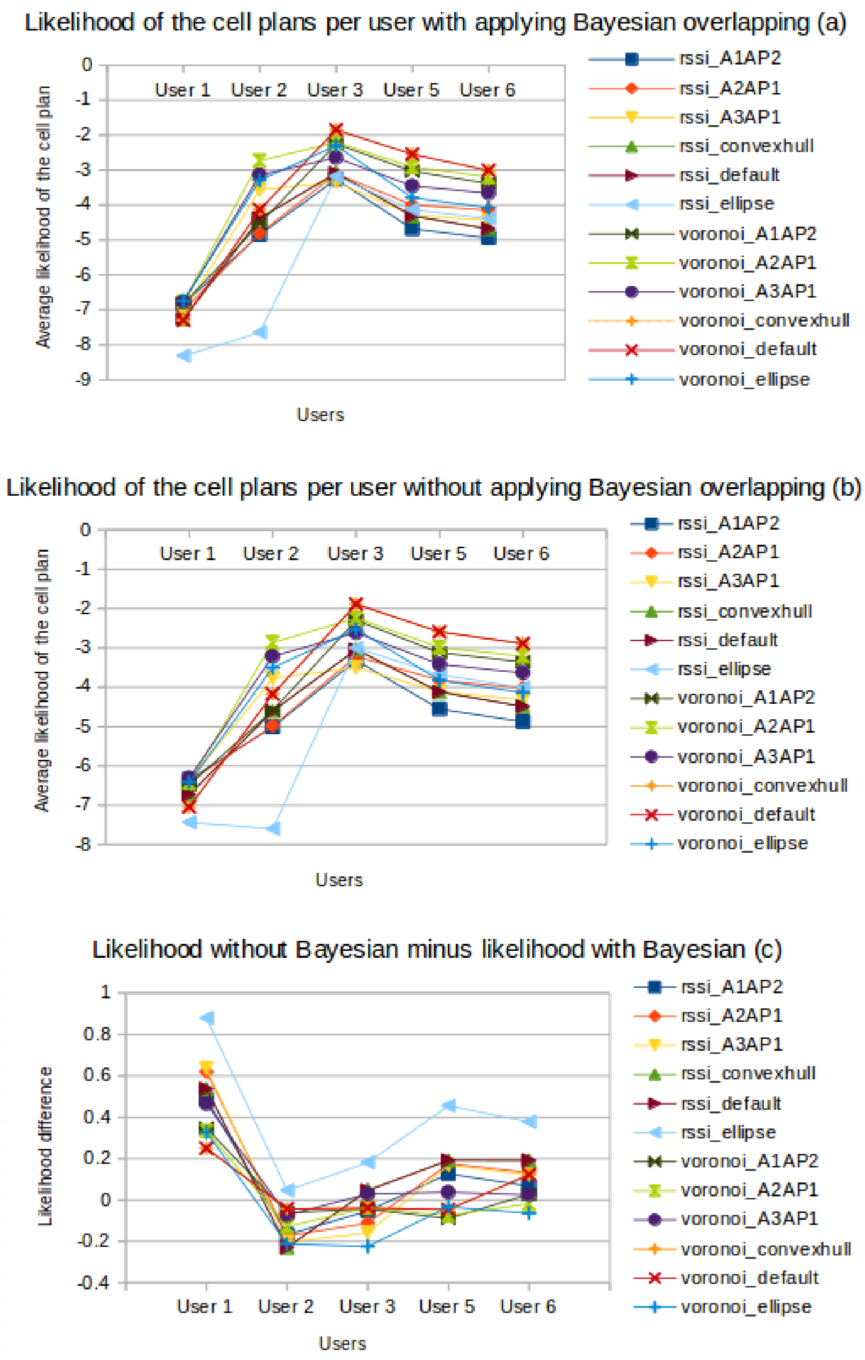


Figure 12: Relative performance of the various cell plan variants: a) contains cell plans with Bayesian overlapping; b) cell plans without Bayesian overlapping. c) shows effect of the Bayesian overlapping.

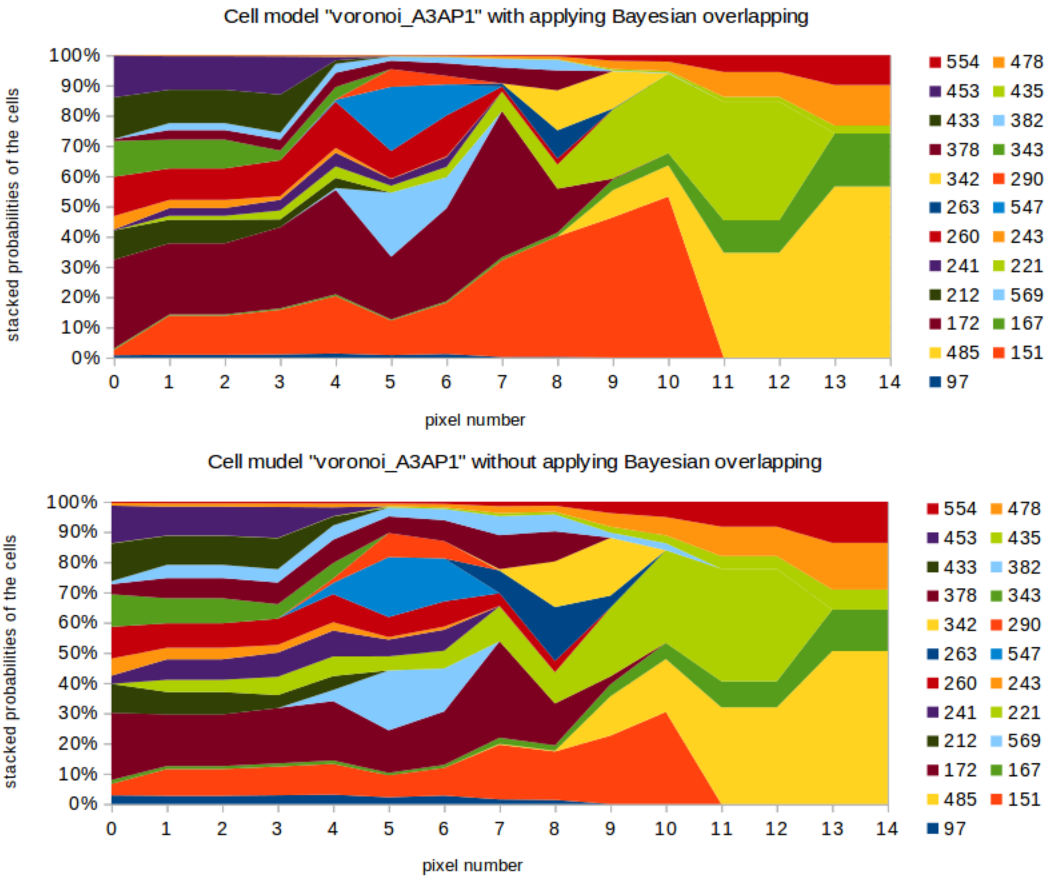


Figure 13: Cells SPDF for sequential line of pixels with and without using Bayesian overlapping cell model, calculated using Formula 5.

4 Conclusions

For this thesis, Cell Plan Calculator was designed and programmed. It was designed to be flexible in order to be able to use the variety of input data. Cell plans were created for passive mobile positioning. Toivo Vajakas and Jaan Vajakas work on SPDF was used to test which cell plan is better for estimating mobile locations. Test data was collected from MNO and test users. Prior SPDF code was used for creating test cases for collected data. Modifications were made to SPDF code to create the stacked probabilities of each cell in each pixel. Results were visualized and analyzed.

4.1 Discussion

When using different observation data, then MLE values are not comparable. Therefore the relative performance of different models can be only evaluated for each user data separately.

Surprisingly all the Voronoi-based cell plans seem to have better SPDF estimates than the RSSI-based. The expectations were that sophisticated RSSI models would be superior to simple model like Voronoi model.

Cell plan “rssi_ellipse” seems to be most inaccurate based on the SPDF in almost most of the user data, this should be certainly avoided in practice.

There are notable differences in the ranking of the cell plan variants between users.

Figure 13 shows how Bayes’ rule changes the likelihood estimations. The results show that accounting for cell overlap effect with Bayes’ rule had in the majority of cases positive effect.

Figure 12 (c) shows that the best cell plans are not improved by Bayes’ rule and worse cell plans are significantly improved.

In Figure 12 “voronoi_convexhull” and “voronoi_default” are practically identical as these methods are internally the same.

4.2 Future plans

Research can be continued in multiple directions.

One way is to apply the algorithms to larger data sets. One might recruit a bigger group of data gatherers for active data collecting, or use crowd-sourcing data. For example, OpenCellId [ope, 2016] collects data with GPS and connected cell’s CGI from volunteers. OpenCellId database has 145273 measurements collected over our test period (120316 measurements in EMT, 19846 in Elisa and 5111 in Tele2).

For more data, it is possible to not to use CDR data at all. Instead use selected data gatherers or crowd-sourced data, that contain GPS location and the connected BTS information for each GPS location.

To further the research with SPDF, extra calculations could be done with Bayesian prior. For example, one could consider GIS layers of roads and buildings. Probability to encounter ME devices should be higher in near roads and buildings. It is possible to group events not only by users but by other attributes, e.g. separate urban and rural events, to test how it will effect MLE values. Too fine-grain subsets may bring up the too small dataset and over-fitting problems.

CDR event type was not taken into account in our analysis. For example, if event type shows the location update, it might suggest that ME is on the edge of the cell, not somewhere in the middle of the cell.

5 Acknowledgment

Research in the thesis was partially supported by a traffic applications research project in Reach-U Ltd and ELIKO and by Location-Based Big Data Algorithms project in STACC. This research has been supported by the European Union through the European Regional Development Fund.

I would like to thank Reach-U Ltd and STACC for giving me an interesting topic for the master thesis. I would also like to thank Jaan Vajakas for Bayesian formulas I got to use for testing cell plan algorithms quality. I would like to thank also my supervisor Toivo Vajakas, who guided my implementation-centric thesis more into a scientific thesis. Finally, I would like to give big thanks to the data gathering participants without whom I had not had the data to test any methods described in this thesis.

References

[ope, 2016] (2016). Opencellid. `http://opencellid.org/`. [Online; accessed 10-August-2016].

[Agrawal and Zeng, 2010] Agrawal, D. and Zeng, Q. (2010). *Introduction to Wireless and Mobile Systems*. Cengage Learning.

[Ahas et al., 2007] Ahas, R., Aasa, A., Mark, Ü., Pae, T., and Kull, A. (2007). Seasonal tourism spaces in estonia: Case study with mobile positioning data. *Tourism Management*, 28(3):898 – 910.

[Ahas et al., 2014] Ahas, R., Armoogum, J., Esko, S., Ilves, M., Karus, E., Madre, J.-L., Nurmi, O., Potier, F., Schmücker, D., Sonntag, U., and Margus Tiru, M. (2014). Feasibility study on the use of mobile positioning data for tourism statistics. Technical report. [Online; accessed 10-August-2016].

[Ahas et al., 2011] Ahas, R., Tiru, M., Saluveer, E., and Demunter, C. (2011). Mobile telephones and mobile positioning data as source for statistics: Estonian experiences. *presentation for NTTS*.

[Becker et al., 2013] Becker, R., Cáceres, R., Hanson, K., Isaacman, S., Loh, J. M., Martonosi, M., Rowland, J., Urbanek, S., Varshavsky, A., and Volinsky, C. (2013). Human mobility characterization from cellular network data. *Communications of the ACM*, 56(1):74–82.

[Calabrese, 2011] Calabrese, F. (2011). Urban sensing using mobile phone network data. `http://researcher.watson.ibm.com/researcher/files/ie-FCALABRE/Urban%20sensing%20using%20mobile%20phone%20network%20data.pdf`. [Online; accessed 10-August-2016].

[Calabrese et al., 2014] Calabrese, F., Ferrari, L., and Blondel, V. D. (2014). Urban sensing using mobile phone network data: A survey of research. *ACM Comput. Surv.*, 47(2):25:1–25:20.

[COAI, 2016] COAI (2016). Cellular network architecture. `http://www.coai.com/indian-telecom-infocentre/telecom-infrastructurenetworks`. [Online; accessed 10-August-2016].

[Csáji et al., 2013] Csáji, B. C., Browet, A., Traag, V. A., Delvenne, J.-C., Huens, E., Van Dooren, P., Smoreda, Z., and Blondel, V. D. (2013). Exploring the mobility of mobile phone users. *Physica A Statistical Mechanics and its Applications*, 392:1459–1473.

[EENA, 2016] EENA (2016). Advanced mobile location is now available in all android phones! `http://eena.org/press-releases/aml-in-android`. [Online; accessed 10-August-2016].

[Ericsson, 2016] Ericsson (2016). Ericsson mobility report. `https://www.ericsson.com/res/docs/2016/ericsson-mobility-report-2016.pdf`. [Online; accessed 10-August-2016].

[GSMArena, 2016] GSMArena (2016). Network coverage - 2g/3g/4g mobile networks. `http://www.gsmarena.com/network-bands.php3`. [Online; accessed 10-August-2016].

[Kwan et al., 2012] Kwan, M., Cartwright, W., and Arrowsmith, C. (2012). Tracking movements with mobile phone billing data: A case study with publicly-available data. In *Advances in Location-Based Services*, pages 109–117. Springer.

[Mountain and Raper, 2001] Mountain, D. and Raper, J. (2001). Positioning techniques for location based services - characteristics and limitations of proposed solutions. *Aslib Proceedings*, 53(10):404–412.

[MVEL, 2016] MVEL (2016). Mvflex expression language. `https://github.com/mvel/mvel`. [Online; accessed 10-August-2016].

[ogr2ogr, 2016] ogr2ogr (2016). Gdal:ogr2ogr. `http://www.gdal.org/ogr2ogr.html`. [Online; accessed 10-August-2016].

[OpenSignal, 2016] OpenSignal (2016). The state of lte. `http://opensignal.com/reports/2016/02/state-of-lte-q4-2015/`. [Online; accessed 10-August-2016].

[Penttinen, 2015] Penttinen, J. T. (2015). *The Telecommunications Handbook: Engineering Guidelines for Fixed, Mobile and Satellite Systems.* John Wiley & Sons.

[Ratti et al., 2006] Ratti, C., Frenchman, D., Pulselli, R. M., and Williams, S. (2006). Mobile landscapes: Using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design*, 33(5):727–748.

[Reach-U, 2016] Reach-U (2016). Demograft. `http://www.reach-u.com/demograft`. [Online; accessed 10-August-2016].

[Saluveer and Ahas, 2014] Saluveer, E. and Ahas, R. (2014). Using call detail records of mobile network operators for transportation studies. *Mobile Technologies for Activity-Travel Data Collection and Analysis*, page 224.

[Saluveer et al., 2012] Saluveer, E., Silm, S., and Ahas, R. (2012). Theoretical and methodological framework for measuring physical co-presence with mobile positioning databases. In *Advances in Location-Based Services*, pages 247–266. Springer.

[Segan, 2015] Segan, S. (2015). Cdma vs. gsm: What's the difference? `http://www.pcmag.com/article2/0,2817,2407896,00.asp`. [Online; accessed 10-August-2016].

[shareTechnote, 2016] shareTechnote (2016). Lte quick reference. `http://www.sharetechnote.com/html/Handbook_LTE_CGI.html`. [Online; accessed 10-August-2016].

[STACC, 2016] STACC (2016). Demograft. `https://www.stacc.ee/edulood/eelviimane-naide/`. [Online; accessed 10-August-2016].

[Tiru, 2014] Tiru, M. (2014). Overview of the sources and challenges of mobile positioning data for statistics. In *International Conference on Big Data for Official Statistics, Beijing [online][date of reference 6 May 2015]< http://unstats.un.org/unsd/trade/events/2014/beijing/Margus%20Tiru*.

[Tondare et al., 2014] Tondare, S. M., Panchal, S. D., and Kushnure, D. T. (2014). Evolutionary steps from 1g to 4.5g. *International Journal of Advanced Research in Computer and Communication Engineering*, 3(4).

[Vajakas and Rõõmusaare, 2016] Vajakas, T. and Rõõmusaare, J. (2016). On optimal spatial probability density estimation of passive mobile positioning events. Unpublished; Accepted for publication by Baltic Electronics Conference.

[Vajakas and Vajakas, 2014] Vajakas, T. and Vajakas, J. (2014). Optimal estimation of spatial density on mobile positioning data with applications to realtime heatmap visualization. Unpublished;presented in Mobile Tartu 2014.

[Vajakas et al., 2015] Vajakas, T., Vajakas, J., and Lillemets, R. (2015). Trajectory reconstruction from mobile positioning data using cell-to-cell travel time information. *International Journal of Geographical Information Science*, 29(11):1941–1954.

[Waadt et al., 2010] Waadt, A., Bruck, G. H., and Jung, P. (2010). *Positioning Systems and Technologies*, pages 177–211. John Wiley & Sons, Ltd.

[Wicker, 2012] Wicker, S. B. (2012). The loss of location privacy in the cellular age. *Communications of the ACM*, 55(8):60–68.

[Zang et al., 2010] Zang, H., Baccelli, F., and Bolot, J. (2010). Bayesian inference for localization in cellular networks. In *INFOCOM, 2010 Proceedings IEEE*, pages 1–9.

A CDR and GPS fix locations with assessments in Estonia and close up in Tartu

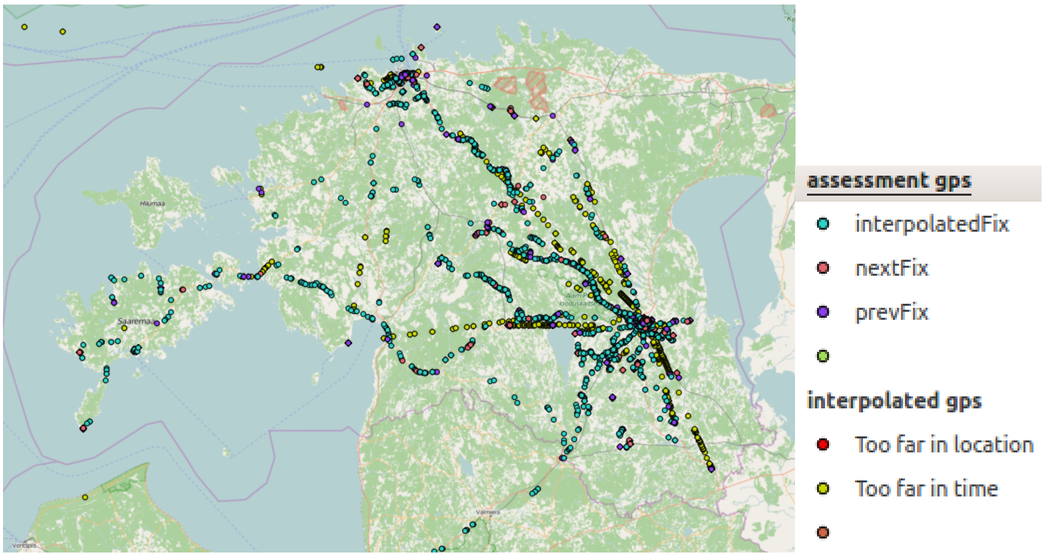


Figure A.1: Users' GPS fix locations with assessments in Estonia.

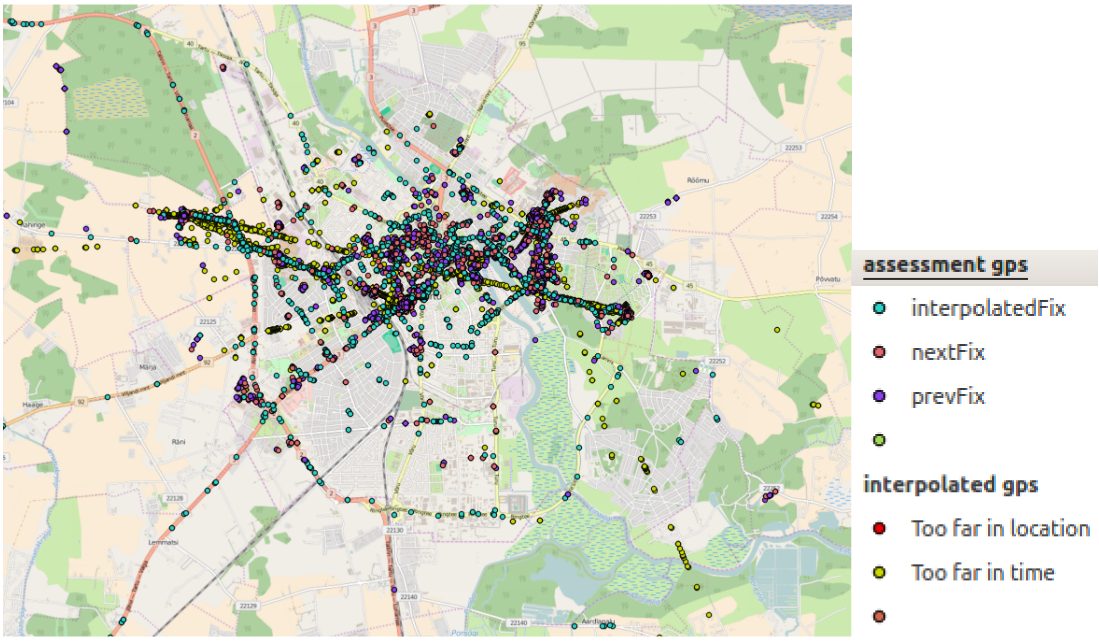


Figure A.2: Users' GPS fix locations with assessments in Tartu.

B Cell Plan Calculator configuration file

```
{
    "rules_location" : "rules.csv",
    "cluster_distance" : 70,
    "stretch_a_direction": 1,
    "stretch_ap_direction": 1,
    "csv_output_path": "cellplan_with_border.csv",
    "type1_cellfiles" : [
        {
            "location" : "celldata_hspa.csv",
            "format" : "csv",
            "separator" : "\|",
            "cgi_field" : "CGI",
                        "technology_generation_value" : "3G",
            "frequency_fields" : "BAND",
            "import_fields" : {
                "lat" : "LATITUDE",
                "lon" : "LONGITUDE",
                "direction" : "AZIMUTH",
                "radius" :"RADIUS",
                "frequency": "BAND"
            },
            "removed_input_fields" : [],
            "coordinate_system" : "DegDec"
        },
                {
                        "location" : "celldata_lte.csv",
                        "format" : "csv",
                        "separator" : "\|",
                        "cgi_field" : "CGI",
                        "frequency_fields" : "BAND",
            "technology_generation_value" : "4G",
                        "import_fields" : {
                                "lat" : "LATITUDE",
                                "lon" : "LONGITUDE",
                "radius" : "RADIUS",
                "direction" : "AZIMUTH",
                "frequency" : "BAND"
                        },
                        "removed_input_fields" : [],
                        "coordinate_system" : "DegDec"
                }
    ],
    "removed_output_fields" : ["mcc","mnc","lac","cellId"],
    "output_order" : [
        "cgi","lat","lon","cluster_id"]
}
```

C Mathematical model of Spatial Probability Density Function

Text of this section is written for the unpublished article [Vajakas and Vajakas, 2014] by Jaan Vajakas.

Random variables are given by underlining (following the van Dantzig convention). Dot is a placeholder for an argument of a function.

For each cell C in a cell plan $\mathcal{C}$ and each pixel x, we want to compute $\mathrm{P}(x|C) = \mathrm{P}(\underline{x} = x \,|\, \underline{C} = C)$, probability that the EM is at location x if it generates an event in cell C.

In the following, it is described how to do it. The raster $\mathrm{P}(\underline{x} = \cdot \,|\, \underline{C} = C)$ is then the SPDF of the cell C.

The probability $\mathrm{P}(x|C)$ is determined by the Bayes' formula:

$$\mathrm{P}(x|C) = \tfrac{\mathrm{P}(C|x)}{\mathrm{P}(x)}, \tag{A.1}$$

where

- $\mathrm{P}(x|C) = \mathrm{P}(\underline{C} = C \,|\, \underline{x} = x)$ is the probability that if an EM is at location x then it is connected to cell C (rather than any other cell or no cell);
- $\mathrm{P}(x) = \mathrm{P}(\underline{x} = x)$ is the Bayesian prior density, i.e. the probability for an ME to be in a pixel x;
- $\mathrm{P}(C) = \mathrm{P}(\underline{C} = C)$ is the probability for ME (at a random location) to be connected to the cell C, i.e. $\mathrm{P}(C) = \sum_i \mathrm{P}(C|x)\,\mathrm{P}(x)$ where x ranges theoretically over the world (in practice, over an area of interests, e.g. one country).

The Bayesian prior $\mathrm{P}(x)$, representing our prior belief, can be constructed from population density data, from road or buildings layers, as people are more likely to be on a road or in a building.

The probability $\mathrm{P}(C|x)$ is computed as follows:

$$\mathrm{P}(C|x) = \mathrm{P}(\mathit{connected}|x) \cdot \mathrm{P}(C|x, \mathit{connected}), \tag{A.2}$$

where

- $\mathrm{P}(\mathit{connected}|x)$ is the probability that ME is connected to the mobile network (i.e. ME is able to generate an event), when it is at location x;
- $\mathrm{P}(C|x, \mathit{connected})$ is the probability that ME is at the location x and is connected to the cell C.

Let $\underline{S}$ denote the active set of cells that can be connected by the ME, i.e. the set of cells actually detectable by the ME at a given time moment. A rough estimate of the conditional probability $\mathrm{P}(C \in \underline{S}|x)$ (that a certain cell C is detectable by the ME if the ME is at the point x) is provided by the MNO in the form of cell polygon: $\mathrm{P}(C \in \underline{S}|x)$ is 1 if location x lies inside the cell's polygon and 0 if outside, so the active set depends deterministically on x. In a more refined model, $\mathrm{P}(C \in \underline{S}|x)$ could have non-zero values to reflect our ignorance of the precise coverage area and the stochastic nature of cell coverage caused by effects like Rayleigh fading and weather changes.

CDR events can only be received when the ME is connected to the network. Estonia is very well covered and connection probability is almost 100%. Actual connection rate is unknown. Therefore, simplify is made and calculate for each raster pixel $\mathrm{P}'(C|x) \approx \mathrm{P}(C, \mathit{connected}|x)$.

For the probability $\mathrm{P}(C|x, \mathit{connected})$ an ad-hoc formula is proposed (in contrast to the other formulas, that have some theoretical justification): namely, the probability is taken to be proportional to the square of the probability $\mathrm{P}(C \in \underline{S}|x)$, i.e.

$$\mathrm{P}(C|x, \mathit{connected}) = \tfrac{\mathrm{P}(C \in \underline{S}|x)^2}{\sum_{c \in C} \mathrm{P}(C \in \underline{S}|x)^2}. \tag{A.3}$$

Formula A.3 expresses the consideration that radio network tries to avoid using the cell with lower SINR, i.e. uses “the winner takes (almost) all calls” line of thinking.

The SPDF estimate quality is evaluated with logarithm of likelihood

$$\sum_i \log \mathrm{P}'(C_i|x_i), \tag{A.4}$$

where i is an index of measurement (one measurement is one CDR with established GPS location).

C.1 Algorithm for computing the Bayesian PDF of cells and the heat map aggregate statistics using the cells

For practical use, the computations are separated into two steps: at first, the PDF is calculated for each cell and after that, the calculated PDFs can be used for various statistical calculations.

This particular implementation was optimized towards heat map calculations, where each cell is given a weight (e. g. the number of people connected to given cell) and raster image is generated, showing spatial density distribution.

For computations, we represented the probability field of each cell as a raster image in the RAM (random access memory). We used the Web Mercator projection. We had a fixed resolution called *full resolution*, which is the resolution we were referring to in the formulas above (when talking about the heatmap pixel x).

In order to save space, Raster data of the cell PDF was stored in computer memory in lower resolution. By *resolution*, we mean pixels per coordinate unit. More precisely, our variable-resolution raster was a grid of tiles, each tile a square matrix of floating-point numbers (pixels) with sidelength a power of two pixels. The admissible resolutions for storing the tiles were the full resolution and the resolutions a power of two times lower than the full resolution.

The rendering was done at a resolution inversely proportional to the square root of the area of the polygon (rounded to the nearest higher admissible resolution, or full resolution if no higher resolution available). This reduction of resolution is justified by the fact the cell shapes have relatively large spatial uncertainty which is proportional to cell size — larger cells have a larger uncertainty of the boundary of the cell's actual service area.

For each cell C, we obtained the raster $\mathrm{P}(C \in \underline{S}|x)$, the probability that cell C was detectable at a given location, by rendering the cell's polygon from the cell plan (1 if the pixel is inside the polygon, 0 if outside, antialiasing on the edges). Optionally we applied a buffer and smoothing effect (blur) to raster obtained from cell plan. These rasters were used in next step for Bayesian PDF calculations.

Using the formulas above, we combined the rasters of the probability fields $\mathrm{P}(C \in S|x)$ (uniform-resolution, but different cells possibly at different resolutions) to produce the variable-resolution rasters $\mathrm{P}(C|x)$ and finally $\mathrm{P}(x|C)$. During the computations, we did not reduce the resolution anywhere. The resulting rasters were stored in memory for use in statistical calculations.

We implemented heatmap calculation functionality based on Bayesian PDF. For heatmap calculation one uses the PDF in an appropriate resolution. If PDF is stored in lower resolution and higher resolution does not exist then available raster is upscaled accordingly.

On optimal spatial probability density estimation of passive mobile positioning events

Toivo Vajakas
Reach-U Ltd;
Institute of Computer Science, University of Tartu
Tartu, Estonia
e-mail: tvajakas@gmail.com

Joosep Rõõmusaare
Reach-U Ltd;
Institute of Computer Science, University of Tartu
Tartu, Estonia

***Abstract* — In passive mobile positioning the cell-level measurements must be translated into geographical location, which can be expressed as spatial probability density function (SPDF).**

In this paper we present the results of a study where we compared different methods to estimate SPDF of passive mobile positioning. The mobile operators provide spatial data about network cells. It is called cellplan. Cellplan can be provided in various formats: location of cell towers and azimuth of antennas, or radio signal strength levels from radio propagation models, etc. Each such source requires specific processing to infer SPDF for passive mobile positioning. We investigated the probabilistic properties of different processing algorithms on different variants of input data. Some investigated methods apply Bayes rules to take into account the effects from overlapping neighbor cells. The results indicate that even on relatively small dataset one can clearly see different accuracy from different processing parameters. The accuracy of SPDF varies significantly depending on processing parameters. and sophisticated radio propagation models did not have significant advantages over simple procedures using Voronoi diagram. (*Abstract*)

Keywords* — *passive mobile positioning; Bayesian location estimate; spatial probability density function; PDF, heat map, location estimation accuracy

I. Introduction

A. Motivation

Passive mobile positioning data, gathered as a by-product by mobile operators, has gained much popularity in human geography studies due to availability of large samples [1]. Mobile positioning data describes the location of a mobile station (MS). A MS can be a phone or a modem for a device such as a security or environmental sensor. Mobile positioning determines the position of a MS with significant spatial uncertainty [2]. Spatial uncertainty can be generally described as spatial probability density function (SPDF) of location

The mobile operators provide spatial data about network cells. It is called cellplan. Cellplan can be provided in various formats: location of cell towers and azimuth of antennas, radio signal strength levels from radio propagation models, etc. Each such source requires specific processing to infer SPDF for passive mobile positioning. In majority of mobile positioning papers the effects of cell area overlap are ignored. We considered applying Bayesian rule to take into account, importance of this effect needed verification.

The aim of this paper is to describe a methodology for such investigation and compare various cellplan SPDF preparation methods.

B. General characteristics of mobile positioning data

Each passive mobile positioning data record has an attribute identifying the mobile network cell the phone was connected to, known as the Cell Global Identity (CGI). A cell is the geographical area where it is possible to connect to one transceiver of a base station. Each cell has limited capacity and therefore operators design smaller cells in regions of high population density. Neighboring cells have considerable overlap. When a mobile phone disconnects from one cell and connects to another the event is called a ‘handover’. For mobile positioning it is important to know the geographical shape of each cell. The actual cell shape depends on many factors, such as antenna radiation pattern and height, network load, signal attenuation on landscape and indoors, signal reflections, radio interference and noise, network configuration parameters such as handover threshold and neighbor cell lists [3]. Fig 1 illustrates the uncertainty present when determining location from the fact that phone is connected to particular cell.

Other location-related attributes in addition to CGI can be collected for improved location accuracy, such as distance to the antenna or signal strength from neighboring cells. CGI data is however still the most scalable passive positioning approach and puts the least load on a network, including for the recently introduced LTE (Long-Term Evolution) networks [4]. In this paper we consider positioning using only CGI data.

Mobile positioning data can be exported from different nodes in the mobile operator’s network, resulting in different levels of detail of the data. The most notable options of passive positioning data are call detail records (CDR) and network subsystem (NSS) event stream [2]. CDR data has been the most widely used option in mobile positioning research. Each

CDR describes a billing-related subscriber activity like starting a call or sending a text message. In some configurations mobile operator provides also periodic update events that report (typically every hour or two) the current location of a mobile. The network also generates location update events when a phone moves from one location area (a group of closely situated base stations) to another [2].

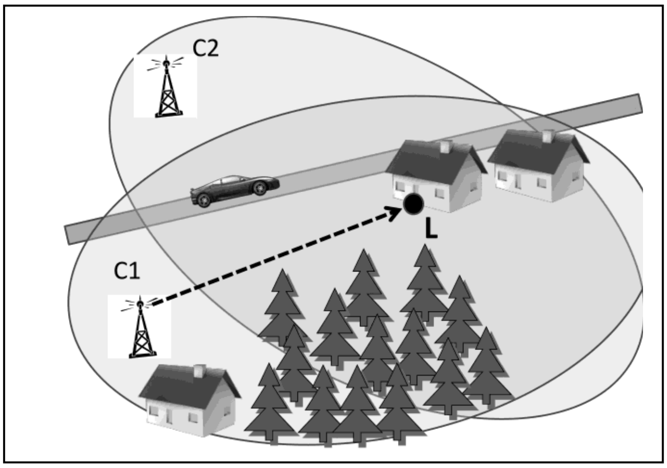


Figure 1. Mobile station location uncertainty problem illustration. Suppose we know that a MS was connected to cell C1 and want to estimate the probability that MS was in location L. The spatial probability distribution of each cell (as defined by cellplan) is shown as filled oval areas. The probability estimate is affected by cell C2 that also covers location L and by prior knowledge that people stay mostly in houses and move on roads and nobody lives in forest.

C. *Related work*

Based on available information the shape of each cell has to be defined to give location estimates for mobile positioning. Cell data provided by mobile operators can be translated to cell shapes as Voronoi polygons by using the assumption that a phone connects to the nearest tower [5]; as best server data polygons by using the assumption that a mobile phone connects to the cell with the strongest signal [6]; or as a raster model based on the assumption that the probability to connect to a cell is a function of distance from the antenna tower [7].

A methodology applying Bayesian methods for defining the SPDF for mobile positioning was given by Zang et al. (2010) [8]. That paper provided a solution for the situation where neighbor antennas are present. The estimate was based on signal-to-interference-and-noise ratio (SINR) calculations. The paper provided a Bayesian probability estimate for situations where CGI is the only information known and additionally it is known that no other cells had good enough SINR. The method was tested against the subset of emergency call data limited to situations where only one cell was within the reach of the MS. They found that the difference between the location measured by GPS and the MLE provided by that method was improved by 20%, compared to baseline method. No direct PDF tests were performed by Zang et al. (2010) [8]. The Bayesian probability formulas in the paper by Zang et al. do not consider the effect that the probability to connect to a concrete cell will be reduced in locations where other cells are also reachable.

D. *Research problem*

We are investigating how the SPDF estimation quality is affected by following factors

- Different input data (tower coordinates vs radio propagation data)
- Post-processing of data (e.g. enlarging and blurring cell boundaries by given factor)
- How much is result affected by cell overlap effects
- How much the result depends on MS and location

II. METHODS

A. *Mathematical model of SPDF*

Given a mobile operator's event logs for some time period, we want to map each event probabilistically to geographical space where the mobile event occurred. We assign spatial PDF to each cell such that PDF defines the probability that the event occurred in any given location. We consider here only discrete probability densities obtained by dividing the area of interest into pixels of appropriate size.

The radio area network (RAN) consists of a finite set of cells, $\mathcal{C}$. For each cell C in the cellplan $\mathcal{C}$ and each pixel x, we want to compute $\mathrm{P}(x|C) = \mathrm{P}\big(\underline{x} = x \mid \underline{C} = C\big)$, the probability that the MS is at location x if it generates an event in cell C. In the following we will describe a method how to do it. (The raster $\mathrm{P}\big(\underline{x} = \cdot \mid \underline{C} = C\big)$ is then what we call the spatial PDF of the cell C.)

The probability $\mathrm{P}(x|C)$ is determined by the Bayes' formula:

$$\mathrm{P}(x|C) = \frac{\mathrm{P}(C|x)}{\mathrm{P}(x)} \qquad (1)$$

where

- ○ $\mathrm{P}(C|x) = \mathrm{P}(\underline{C} = C \mid \underline{x} = x)$ is the probability probability that if a MS at location x then it is connected to cell C (rather than any other cell or no cell);
- ○ $\mathrm{P}(x) = \mathrm{P}(\underline{x} = x)$ is the Bayesian prior density, i. e. the probability for a person (or more precisely, a mobile station) to be in pixel x;
- ○ $\mathrm{P}(C) = \mathrm{P}(\underline{C} = C)$ is the probability for mobile station (at a random location) to be connected to cell C, i. e. $\mathrm{P}(C) = \sum_x \mathrm{P}(C|x)\mathrm{P}(x)$ where x ranges theoretically over the world (in practice, over an area of interest, e. g. one country).

The Bayesian prior $P(x)$, representing our prior belief, can BE constructed e. g. from population density data, road and building layers (people are more likely to be on road or in a building).

The probability $P(C|x)$ is computed as follows:

$$\mathrm{P}(C|x) = \mathrm{P}(\text{connected} \mid x) \cdot \mathrm{P}(C \mid x, \text{connected}) \qquad (2)$$

where

- P(connected | x) is the probability that the MS is connected to the RAN at all (i.e., is able to generate an event) if it is at location x,
- P(C | x, connected) is the probability that if a MS is at location x and is connected then it is connected to cell C.

Let $\underline{S}$ denote the active set of the MS, i. e. the set of cells actually detectable by the MS at a given time moment. A rough estimate of the conditional probability $\mathrm{P}(C \in \underline{S} \,|\, x)$ (that a certain cell C is detectable by the MS if the MS is at point x) is provided by the mobile operator in the form of cell polygon: $P(C \in \underline{S} \,|x)$ is 1 if location x lies inside the cell's polygon and 0 if outside, so the active set depends deterministically on x. In a more refined model, $P(C \in \underline{S} \,|x)$ could have non-zero values to reflect our ignorance of the precise coverage area and the stochastic nature of cell coverage caused by effects like Rayleigh fading and weather changes.

We can only receive CDR events when the mobile phone is connected to network. Also, Estonia is very well covered and connection probability is almost 100%. We don't know actual connection rate. Therefore, we simplify and calculate for each raster pixel $\mathrm{P}'(C \,|\, x) \approx \mathrm{P}(C, connected \,|\, x)$.

For the probability P(C | x, connected) we propose an ad-hoc formula (in contrast to the other formulas in this paper, which have some theoretical justification): namely, we take the probability to be proportional to the square of the probability $\mathrm{P}(C \in \underline{S} \,|x)$, i. e.

$$\mathrm{P}(C \mid x, \text{connected}) = \frac{\mathrm{P}(C \in \underline{S} \,|x)^2}{\prod_{C \in \mathcal{C}} \mathrm{P}(C \in \underline{S} \,|x)^2} \qquad (3)$$

B. *Description of test data*

Data consists of GPS measurements of individual persons who have installed GPS track recording software into their mobile phones, and CDR data for same persons from mobile operator.

For each CDR record we found from GPS track the location of person for given time moment. The records not covered by GPS track were ignored. We used 4 personal tracks from same 8 month time period. Example of data is given on Fig 2.

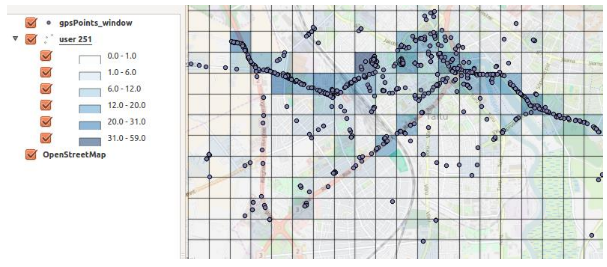


Figure 2. Measured data used for model accuracy estimation. On the left is colorscale for count of measurements in grid cell. Each grid cell is 630m square

C. *SPDF model quality assessment criteria*

We can only receive CDR events when the mobile phone is connected to network. Also, Estonia is very well covered and connection probability is almost 100%. We don't know actual connection rate. Therefore, we simplify and calculate for each raster pixel $\mathrm{P}'(C \,|\, x) \approx \mathrm{P}(C, connected \,|\, x)$.

The quality of SPDF estimate is evaluated with logarithm of likelihood

$$\textstyle\sum_i \log \mathrm{P}'(C_i \mid x_i) \qquad (4)$$

where *i* is index of measurement (one measurement is one CDR with established GPS location).

III. Results

A. *SPDF variants from cellplans*

We divided test area into quadratic pixels of size 630 meters. For all cellplan variants the values of raster $\mathrm{P}'(C \,|\, x)$ were calculated. As illustration on Fig 3 is cross-section of area, showing relative probability of each cell in given point.

B. *Model likelihood calculations*

Using the calculated SPDF rasters we calculated with formula (4) the likelihood of data given each SPDF variant tested. There were two cellplan datasets for same network, one based on RSSI (Received Signal Strength Indication) data and another on tower+azimuth (used to construct Voronoi polygons). Derived variations include:

- rssi_default –RSSI data, unchanged
- rssi_A1AP2 –twice stretched beam width
- rssi_A2AP1 –twice stretched beam along azimuth
- rssi_A3AP1 –three times stretched along azimuth
- rssi_convexhull –convex hull over original geometry
- rssi_ellipse –original approximated with ellipse
- voronoi_default – tower coordinate data, Voronoi constructed
- voronoi_A1AP2 – stretched twice width of beam
- voronoi_A2AP1-- stretched twice along the beam
- voronoi_A3AP1—stretched three times along the beam
- voronoi_convexhull -- original
- voronoi_ellipse

The results are shown on Fig 4.

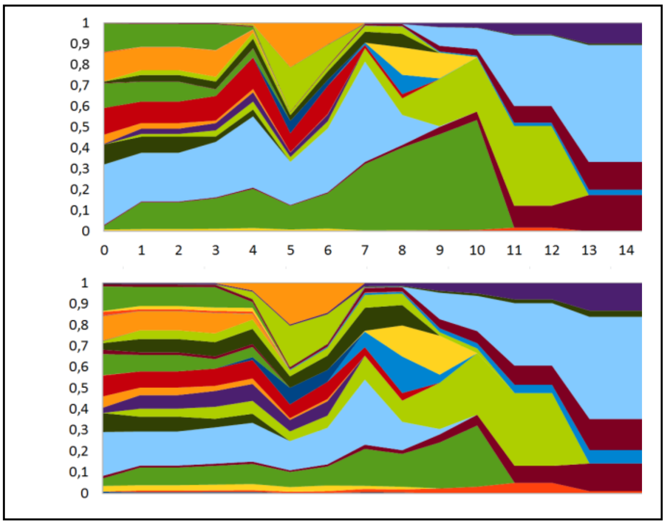


Figure 3. Illustration of SPDF of all cells along cross-section of test area along a line, showing relative probability of each cell in given location. Horizontal axis is pixel number (each pixel is 630m) and vertical axis is stacked probabilities of cells. Upper chart is generated with applying Bayesian overlapping cell model, lower drawing without considering overlapping effects.

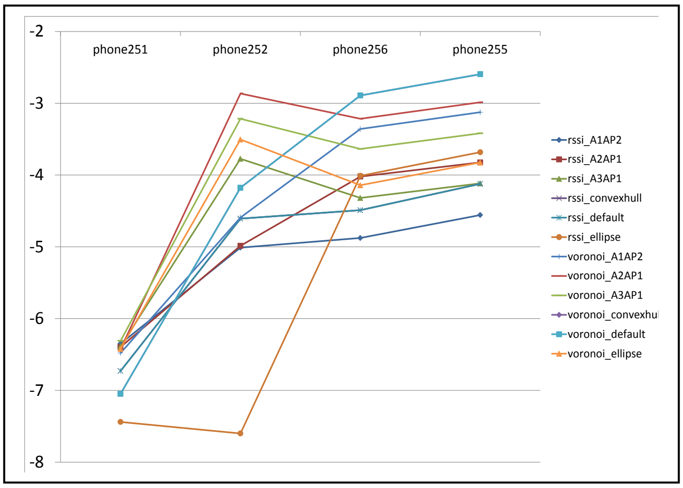


Figure 4. Relative performance of various cellplan variants. Horizontal axis – different test phone tracks (subsets of positioning data, with different spatial distribution). Vertical axis – average log P(C|x) for CDR records of given track.

IV. Discussion

The main findings are

- Some processing variants performed significantly better than the others for certain situations, but there was no single best method for all situations. When backed with large test dataset one could develop a mixed processing procedure using different method for different areas but it is limited by overfitting concerns.
- SPDF calculated from RSSI data was not superior to simplistic Voronoi-based SPDF. We had expected that RSSI input enables much better estimation than tower location and azimuth data.
- The modest dataset consisting of four personal tracks characterizes some location sufficiently but is not sufficient to give overall picture.
- Accounting for cell overlap effect with Bayes rule had in majority of cases positive effect. The likelihood value depended much more on location than Bayes, but this need not mean that applying overlap correction with Bayesian rule is insignificant. Due to the methodology of comparison likelihood is related to probability density, and in areas with larger cells the SPDF values are expected to be significantly lower, thus it need not be the flaw of SPDF estimation.

In future work we plan to analyze specific situations where performance of one or other SPDF estimation variant degrades and optimize the methods accordingly. Also we plan to investigate effects of previous state, e.g. MS approaching cell, or stationary in neighborhood of cell.

Acknowledgment

This research was partially supported by traffic applications research project in Reach-U Ltd and ELIKO and by Location-Based Big Data Algorithms project in STACC. This research has been supported by the European Union through the European Regional Development Fund.

References

[1] J. Steenbruggen, E. Tranos and P. Nijkamp, "Data from mobile phone operators: a tool for smarter cities?," Telecommunications Policy, 2014.

[2] E. Saluveer and R. Ahas, "Using Call Detail Records of Mobile Network Operators for Transportation Studies," in Mobile Technologies for Activity-Travel Data Collection and Analysis, Hershey, PA, IGI Global, 2014, p. 224.

[3] M. Amirijoo, "Neighbor Cell Relation List and Physical Cell Identity Self-Organization in LTE," in 2008 IEEE International Conference on Communications Workshop, 2008.

[4] S. Cherian and A. Rudrapatna, "LTE Location Technologies and Delivery Solutions," Bell Labs Technical Journal, vol. 18, no. 2, p. 175–194., 2013.

[5] R. Ahas, A. Aasa, A. Roose, S. Silm and Ü. Mark, "Evaluating passive mobile positioning data for tourism surveys: an Estonian case study," Tourism Management, vol. 29, no. 3, pp. 469-486, 2008.

[6] F. Calabrese, "Urban sensing using mobile phone network data.," in Ubicomp 2011 Tutorial, 2011.

[7] F. Calabrese, Using cell-phone data to understand urban dynamics in the city of Amsterdam, MIT SENSEable City Laboratory., 2008.

[8] H. Zang, F. Baccelli and J. Bolot, "Bayesian inference for localization in cellular networks," in 2010 Proceedings IEEE INFOCOM, 15–19 March 2010, San Diego, CA, USA, 2010.

E Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, Joosep Rõõmusaare (date of birth: 29th of July 1988),

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

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

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

Probabilistic Location Estimate of Passive Mobile Positioning Events

supervised by Toivo Vajakas

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

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

Tartu, 11.08.2016