UNIVERSITY OF TARTU
FACULTY OF MATHEMATICS AND COMPUTER SCIENCE
Institute of Computer Science
Msc. Software Engineering

Mariano Hedberto Jofre

Placement and Movement Episodes Detection using Mobile Trajectories Data

Master's thesis (30 ECTS)

Supervisor(s): Dr. Amnir Hadachi
Msc. Elis Kõivumägi

May 2015

Placement and Movement Episodes Detection using Mobile Trajectories Data

Abstract

This thesis presents a trajectory episode matrix to enable the detection of placement and movement episodes from mobile location data. The data used in this work is very sparse in time and space. Therefore, the estimation of user's placement and movement patterns poses a big challenge. The presented approach performs data analysis to find meaningful locations and introduces an algorithm to detect movement and placement episodes. To perform the analysis and visualize the results a statistical analysis tool was developed with R. The work done as a result of this thesis can be used to improve the identification of the meaningful locations and to help predicting the semantic meanings of mobile user's patterns.

Keywords:

Urban mobility, movement, placement, pattern, event, episodes, trajectories, time spatial data, business intelligence, R, cloud, mobile positioning, passive data, antenna, cells, centroid, CDR, Cell Plan, location-based services, anchor points, modelling, ping-pong handover, clustering.

Asukohaandmetest seisu- ning liikumisepisoodide tuvastamine

Lühikokkuvõte

Teostatud töö eesmärgiks on tuvastada asukohaandmetest seisu- ning liikumisepisoode kasutades selleks trajektoori ülekattuvusmaatriksit. Antud töös kasutatud andmed on väga hajusad nii ajalises kui ka geograafilises mõttes. Seetõttu on antud ülesanne suur väljakutse. Välja pakutud lahenduse raames teostati andmeanalüüs mille raames tuvastati kasutajatele tähtsad asukohad ning pakuti välja algoritm, mille abil tuvastda seisu- ning liikumisepisoodid. Andmete analüüsimiseks ning visualiseerimiseks kasutati R-i.

Võtmesõnad:

Linnasisene liikuvus, liikumine, asetus, muster, sündmus, episood, trajektoorid, aeg-ruumilised andmed, äriintellekt, R, pilvetehnoloogia, mobiilpositsioneerimine, passiivsed andmed, antenn, cell, raskuskese/kaldenurk, CDR, Cell Plan, asukoha-põhised teenused, ankrupunktid, modelleerimine, ping-pong ülekanne, klasterdamine.

ACKNOWLEDGEMENTS

This thesis was made during the last nine months of my Master of Software Engineering as part of this Master´s program. I am deeply grateful to the University of Tartu - Faculty of Computer Science for giving me the opportunity to pursue my studies. This thesis is made possible by the cooperation and assistance of different people and organizations. Here I mention some of them.

First I want to thank my thesis supervisors, Dr. Amnir Hadachi and MSc Elis Kõivumägi for their support, teaching, guidance, commitment and patience. I am deeply grateful for having had the opportunity to develop my thesis in the Distributed System Group - Institute of Computer Science of the University of Tartu and STACC - Software Technology and Application Competence Center.

I want to extend my gratitude to the other members of the Distributed System Group, Artjom Lind and Oleg Batrashev for their help, teachings and fellowship.

Last but not least, I want to thank my family for all their love, unconditional support and help. My wife Ximena to be my partner and my life. My son Lautaro in whom one always reflected. My parents for their guidance and support.

LIST OF FIGURES

LIST OF ABBREVIATIONS

CDR: Call Detail Records

KDD (Knowledge Discovery in Databases)

GKD: Geographic Knowledge Discovery

GPS: global position system

DBSCAN: Density-based spatial clustering of applications with noise

V-DBSCAN: Varied density spatial clustering of applications with noise

ST-DBSCAN: Spatial-Temporal Density Based Clustering

Table of Contents

1. INTRODUCTION

1.1 Motivation

Discovering meaningful locations and recognizing user's movement and placement patterns from mobile location data is necessary to understand social patterns, mobility behavior and human mobility. Such comprehension can give us solutions to demanding issues in a wide range of fields, such as mobile network planning, urban planning [2, 3, 10] and behavior of customers [4]. There have been numerous approach and studies on meaningful locations and movement pattern analysis from mobile phone records [1, 3, 6, 7, 12, 13, 14, 16, 20, 28]. Mobile operators have the capability to collect thousands of location events per user per month. Detection of the episodes to find meaningful locations and discovering placement and movement patterns requires new approaches.

We use the combination of CDRs and cell plan as input data. The CDR data has information about the interaction between the mobile user and the network. Each interaction is an event. The cell plan has cell's geographical position data and also has a model representation of coverage area as a polygon.

From the study of the input data and the frequency event analysis we find that user has some regular placement and trajectory patterns. We also find that some events show, for example, an impossible trajectory or the time-distance can't be possible. This situation is created by wrong input data or because of a ping-pong handover phenomenon in the mobile network

An episode is defined by the intersection between trajectory events polygons. By the detection of placement and movement episodes it is possible to get more accurate user trajectory patterns. We believe that by calculating the "trajectory episode matrix" we could accomplish our prediction to detect placement and movement patterns from mobile trajectories data.

Finally, there is also a requirement of running some clustering analysis to improve the episode spatial distribution.

1.2 Thesis Rationale

1.2.1 Research Questions

The data used during the research originates from mobile location stream. For mapping location events to physical world a cell plan from mobile operator is used, however to get useful information from that is mandatory to discover patterns and answer the following research questions:

(a) What are the episodes placement/movement for every trajectory user?
(b) What are the patterns in placement episodes?
(c) What are the most frequent trajectories of displacements?

1.2.2 Research Objectives

All of these issues are crucial for the analyses of movement episodes from mobile trajectories patterns and to get accurate results. The enumerations of the previous research questions are echoed in these research questions.

Objective 1: Try to detect and extract the placement episodes pattern from the combination of CDR and Cell Plan data.

Objective 2: Create a matrix to show the trajectory of user's episodes per day. The episodes are defined by the overlapping of the trajectory cells polygons.

Objective 3: Identify different placement episodes as a result of the overlapping of consecutive events.

Objective 4: Implement a framework to analyze and visualize clustering analysis dynamically.

Objective 5: Show the most frequent placement and movement episodes for each user and remove outliers from the input data.

1.2.3 Structure of the Thesis

This thesis is organized as follows: Chapter 2 will present the background and related work relevant for the presented research. In Chapter 3 we present the methodology of the analysis approach and the technical solution of it. In this chapter is defined the trajectory episode matrix. Chapter 4 we present the results of the data analysis and we discuss on it and finally on Chapter 5 we conclude the thesis by presenting conclusions and future work ideas

2. BACKGROUND AND RELATED WORK

This chapter will give an overview of the theoretical background and related work of the topic. We have divided our related work section into six subsections. The first section will review the elementary definitions about the movement analyses and a description of their components. The second section will review the concept of Movement and Placement. The third part will review the Geographic Knowledge Discovery (GKD) process to get information from spatial data. The fourth section will review related work on Mobility data mining using CDR´s data. The fifth part of this chapter will review on the CDR and Cell Plan input data. Finally, the last section will review the technical solution to perform the analyses with R.

2.1 Key terms - Definitions

Definitions of principal key terms to define movement and placement:

(a) **Movement:** is defined as “a change in the spatial location of the whole individual in time”, that is, as whole-body movement [11]. This thesis uses the Eulerian frame of reference to characterize the flow movements, which has the temporal and spatial dimensions.

(b) **Trajectory:** is the movement of an object and is a time sequence ordered set of positions [11]. Related with the observations of trajectory movements is about the series of consecutive tuples coordinates-timestamp.

(c) **Moving individual:** is defined as the individual position changes over time. It is called Moving Point Object (MPO) in which the location of the object in time is considered. The geographic coordinates define the location. The time-georeferenced objects are defined by a tuple of Cartesian coordinates and time. The particular case of this thesis each geographical event is defined by a set of Cartesian coordinates which represent a polygon.

(d) **Polygon:** is a geographical representation of the cell coverage area in which represents a modeling coverage area of every cell. It belongs to the antennas. The shape and size of the polygon is up to the cell direction, power and type of cell signal.

(e) **Centroid representation:** each polygon has a geometric center or an arithmetic mean of the polygons coordinates. For the aim to calculate different movement features like distance and time distance is important to simplify the model with centroids instead of polygons.

(f) **Georeferenced movement:** movement objects or processes that move about in geographic space.

(g) **Event:** it is also calls as positioning event. A passive location data acquisition happens each time the user interacts with some mobile features or when the mobile company updates the network location. Example of user events are making/answering calls, sending/receiving text messages, initiating a data session, etc.

(h) **Episode:** for this thesis the state of the users could have to labels, placement and movement. So, each time that a placement or movement happens define an episode. An objective of this thesis is to identify the different user episodes along the trajectory.

2.2 Concept of Movement and Placement

The episodes approach needs a deep comprehension about the concept of movement and placement. This is means that is important to understand when the user is settled or moving. In particular since we are using CDR data related with a cell plan provided by the mobile companies.

The model of movement framework [11] helps to understand and define clearly the movement concept and characteristics. This model has four principal components:

(a) **Internal state:** motivation of user to move.

(b) **Movement characteristics:** this are divided in motion capacity (coordinates and timestamp) and navigation capacity (trajectory). This data is collected from the Cell Plan and CDR.

(c) **Movement path:** it defines mainly two types of movement, continuous path and discontinues path. This thesis analyses discontinues path because the coordinates and time data is updated only when an event happens.

(d) **External factors**: are surrounding environment defined by spatial constraints, important places, weather, etc.

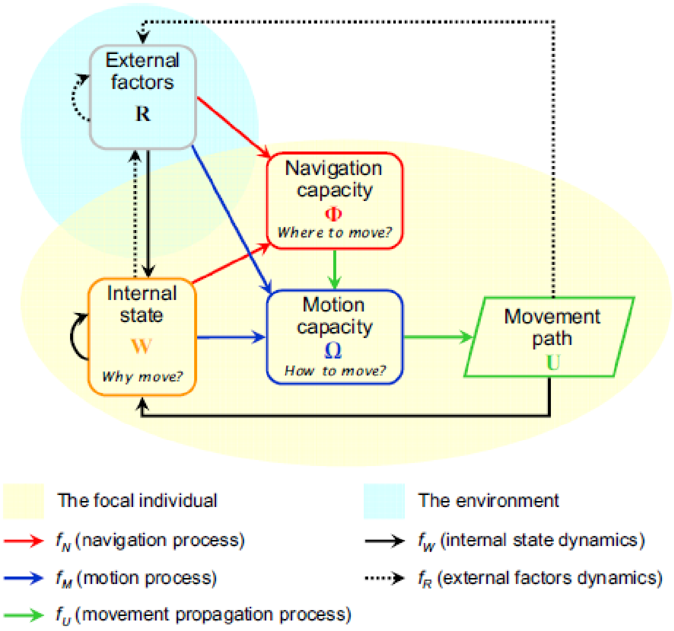


Figure 1 - Relation between the movements' factors [11]

Elements, which define the movement, can be related with the motivations and environment threshold by one side, and could also be defined by intrinsically features like geographic position, time, speed, etc.

The other movement state that cares on the aim of this thesis is the placement. The placement is related with the urban spatial-temporal structure (USTS) [8], which defines the space S object.

This space has 3 dimensions, two spatial dimensions and one time dimension.

$$S \triangleq \{s = (t, x, y) | t \in [t_o, t_1], (x, y) \in A\} = [t_o, t_1] \times A$$

t: time
x: longitude
y: latitude
$[t_o,t_1]$: time period of interest
A: set of Latitude and Longitude to study

Finally is important to define, as the aim of this thesis placement and movement related with a high frequency visit cell.

Placement Pattern: when the object, in the thesis a mobile user, stays fiscally in a certain place and he doesn´t change his coordinates, and this behavior is discovered for any time-spatial mining process.

Movement Pattern: when the object, in the thesis the user, has a fiscally coordinates change and this change and this behavior is discovered for any time-spatial mining process.

2.3 Geographic Knowledge Discovery (GKD)

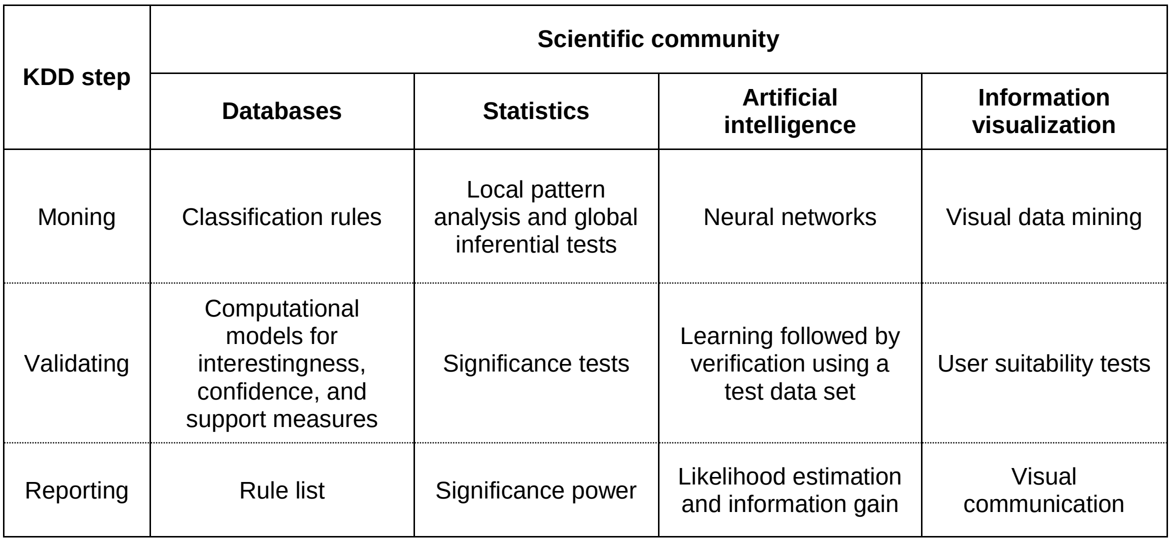
The "process of obtaining information through data mining and distilling this information into knowledge through interpretation of information and integration with existing knowledge" [30] is the formal definition of Knowledge Discovery in Databases (KDD) and to achieve the objectives of this thesis; it is necessary to apply some technique KDD in the field of geographic analysis. The KDD techniques process data and, after that discover hidden information and interesting patters or behaviors.

The KDD give us a clearly step-by-step framework to get and handle the raw data, the posterior processing, and finally the possibility to extract useful information.

The spatio-temporal data has a particular approach of KDD, which it calls Geographic Knowledge Discovery (GKD). The spatio-temporal data include moving objects and it needs some sophisticated data mining approaches and techniques to analyze data properly.

The KDD is also a framework, which has several steps to pursuit.

- **(a)** Data selection
- **(b)** Data preprocessing
- **(c)** Data enrichment
- **(d)** Data reduction and projection
- **(e)** Data mining and pattern recognition
- **(f)** Reporting



KDD step	**Scientific community**			
	Databases	**Statistics**	**Artificial intelligence**	**Information visualization**
Moning	Classification rules	Local pattern analysis and global inferential tests	Neural networks	Visual data mining
Validating	Computational models for interestingness, confidence, and support measures	Significance tests	Learning followed by verification using a test data set	User suitability tests
Reporting	Rule list	Significance power	Likelihood estimation and information gain	Visual communication

Figure 2 - KDD steps related with the technique used in the Scientific Community [30]

2.4 Mobility data mining using CDR´s data

2.4.1 Analyze movement and patterns

The paper from Fosca Giannotti “Unveiling the complexity of human mobility by querying and mining massive trajectory data”[18] used car GPS collected data form to analyze movement and patterns. It is a detailed and interesting approach to spatio-temporal Data Mining. There are some clear definitions of basic statistics to represent the movements of the users.

- **(a)** Trip length: Euclidean distance with the Latitude and Longitude as coordinates.
- **(b)** Trip duration: time difference between Latitude and Longitude as coordinates.
- **(c)** Correlation of length and speed trips

(d) Radius of gyration: $rg(T) = \sqrt{\frac{1}{n} * \sum(euclidean\ distance - \ cm(T))^{2}}$

The radius of gyration helps to show the preferred location of a user in his movement patterns.

(e) Spatio-temporal analysis of density: this is spatial distribution density heatmap. For our thesis task the Trip length will be called “distance” and the “Trip duration” will be called “time-distance”.

Cm: center mass represent the centroid value of the Latitude and Longitude as coordinates, which belongs to a path.

2.4.2 Algorithms to analyze movement patterns

The frequent movement patterns collected from CDR data are the result of the analyses to discover the frequent trajectories and eliminate the infrequent trajectories or places. The algorithm LS – Large Sequence [19] is a regression based approach. The CDR data to perform the algorithm LS is similar that the data available from this thesis project. It has the ID User, Date/Time and ID Cell features.

The next step, after the regression approach, is to perform the time-spatial data with the algorithm TC (Time Clustering). It is used to group the CDR with a close occurrence time.

The paper “A regression-based approach for mining user movement patterns from random sample data” [19] gives a solid knowledge to identify patterns throw the LS and TC algorithm, however I doesn´t consider the overlapping or ping-pong handover issue. The other limitation is that they perform the algorithms with sample random data. Sometimes, run algorithms with “for loop” instances would generate problems with a big amount of data or calculus.

The trajectory patterns could be found with the regression and cluster algorithms (LS ans TC) [19] but also been other methods like interpolation [20]. Yihong Yuan and Martin Raubal (2010) [20] found patterns with interpolation and the correlation between mobile usage and trajectory patterns. This scope just focus in the average radius spread of any user in a city, so this spatial clustering just can show spatial dispersion in a map.

2.4.3 Clustering Algorithms for Spatial dataset

Clustering is very important data mining method to identify classes from a dataset. For the particular case of this thesis also there are many spatial clustering techniques to handle spatial and even time-spatial datasets.

Is important to study which of the clustering techniques fit better to perform the analyses with our particular input data and, in the process deeply understand what the strengths and weaknesses of each method.

2.4.3.1 Considerations for clustering spatial data

To choose what is the most suitable classification method to analyze spatial data is necessary to understand the strength and weakness of every clustering method

Main limitations for clustering spatial data:

(a) k-partitioning clustering algorithms like k-means are bounded to clusters with a convex shaped.

(b) Another important limitation is that in some cluster algorithm is a requirement to have previous knowledge of the database to use particular input parameters. For algorithms like k-means is better to have a priori domain knowledge of the dataset to avoid undesirable results.

(c) If the clustering technique derives from a polynomial, the complexity with large datasets means that the algorithm is not scalable.

To overcome the above restrictions may use algorithms that are based on density.

For the density algorithm the clusters have an increased density with respect to the points they possess. Also, any single points around the clusters are outliers and do not belong to any clusters because there been in a low concentration area.

2.4.3.2 DBSCAN: Density Based Spatial Clustering of Applications with Noise

DBSCAN clustering locates regions of high density that are separated from one another by regions of low density. DBSCAN finds clusters of arbitrary shapes, with high quality noiseless output cluster. [10, 32]

The density is estimated for a particular data point by counting the number of point with a distance Eps of that point. The point also counts. The density of any cluster will depend will depend on the defined radius [32].

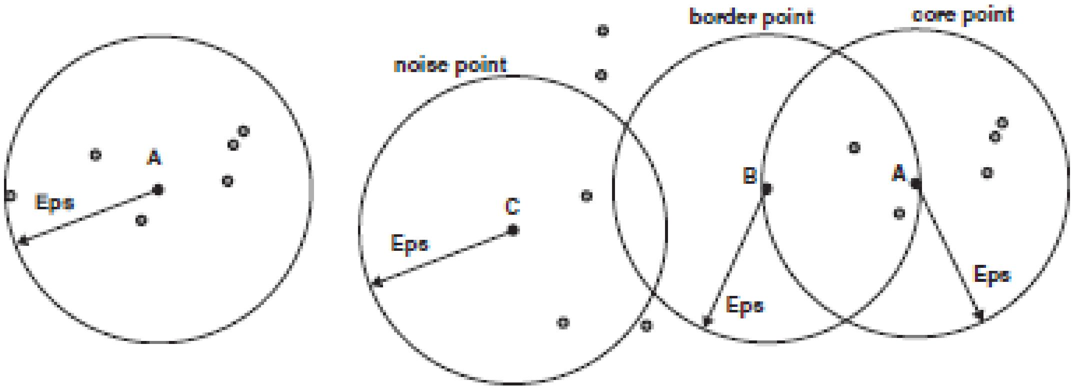


Figure 3 - DBSACAN Core, border and noise point [32]

With the DBSAN algorithm the first step is to label all points as core, border or noise. The second step eliminates the noise points. On the third step put and edge between all core points that are within Eps of each other. The fourth step separate clusters with the connected core points and finally assign each border point to one cluster which it is associated [32].

Definitions

(a) Minimum number of points/Minimum size of a cluster (Minpts): Define whether a neighborhood is denser or not. Define the density threshold of the regions.

(b) Distance of point p within given Eps

Neighborhood point p within Eps value that is referred NEps (p).

$$NEps\ (p) = \{\ q \in D\ |dist(p, q) \geq Minpts\}$$

$Eps: radius$
$D: database$
$q: Core\ point$
$p: Border\ point$
$Minpts: Minimun\ number\ of\ points\ in\ cluster$

(c) Core point q condition

$Number\ of\ NEps(p) \geq Minpts$

(d) Directly density reachable points = core point q and border point p

(e) Core point: Minimum numbers of points are needed within Eps-neighborhood

$$NEps(q) \geq Minpts$$

(f) Border Point: Eps-neighborhood of border point has less point than the Eps of core point. $p \in NEps(q)$

(g) Density reachable points: Point p is referred as density reachable from another point q in order to Eps and Minpts. If there is a connected chain of point

$$p_{n+1}\ is\ directly\ recheable\ from\ p_n$$

(h) Noise: When a point is not a core point, a border point and not belongs to any cluster.

(i) Time Complexity: O(n^2) for each point it has to be determined if it is a core point

(j) n: is the number of objects to be clustered

(k) Time Complexity: O(n2). For each point it has to be determined if it is a core point

(l) Space Complexity: O(n)

The algorithm requires up to two parameters:

- **(a)** Density metric
- **(b)** Minimum size of a cluster

As a result, no need to estimate the number of clusters a priori as in another cluster, like k-means.

Strengths

- **(a)** The algorithm find arbitrary shape cluster.
- **(b)** The algorithm can identify and remove noise from the dataset.
- **(c)** It is requires only two parameter which are mostly insensitive ordering of the point in the database.

Weaknesses

- **(a)** Multi density dataset are not complete by DBSCAN.
- **(b)** Run time complexity is high.
- **(c)** The algorithm can´t process datasets accurately with big densities difference.

2.4.3.3 FDBSCAN: Fast DBSCAN

The aim of the FDBSCAN algorithm is to improve the speed process of the DBSCAN algorithm. While the algorithm is faster than the DBSCAN algorithm, it is loos accuracy in the results [10].

Strengths

(a) Better speed performance than the DBSCAN

Weaknesses

(a) Object loss is higher than the basic DBSCAN.

2.4.3.4 VDBSCAN: Varied density spatial clustering of applications with noise

The algorithm aim of VDBSCAN is remove the input radius parameter, called Eps, of DBSCAN algorithm, in addition to the possibility of finding clusters of varying densities [15, 33].

The VDBSCAN process is divided into two main parts, finding Eps, and finding clusters.

In the first part the algorithm draw a graph that contains the distance of the k-th nearest neighbor of all objects to be analyzed in ascending order. K is the minimum number of points that a cluster must have. The picture shows this algorithm step. Each jump of values in the graph defined an Eps radius [33].

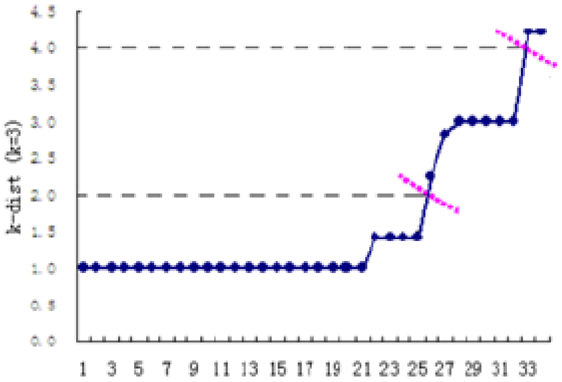


Figure 4 - Chart to find the minimum number of k that a cluster must have

In the second part of the algorithm, the DBSCAN algorithm is executed for each Eps value. This combination makes possible to cluster with different densities.

This algorithm detects cluster with not uniform density and also selects several values of input parameter Eps for different densities. The Minpts is also automatically generated upon the characteristics of the datasets.

The VDBSCAN algorithm provides a set of different Eps for a user-specified MinPts that can recognize clusters with varying density [10].

Algorithm steps:

(a) Choosing parameters Eps

(b) Cluster with varied densities.

Algorithm process:

1. Calculates and stores k-dist for each project and partition the k-dist plots.
2. The number of density is given by k-dist plot.
3. The parameter Eps is selected automatically for each density.
4. Scan the dataset and cluster different densities using corresponding Eps.
5. Display the valid cluster with respect to varied density.

Strengths

(a) Gives different Eps for user specified parameter that recognizes clusters with varying density.
(b) Has the same time complexity as DBSCAN.
(c) Can identify clusters with different density, which is not possible in DBSCAN algorithm.

Weaknesses

(a) When needs a K from users, less accuracy of the algorithm

2.4.3.5 ST-DBSCAN: Spatial-Temporal density based clustering

This algorithm is a modification DBSCAN algorithm and can found clusters respect to non-spatial, spatial and temporal values of the objects [10].

Features

(d) ST-DBSCAN cluster spatial-temporal with non-spatial, spatial and temporal attributes.
(e) To solve conflicts in the border objects compares the average value of a cluster with new incoming value.
(f) DBSCAN can´t detect noise points when the density is heterogeneous, however the ST-DBSCAN can assign density factor to each cluster and overcome the DBSCAN limitation.

Strengths

(a) The Spatial-temporal data is stored as temporal slices of the spatial dataset.
(b) This algorithm can be used in many geographic information systems

Weaknesses

(a) The input is not generated automatically.

Algorithm	Input Parameter	Arbitrary Shape	Varied Density
DBSCAN	Radius, Minpts	YES	NO
FDBSCAN	Radius, Minpts	YES	NO
VDBSCAN	Automatically generated.	YES	YES
ST-DBSCAN	(x,y,t)	YES	NO

2.4.4 Visual analytics approach

A simple and clear visual analysis is necessary to discover patterns and to show the results. The paper “A Visual Analytics Approach for Extracting Spatio-Temporal Urban Mobility Information from Mobile Network Traffic”, written by Günther Sagl, Martin Loidl and Euro Beinat [22], use a data called mobile network traffic.

There is a classification and definition of mobility patterns in:
(a) Temporal Mobility Patterns
(b) Spatial Mobility Patterns
(c) Similarity and Symmetry of Mobility
(d) Exceptional Mobility pattern

They analyze, find and show patterns with the use of charts like [22]:
(a) Areaspline Chart for daily movement flow.
(b) Spatial Density Map for spatial mobility patterns and show by regions. The chart density heatmap is almost the most frequent representation to have a macroscopic view of the geographical or time-spatial distributions of events. It is efficient to show more values and to have comparative view of them. [26]
(c) Polar charts to show the similarity of patterns between different days of the week or regions and finally, perhaps the most interesting, they use Stacked Bar charts to show exceptional mobility patterns. This visualization is useful because the can show a large amount of cells in one graph.

2.4.5 Ping pong handover

To analyze the overlapping phenomenon and understand how it is works is important to understand the ping-pong handover and how mobile companies manage or handle it.

The Ping-Pong handover happens when exists several unwanted handovers of a mobile network in the active connection, from a cell or antenna to the mobile user [27].

The mobile companies “manage” the ping-pong handover to, mainly, avoid overcrowd users in a particular area at any time. The ping-pongs, repeated handovers, could be wanted or unwanted. If the ping-pongs are managed by the cell companies is wanted, otherwise and because the overlapping issue is unwanted [27].

The paper written by Zoltán Fehér, András Veres and Zalán Heszberger called “Ping-pong Reduction using Sub cell Movement Detection” [21] describes a methodology to classify ping-pongs in cellular networks. There is also combined with a method of subcells to detect when handover threshold tuning (pinning) is more effective to apply without increase the risk of latency or failed connections.

They apply the “Sub Cell Movement Detection Method” to verify if the mobile user is moving, and therefore changing antenna. Or if instead is performing ping-pongs when it is stationary.

There are two data sources that should be different if the user is moving or not; one is related with the ping-pong detection (detection of repeated handovers) and the other one is related with movement detection methods.

The term “terminal” defines the movement state of any user, and applies “SFN-CFN Observed Time Difference” method.

$$CFN - SFN = Tm$$

$$Toffset = (SFN1 - SFN2) = Tm1 - Tm2$$

CFN: Cell frame number SFN: System Frame Number

Tm: time arrival of system frames T_{offset}: time difference from two base stations

Previous conclusions:

(g) User moving: If any pair of stations shows change in the T_{offset}.
(h) User stationary: If none of the pairs of stations shows change in the T_{offset}.

The algorithm is based on the handover history.

The algorithm focuses on the short path sequence as:

$$Cell\ A\ \rightarrow Cell\ B\ \rightarrow Cell\ A$$

T_{pp}: time path period. It could take any value, however is has to be a small as possible to show that a ping-pong handover is happening. This paper defined T_{pp} = 30 seconds. For any user we can also define a source and target.

If in a period of time minor or equal than T_{pp}, there are several back and forth and the sequences are concatenated, so there is ping-pong sequence.

There is also possible to define more complex path and ping-pong handovers with more Cell ID and Cell combinations.

The figure shows the, graphically, the ping-pong sequence for just to Cell with a defined T_{pp}.

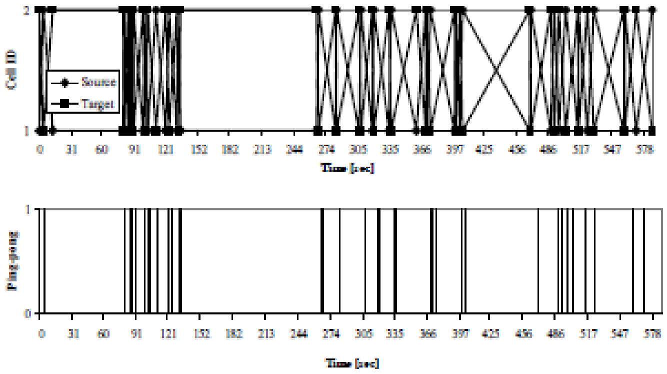


Figure 5 – Ping-pong sequence graph [21]

The next figure shows a two dimension display of the ping-pong handovers that could happened in a period of time. So in this charts we can see the user movement and also the ping-pong handover. The main limitations of these charts are that only shows the transition of two known cells.

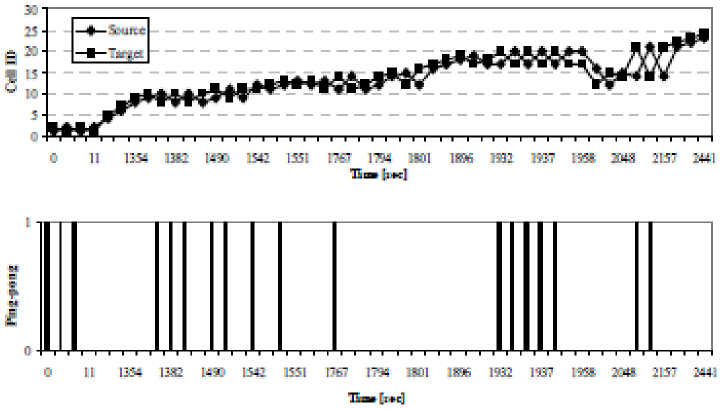


Figure 6 - Time sequence graph to discover ping-pong [19]

To detect which ping pong handover belong or not to a movement episode they use a decision rule engine. This is based on the handover tuning defined by the mobile operator, Key performance indicator and the SON (Self Organization Network Algorithm) [23] [24].

The Spatial-temporal analysis is based on the spatial time series models of a single variable (STARIMA model class), which are the bases for the movement flows [25]. To analyses cell sequence of ping-pong handover the STARIMA approach looks viable, mainly because each LatLong coordinates with his time could be modeled as a linear combination of their previous spatial-temporal tuples.

From the related worked analyses is important to remark that is necessary to create a framework to link the most frequent visit cell from user, the most frequent trajectory user and the ping-pong handover for unwanted movements. None of the related works solve these three issues together. This is the scope of the thesis “Detecting placement and movement episodes from mobile trajectories”.

2.5 Data

The input data of the present thesis are the Cell Plan and the CDR. The combination of these input data are the source for the thesis process analyses. The Cell Plan has data related with the geographical position of the antennas and cells, and also has their modeling representation of polygons. As I mentioned before the polygons are related with the type of cell (GSM, 3G, 4G, etc), power and geographical coordinates. The CDR has information related with the interaction of the user mobile and the network through events.

2.5.1 Cell Plan

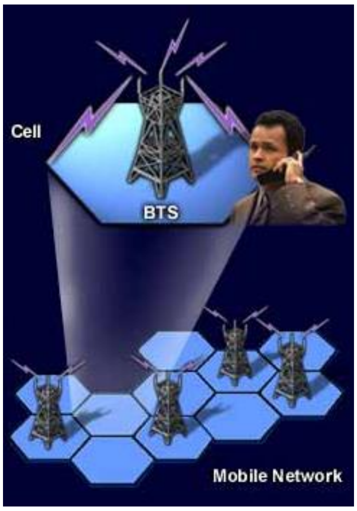


Figure 7 - Nominal Cell plan representation [9]

As I mention before the input data for the thesis analyses came from two sources. The first one, the cell plan, is a graphical representation of a mobile network [9]. The main elements are the layout, cell ID, and some technical data related with the power and frequency of the cell.

A nominal cell plan is often presented as a pattern of hexagonal cells of different sizes; however the input cell plan of this thesis is a Voronoi layout with different number of sides.

2.5.2 Call Detail Record data (CDR)

The other necessary data source to perform mobile time spatial analyses is the CDR (Call Detail Record).

This data input is a passive location data [18] is acquired after a positioning event has occurred. The CDR has all the detailed information related with the user who calls [17] like User ID, Call Origin, Call Destiny, Call Duration, Cell ID and any other data related with the billing which could cares to the mobile company. [17] For this reason is useful for service cost, user habits, user movement or placement, marketing issues fraud detection, etc. [17]

Unfortunately there is not a common or standard format of CDR data because it depends on the mobile service provider criteria, technology to recover data and output format (text file o binary). For this reason it is necessary to parse and perform an ETL process to prepare the raw data.

There is also another consideration with the CDR data, and is related with the availability. So there could be processed in a batch process or it could be a Stream Data.

The Stream Data is a real-time, continuous, ordered sequence of items. It is not possible to manage the arrive order and also is not possible to store locally a stream completely. [13]

The Stream Data should have the follows elements [14]:

(i) Data points arrive continually
(j) Data stream is unbounded
(k) Data points have to be proceeding at once in the arriving order and as fast as the arriving rate.

To use Stream Data for the present thesis, the most efficient way to handle it is to create a batch with a presetting size and then perform the analyses. [17]

There is also another classification of input data and is related with the way to collect it. The Active mobile positioning [17] is used for mobile position. The location of the mobile phone is determined with a radio wave query [14,16]. This procedure requires a permit from the mobile user.

The SPM (Social Positioning Method) [16] study, just for Estonia, how accurate could be an active mobile positioning. Despite of it, the active mobile positioning fit better with a restricted data size mainly because it requires a permit from the mobile user.

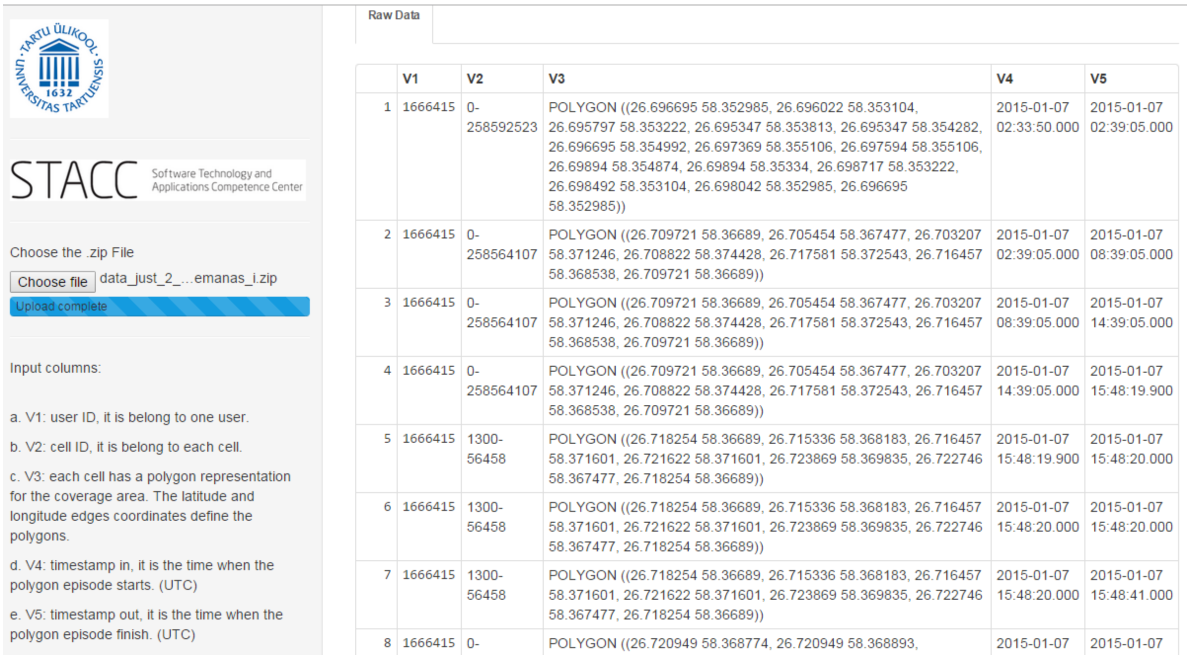
The other way to collect data is the passive mobile positioning. The data is collected automatically by the mobile companies and is like the CDR data [17]. There is not necessary any kind of user permission.

2.5.3 Aggregated input data

The input data is a combination of CDR and Cell Plan data, it is provided by mobile companies. The current amount of data what the framework is performing has five variables and 124828 lines. Each one represents an event. The input data has 19 users and 14 consecutive weeks, starting from 07/01/2015.

The variables are as follows:

(a) User ID: it defines each mobile user
(b) Cell ID: it defines the cell
(c) Movement episode polygon: the raw data has the POLYGON shape with sets of “Latitude and Longitude” coordinates and define a complete movement episode.
(d) Movement episode time in: start date/time of the movement episode.
(e) Movement episode time out: finish data/time of the movement episode.



STACC Software Technology and Applications Competence Center
Choose the .zip File
Choose file
data_just_2_...emanas_i.zip
Upload complete
Raw Data

Input columns:

a. V1: user ID, it is belong to one user.

b. V2: cell ID, it is belong to each cell.

c. V3: each cell has a polygon representation for the coverage area. The latitude and longitude edges coordinates define the polygons.

d. V4: timestamp in, it is the time when the polygon episode starts. (UTC)

e. V5: timestamp out, it is the time when the polygon episode finish. (UTC)

	V1	V2	V3	V4	V5
1	1666415	0-258592523	POLYGON ((26.696695 58.352985, 26.696022 58.353104, 26.695797 58.353222, 26.695347 58.353813, 26.695347 58.354282, 26.696695 58.354992, 26.697369 58.355106, 26.697594 58.355106, 26.69894 58.354874, 26.69894 58.35334, 26.698717 58.353222, 26.698492 58.353104, 26.698042 58.352985, 26.696695 58.352985))	2015-01-07 02:33:50.000	2015-01-07 02:39:05.000
2	1666415	0-258564107	POLYGON ((26.709721 58.36689, 26.705454 58.367477, 26.703207 58.371246, 26.708822 58.374428, 26.717581 58.372543, 26.716457 58.368538, 26.709721 58.36689))	2015-01-07 02:39:05.000	2015-01-07 08:39:05.000
3	1666415	0-258564107	POLYGON ((26.709721 58.36689, 26.705454 58.367477, 26.703207 58.371246, 26.708822 58.374428, 26.717581 58.372543, 26.716457 58.368538, 26.709721 58.36689))	2015-01-07 08:39:05.000	2015-01-07 14:39:05.000
4	1666415	0-258564107	POLYGON ((26.709721 58.36689, 26.705454 58.367477, 26.703207 58.371246, 26.708822 58.374428, 26.717581 58.372543, 26.716457 58.368538, 26.709721 58.36689))	2015-01-07 14:39:05.000	2015-01-07 15:48:19.900
5	1666415	1300-56458	POLYGON ((26.718254 58.36689, 26.715336 58.368183, 26.716457 58.371601, 26.721622 58.371601, 26.723869 58.369835, 26.722746 58.367477, 26.718254 58.36689))	2015-01-07 15:48:19.900	2015-01-07 15:48:20.000
6	1666415	1300-56458	POLYGON ((26.718254 58.36689, 26.715336 58.368183, 26.716457 58.371601, 26.721622 58.371601, 26.723869 58.369835, 26.722746 58.367477, 26.718254 58.36689))	2015-01-07 15:48:20.000	2015-01-07 15:48:20.000
7	1666415	1300-56458	POLYGON ((26.718254 58.36689, 26.715336 58.368183, 26.716457 58.371601, 26.721622 58.371601, 26.723869 58.369835, 26.722746 58.367477, 26.718254 58.36689))	2015-01-07 15:48:20.000	2015-01-07 15:48:41.000
8	1666415	0-	POLYGON ((26.720949 58.368774, 26.720949 58.368893,	2015-01-07	2015-01-07

Figure 8 - Raw data - Input data

2.6 Overview of R

To process and analyse the data and also shows the results we used R software. **R** is also a programming language for statistical computing and graphics.[31].

R functionality is extendable by different libraries and packages. For graph we "googleVis", "plotGoogleMaps", "maptools", "RColorBrewer", "rCharts", "rHighcharts", "leafletR", "gplots", "ggmap", ”colorRamps” and “ggplot2".

For the data parsing and arrangement "data.table", "plyr", "splitstackshape", "R.utils", "reshape”, "grid", "xtable", “reshape2", "lattice", "caTools”, "dplyr").

The statistical analysis is performing by the R base and the following packages "fpc" and "MASS".

Finally the server and html publication is performing by the "shiny" and “rServe”.

3. METHODOLOGY

3.1 System Architecture

3.1.1 Movement / Placement Analyses Framework

The system, in order to get a better understanding of user trajectory and placement/movement patterns, is designed to process different size input data and perform different analyses. It also shows the output in graphs, tables and spatial representations. The data is analyzed according to the following processes (for more details about system functionalities check the Appendix I):

(a) Cell Frequency Analyses
(b) Centroid Analyses
(c) Ping-pong Handover Analyses
(d) Ping-pong Handover Episodes
(e) Cells Overlap – Polygon Intersection
(f) Cells Overlap – Episodes
(g) Cells Overlap – Trajectory map with polygons
(h) Cells Overlap – Episodes Time – Interval Tables
(i) Cells Overlap – Trajectory Analyses – Polygon Distance
(j) Cluster Analyses

The aim of the “Cell Frequency Analyses” is to have an initial approach to find meaningful location from the events for each user. There is a visual map distribution of polygons and centroids by day and by hour. And finally there are tables and graphs with the most visited cell by week and day.

The “Centroid Analyses” helps to understand the nature of data but also aloud to perform a first approach to eliminate outliers. The data is filtered with a Time-distance threshold between consecutive trajectory events. The system has a reactive input to change time-distance filter and compare results.

The variable “centroid time difference” from the “Centroid Analyses” is the main input to perform the “Ping-pong Handover Analyses” between events. A ping-pong handover happens when there is an A-B-A trajectory pattern and the transition time for this pattern is less than a certain value. As a convention, the “pathABA time” threshold is 30 second. It means that any A-B-A trajectory with less or equal than 30 seconds is a ping-pong handover. Identify the ping-pong handovers is important because is an important approach to eliminate outliers or wrong input data.

Each event is defined with Cell ID, User ID, polygon and timestamp. In the trajectory path of each user could exist polygons intersections between consecutive trajectory events. An episode happens when there is an overlapping of more than 20% between consecutive trajectory polygons. The percentage threshold was being defined by convention. The “Cells Overlapp – Polygon Intersection analyses” is the most important part of this thesis because it sets an overlapping matrix to show the different episodes in the trajectory. The matrix also helps to find candidates to be placement or movement episodes.

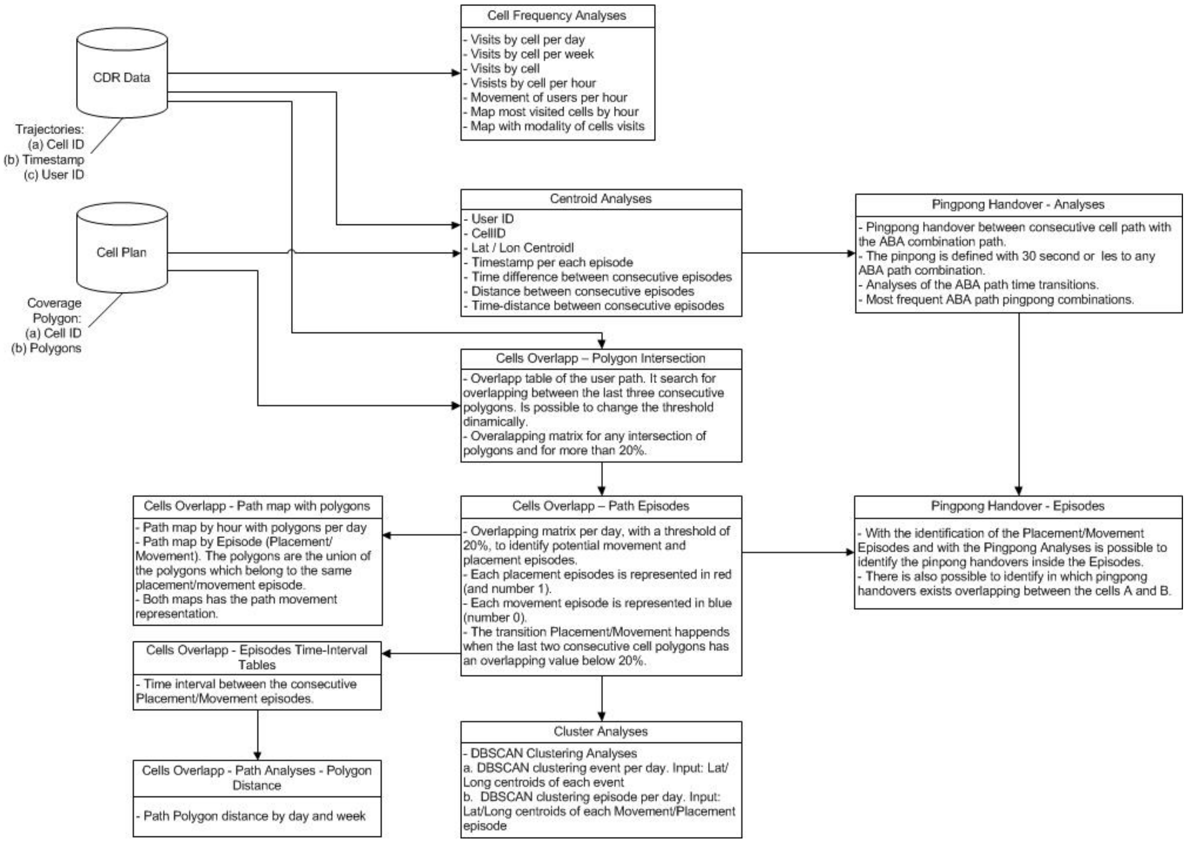

CDR Data
Trajectories:
(a) Cell ID
(b) Timestamp
(c) User ID
Cell Plan
Coverage Polygon:
(a) Cell ID
(b) Polygons
Cell Frequency Analyses
- Visits by cell per day
- Visits by cell per week
- Visits by cell
- Visists by cell per hour
- Movement of users per hour
- Map most visited cells by hour
- Map with modality of cells visits
Centroid Analyses
- User ID
- CellID
- Lat / Lon Centroidl
- Timestamp per each episode
- Time difference between consecutive episodes
- Distance between consecutive episodes
- Time-distance between consecutive episodes
Pingpong Handover - Analyses
- Pingpong handover between consecutive cell path with the ABA combination path.
- The pinpong is defined with 30 second or les to any ABA path combination.
- Analyses of the ABA path time transitions.
- Most frequent ABA path pingpong combinations.
Cells Overlapp – Polygon Intersection
- Overlapp table of the user path. It search for overlapping between the last three consecutive polygons. Is possible to change the threshold dinamically.
- Overalapping matrix for any intersection of polygons and for more than 20%.
Cells Overlapp - Path map with polygons
- Path map by hour with polygons per day
- Path map by Episode (Placement/ Movement). The polygons are the union of the polygons which belong to the same placement/movement episode.
- Both maps has the path movement representation.
Cells Overlapp – Path Episodes
- Overlapping matrix per day, with a threshold of 20%, to identify potential movement and placement episodes.
- Each placement episodes is represented in red (and number 1).
- Each movement episode is represented in blue (number 0).
- The transition Placement/Movement happends when the last two consecutive cell polygons has an overlapping value below 20%.
Pingpong Handover - Episodes
- With the identification of the Placement/Movement Episodes and with the Pingpong Analyses is possible to identify the pinpong handovers inside the Episodes.
- There is also possible to identify in which pingpong handovers exists overlapping between the cells A and B.
Cells Overlapp - Episodes Time-Interval Tables
- Time interval between the consecutive Placement/Movement episodes.
Cells Overlapp - Path Analyses - Polygon Distance
- Path Polygon distance by day and week
Cluster Analyses
- DBSCAN Clustering Analyses
a. DBSCAN clustering event per day. Input: Lat/ Long centroids of each event
b. DBSCAN clustering episode per day. Input: Lat/Long centroids of each Movement/Placement episode

Figure 9 – Detailed System Architecture Overview

The posterior process is the “Cells Overlapp – Path Episodes” and gets the overlapping matrix by day. It clearly identifies the episodes in the trajectory. As the key product of this thesis, the overlapping matrix per day is the main input for the following analysis.

With the different Episodes identified is possible to perform a “Ping-pong handover – Episode” analysis inside each episode trajectory and between them. If there is a ping-pong handover between different episodes, it means that could not be a user possible physical displacement. It is just a ping-pong between cells.

The union of the event polygons, which belong to the same episode, became a new polygon with timestamp and polygon coordinate variables. Then, through “Cells Overlap – Trajectory map with polygons process” the system shows maps like:

(a) Trajectory event maps by week, day, and hour.
(b) Trajectory episodes maps by week, day, and hour.

With these two time-spatial representations is possible to identify which episodes are placement or movement. The analytic tool can filter by period of time the events or the episodes.

There are two subsequent processes, which analyze the trajectory episode as an object. The aim of “Cells Overlapp – Episodes Time interval” and “Cells Overlapp – Path Analysis – Polygon Distance” is to calculate the time-distance between trajectory episodes. With this value is possible to filter, with a time-distance threshold, values that are not physically possible to happen.

The last process possible to perform with this thesis analysis tool is the “Cluster Analyses”. It performs three main DBSCAN clustering:

(a) DBSCAN clustering event per day
(b) DBSCAN clustering episode per day
(c) DBSCAN Clustering Analyses to all days episodes: it perform the analyses with all the same days together.

This approach, gives accurate results to identify different episodes and help to discover placement episodes. The analytic tool developed can perform DBSCAN clustering analysis with reactive inputs like “Distance”, “Minimal Points” and “Zoom”.

For each one of these items is possible to see the results in a map.

3.1.2 Technical solution Framework

A cloud environment is deployed and also is possible to perform simultaneous analyses from different locations. The main aspects from the technical point of view and the implementation of the data analysis tool developed for this thesis are the following.

The technical solution has the follow elements:
(a) Data mining analyses: mainly with R script. The objective is to have more control on the features of the used algorithms.
(b) Data visualization framework: R library/package “Rserve” aloud to R to work as a server. With the R package “Shiny” is possible to create an html framework to deploy WEB visualization.
(c) With the server and html framework solved: The task to show the results properly is possible with html code, the R packages ggplot2, googlevis, rChart and leafletR.
(d) Visualization approach: To improve the geographic visualization features with java and with the js library leaflet.

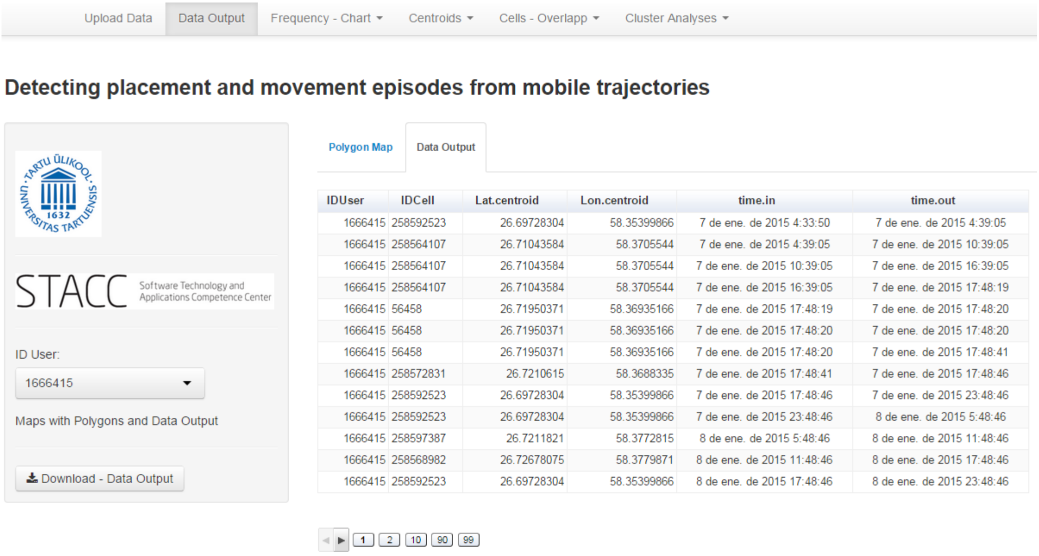
(e) Geographical class or objects: The creation of particular geographical class or objects to work with and time spatial data is a transversal technical requirement. For instance in some cases is necessarily to create spatial polygons or LatLon objects, and in other cases GeoJSON objects.

3.2 Placement / Movement Analyses Methodology

Each one of the previous step analyses has a methodology to calculate the different variables and outputs. Because these processes are linked each other I am going to introduce in the methodology of each one.

3.2.1 Data analysis

The first step is to get the polygon centroid of each event. The aim is to simplify the distance calculus between events.



Detecting placement and movement episodes from mobile trajectories

STACC Software Technology and Applications Competence Center

ID User: 1666415

Maps with Polygons and Data Output

IDUser	IDCell	Lat.centroid	Lon.centroid	time.in	time.out
1666415	258592523	26.69728304	58.35399866	7 de ene. de 2015 4:33:50	7 de ene. de 2015 4:39:05
1666415	258564107	26.71043584	58.3705544	7 de ene. de 2015 4:39:05	7 de ene. de 2015 10:39:05
1666415	258564107	26.71043584	58.3705544	7 de ene. de 2015 10:39:05	7 de ene. de 2015 16:39:05
1666415	258564107	26.71043584	58.3705544	7 de ene. de 2015 16:39:05	7 de ene. de 2015 17:48:19
1666415	56458	26.71950371	58.36935166	7 de ene. de 2015 17:48:19	7 de ene. de 2015 17:48:20
1666415	56458	26.71950371	58.36935166	7 de ene. de 2015 17:48:20	7 de ene. de 2015 17:48:20
1666415	56458	26.71950371	58.36935166	7 de ene. de 2015 17:48:20	7 de ene. de 2015 17:48:41
1666415	258572831	26.7210615	58.3688335	7 de ene. de 2015 17:48:41	7 de ene. de 2015 17:48:46
1666415	258592523	26.69728304	58.35399866	7 de ene. de 2015 17:48:46	7 de ene. de 2015 23:48:46
1666415	258592523	26.69728304	58.35399866	7 de ene. de 2015 23:48:46	8 de ene. de 2015 5:48:46
1666415	258597387	26.7211821	58.3772815	8 de ene. de 2015 5:48:46	8 de ene. de 2015 11:48:46
1666415	258568982	26.72678075	58.3779871	8 de ene. de 2015 11:48:46	8 de ene. de 2015 17:48:46
1666415	258592523	26.69728304	58.35399866	8 de ene. de 2015 17:48:46	8 de ene. de 2015 23:48:46

Figure 10 - Input data with centroid polygons

The centroid of each polygon is calculated with R in two steps:

(a) Create an sp (Spatial polygon) object
The sp object contains all the features of a spatial object.
(b) Calculate the latitude and longitude centroids of each polygon event with the gCentroid function from the “rgeos” R package.

I use the “rgeos” package for the centroid calculation mainly of performance and speed issues.

3.2.2 Polygon map of user trajectory

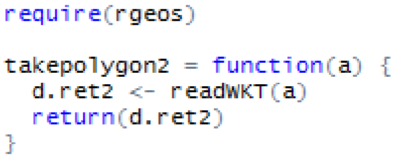
The input data to draw the polygons is defined by the coordinates of every polygons vertex and has the WKT (Well know text) format. The WKT is a markup language to represent vector geometry objects and spatial reference systems. However to run it in R is better to create “sp” spatial objects in order to use his computational capabilities with the rgeos R package.

Example of polygon input data:

POLYGON ((26.696695 58.352985, 26.696022 58.353104, 26.695797 58.353222, 26.695347 58.353813, 26.695347 58.354282, 26.696695 58.354992, 26.697369 58.355106, 26.697594 58.355106, 26.69894 58.354874, 26.69894 58.35334, 26.698717 58.353222, 26.698492 58.353104, 26.698042 58.352985, 26.696695 58.352985))

The sp class doesn't hold data, it only has common information to all the spatial polygons subclasses like bounding box and reference system. Examples of sp subclasses are classes: SpatialPoints, SpatialPolygons, SpatialLines, etc.

To create the sp class objects I run the follow R code:



```
require(rgeos)

takepolygon2 = function(a) {
  d.ret2 <- readWKT(a)
  return(d.ret2)
}
```

With sp object created is easy to perform spatial objects operations with polygons in R.

In the following sections we will show how we can create objects of these classes from scratch or from other classes, and which methods and functions are available for them.

The spatial objects sp created could have a data frame with the coordinates values and is even possible to attach other features like coordinate system. The spatial object is also part of a data frame so it is linked with to the others spatial-temporal features of the present thesis like timestamp, Cell ID, User ID.

For the creation of episodes from the user trajectory is necessary to perform to basic mathematical operations with the polygons:

(a) Intersection
(b) Union

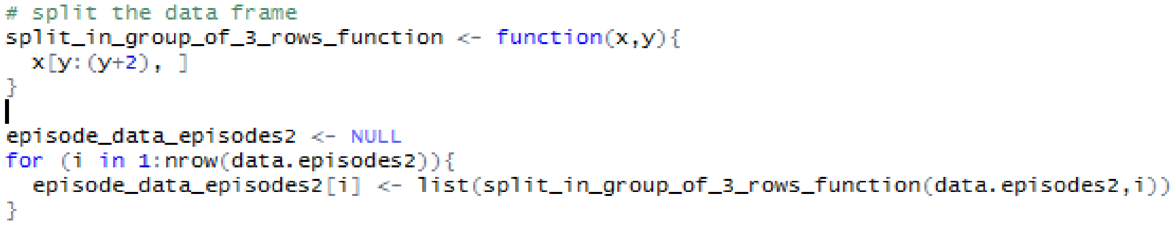
3.2.2.1 Cells Overlapping Analyses – Polygon Intersection operations

This is the key part of the thesis project because the main product is the overlapping matrix. In the previous chapter I mentioned that the episode object is a product of the intersection of consecutive trajectory polygons.

The trajectory is defines by Cell = {Cell$_1$,…, Cell$_i$, Cell$_{i+1}$,…, Cell$_{n-1}$, Cell$_n$}

The intersection between polygons is calculated between the event trajectory of Cell$_i$, Cell$_{i+1}$, Cell$_{i+2}$.

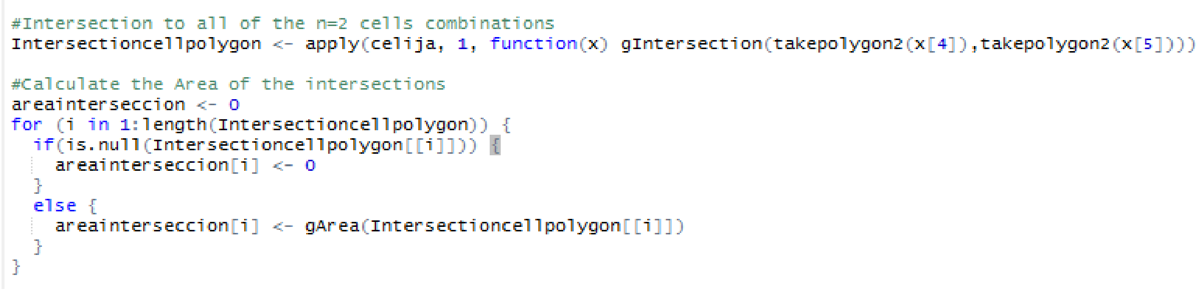
And the algorithm to split the data and generate the Cell$_i$, Cell$_{i+1}$, Cell$_{i+2}$ patterns is the following:



```
# split the data frame
split_in_group_of_3_rows_function <- function(x,y){
  x[y:(y+2), ]
}

episode_data_episodes2 <- NULL
for (i in 1:nrow(data.episodes2)){
  episode_data_episodes2[i] <- list(split_in_group_of_3_rows_function(data.episodes2,i))
}
```

The other part of the process is about to calculate the intersection percentage of the polygons in combinations of two Cell from the three element trajectory groups Cell$_i$, Cell$_{i+1}$, Cell$_{i+2}$ with the following algorithm:



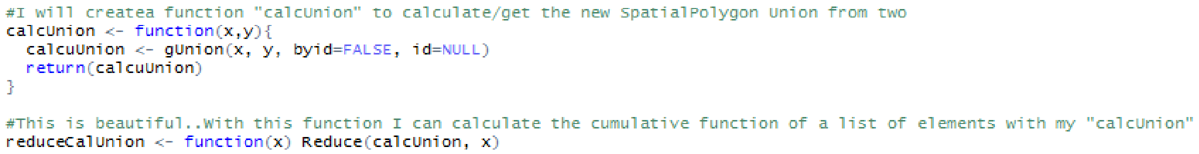
```
#Intersection to all of the n=2 cells combinations
Intersectioncellpolygon <- apply(celija, 1, function(x) gIntersection(takepolygon2(x[4]),takepolygon2(x[5])))

#Calculate the Area of the intersections
areainterseccion <- 0
for (i in 1:length(Intersectioncellpolygon)) {
  if(is.null(Intersectioncellpolygon[[i]])) {
    areainterseccion[i] <- 0
  }
  else {
    areainterseccion[i] <- gArea(Intersectioncellpolygon[[i]])
  }
}
```

3.2.2.2 Path map by hour with polygons – Polygon union operations

An episode is created when there is a polygon intersection between consecutive trajectory events, this episodes could be represented as new polygon which is defined as the union of all the event polygon which belong to the episode.

If there is any intersection between the events trajectory, according with the previous algorithm, is possible to calculate a new polygon, which is the union of all the same, event polygons.



```
#I will createa function "calcUnion" to calculate/get the new SpatialPolygon Union from two
calcUnion <- function(x,y){
  calcuUnion <- gUnion(x, y, byid=FALSE, id=NULL)
  return(calcuUnion)
}

#This is beautiful..With this function I can calculate the cumulative function of a list of elements with my "calcUnion"
reduceCalUnion <- function(x) Reduce(calcUnion, x)
```

3.2.3 Centroids – Create new variables to perform analyses

This step is about parsing and reshapes the input data to later split and get different views from data.

As part of the reshape data process I created new variables. I identified the “time in” and “time out” of each event to calculate the time transition between events, this variable calls “time_difference_centroids”.

I use the Cartesian distance algorithm to calculate the coordinate distance between consecutive centroid events in the trajectory Whit this algorithm is possibly to have an accurate distance value because there is a correction with the earth radio. This variable calls “distance_centroids”.

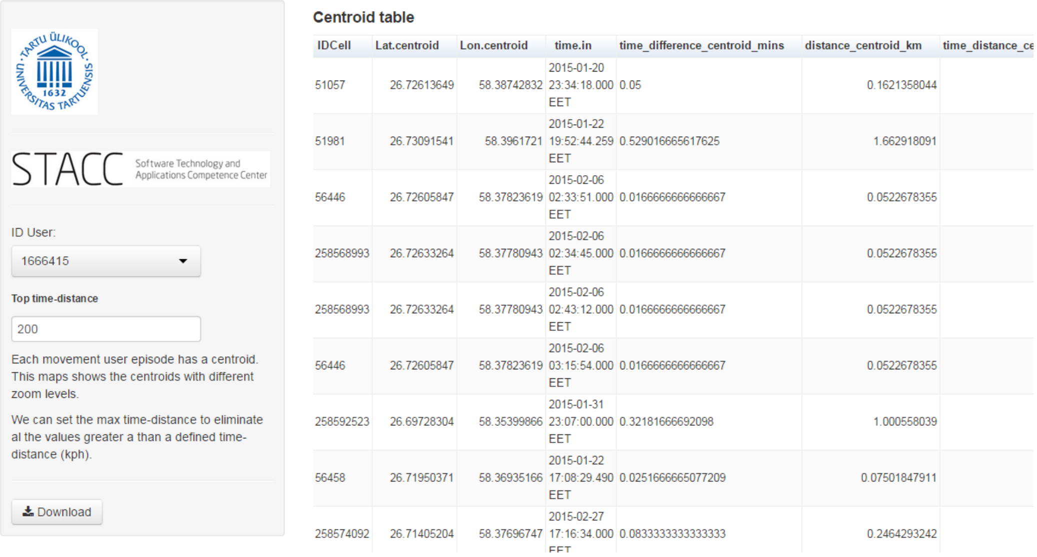
The combination of these variables is the “time_distance_centroid”, this is the first simple analyses to identify possible outliers in the data set.

$$time\ distance\ centroid\ (\frac{km}{h}) = \frac{distance\ centroids\ (km)}{time\ difference\ centroids\ (h)}$$

The system has time-distance threshold value that can exclude transition events. These outliers are the time-distance trajectory transitions, which are too big to be considering part of the real user movement.

For instance the time-distance values over 200 km/h could be definitely consider as outliers because it wouldn´t be a real user movement value.

The output table for this analysis is shown in the figure 11.



ID User: 1666415

Top time-distance: 200

Each movement user episode has a centroid. This maps shows the centroids with different zoom levels.

We can set the max time-distance to eliminate al the values greater a than a defined time-distance (kph).

Centroid table

IDCell	Lat.centroid	Lon.centroid	time.in	time_difference_centroid_mins	distance_centroid_km	time_distance_ce
51057	26.72613649	58.38742832	2015-01-20 23:34:18.000 EET	0.05	0.1621358044	
51981	26.73091541	58.3961721	2015-01-22 19:52:44.259 EET	0.529016665617625	1.662918091	
56446	26.72605847	58.37823619	2015-02-06 02:33:51.000 EET	0.0166666666666667	0.0522678355	
258568993	26.72633264	58.37780943	2015-02-06 02:34:45.000 EET	0.0166666666666667	0.0522678355	
258568993	26.72633264	58.37780943	2015-02-06 02:43:12.000 EET	0.0166666666666667	0.0522678355	
56446	26.72605847	58.37823619	2015-02-06 03:15:54.000 EET	0.0166666666666667	0.0522678355	
258592523	26.69728304	58.35399866	2015-01-31 23:07:00.000 EET	0.32181666692098	1.000558039	
56458	26.71950371	58.36935166	2015-01-22 17:08:29.490 EET	0.0251666665077209	0.07501847911	
258574092	26.71405204	58.37696747	2015-02-27 17:16:34.000 EET	0.0833333333333333	0.2464293242	

Figure 11 - Centroid analyses and time-difference threshold

3.2.4 Ping-pong Handover

It is important for this thesis project to understand the nature of data and work with different approaches to analyze it. A ping-pong handover happens when a particular trajectory is repeated in a certain period of time. For this thesis I defined the trajectory sequence A-B-A and a period time of 30 seconds.

The ping-pong handover must meet two conditions:

(a) Trajectory pattern A-B-A
(b) Trajectory A-B-A time = 30 seconds

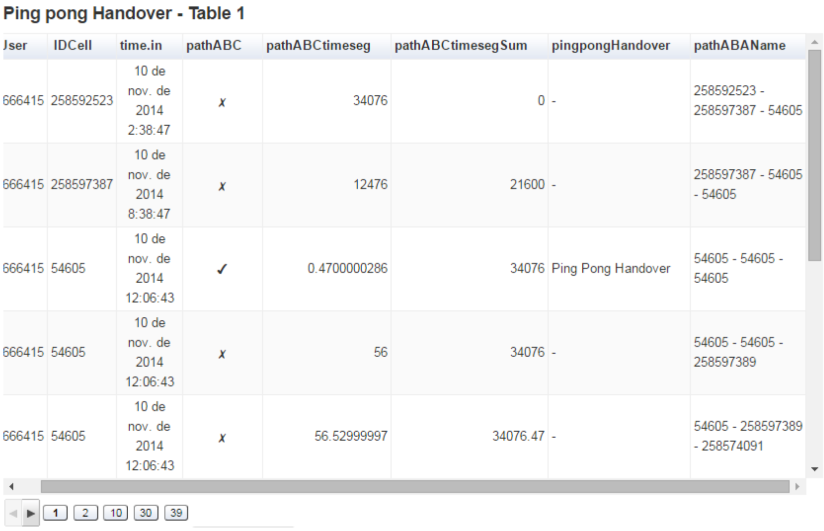
I set these two conditions to meet, the first one is because I don´t have data about radio frequency intensity or signal characteristics. And the trajectory ABA time of 30 seconds is defined as a jump between cells that is not to happen with the user movement.

In order to discover ping-pong handovers in the data set by user,

(a) Calculated the transition time between consecutive events
(b) Get the cumulative transition time of the trajectory events.
(c) Identify the events with the trajectory pattern A-B-A.
(d) And finally identify the trajectory, which meets the ping-pong handover conditions.

The ping-pong handover analysis is very important because is useful to identify each events which not represent real user movements.

According with the episode analyses, if there is a ping-pong handover situation between two different episodes, with the elimination of the jump cell event there is a possibility that the episodes were not actually separated.



Ping pong Handover - Table 1

Jser	IDCell	time.in	pathABC	pathABCtimeseg	pathABCtimesegSum	pingpongHandover	pathABAName
666415	258592523	10 de nov. de 2014 2:38:47	✗	34076	0	-	258592523 - 258597387 - 54605
666415	258597387	10 de nov. de 2014 8:38:47	✗	12476	21600	-	258597387 - 54605 - 54605
666415	54605	10 de nov. de 2014 12:06:43	✓	0.4700000286	34076	Ping Pong Handover	54605 - 54605 - 54605
666415	54605	10 de nov. de 2014 12:06:43	✗	56	34076	-	54605 - 54605 - 258597389
666415	54605	10 de nov. de 2014 12:06:43	✗	56.52999997	34076.47	-	54605 - 258597389 - 258574091

Figure 12 - Ping-pong handover table

Ping-pong handover algorithm:

$$for\ \left(i\ in\ 1{:}\,nrow\left(data_{to_{pingpong}}\right)\right)\left\{data_{to_{pingpong\$pingpongHandover[i]}} < -\,ifelse\left(\left(data_{to_{pingpong\$pathABC[i]}} == TRUE\right)\&\&\left(data_{to_{pingpong\$pathABCtimeseg[i]}} < 3.000000e+01\right),\ \text{Ping Pong Handover, -}\right)\right\}$$

3.2.5 Cluster Analyses

The implementation of the DBSCAN algorithm were through the R package fpc, which can perform the DBSCAN algorithm with the following characteristics.

dbscan(data, eps, MinPts) and time complexity is o(n^2)

The clusters require MinPts within a maximum distance (eps) around one of its members. Any point within eps around any point, which satisfies the seed condition, is a cluster member.

Noise: when points may not belong to any clusters.

With this algorithm were developed a reactive tools change DBSCAN input and perform several test. We can change the “Distance” and “Minimal points”. So it is possible to adjust the clustering results.

The output is shown in a 2 dimensional chart with the latitude and longitude as axes.

There is also available a map output, which is similar than the chart but with a map representation layer. It is also possible to set chart size and the zoom.

As I mentioned in the first part of the methodology chapter three DBSCAN analyses are available.

The DBSCAN clustering analyses has three steps in order to compare the results and show the improvements in the definition of placement and movement episodes detection.

(a) DBSCAN Clustering event analyses by day
(b) DBSCAN Episode Movement/Placement Clustering analyses
(c) DBSCAN Clustering Analyses to all days episodes

3.2.5.1 DBSCAN Clustering event analyses by day

The first DBSCAN clustering for every user is by day. This clustering is helpful, through the clustering centroid to found meaning full locations. In the output chart and map we can see the clustering centers and the points, which belong to the center. In this representation the outliers scatter are not shown.

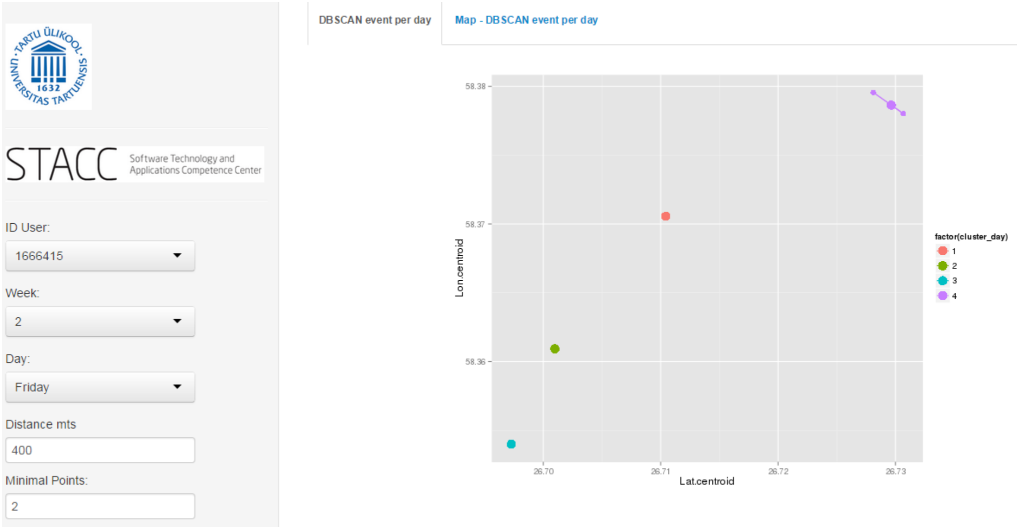

ID User: 1666415
Week: 2
Day: Friday
Distance mts 400
Minimal Points: 2

Figure 13 - DBSCAN Clustering event analyses by day

3.2.5.2 DBSCAN episode Movement/Placement clustering analyses by day

The second step, in order to get accurate results and a better definition of meaningful locations is the DBSCAN clustering episode by day. In this case the events belong to a Movement/Placement episode. I repeated the analyses with the same inputs values.

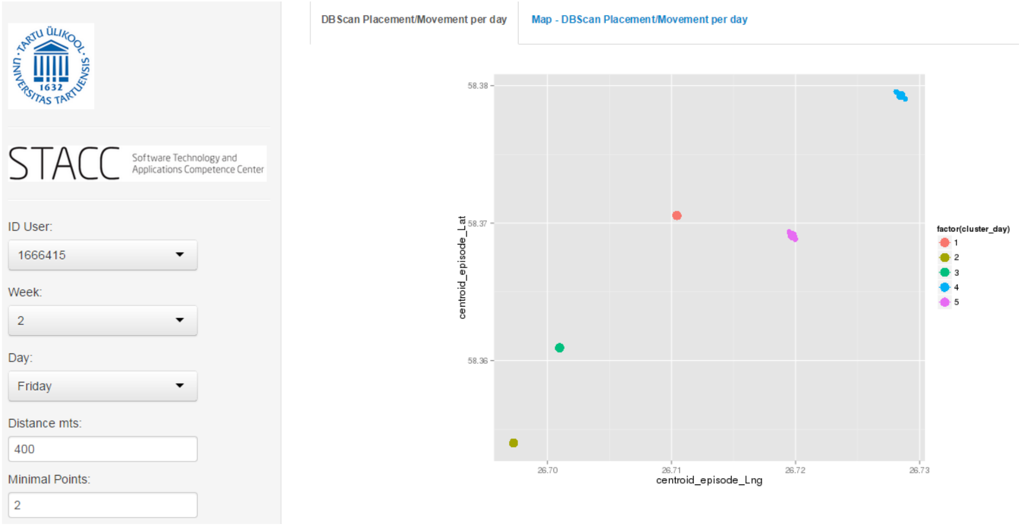
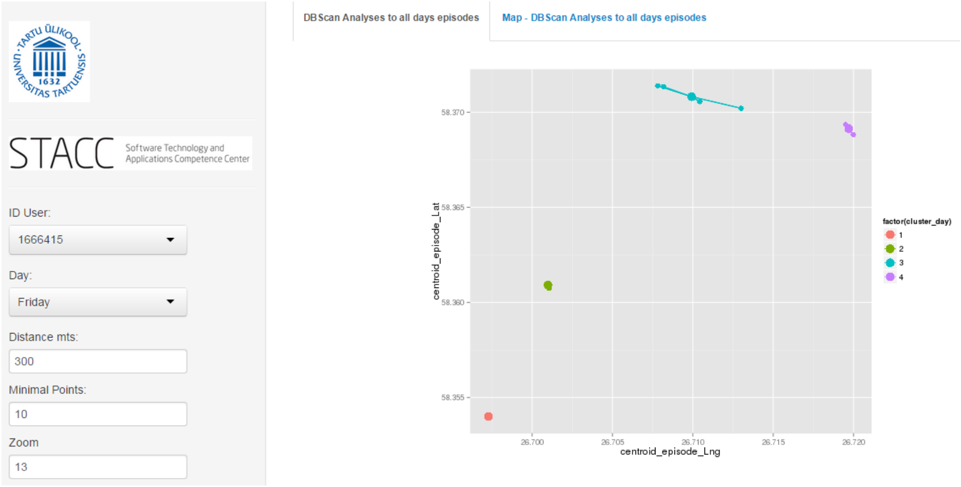

ID User: 1666415
Week: 2
Day: Friday
Distance mts: 400
Minimal Points: 2
STACC
Software Technology and Applications Competence Center
DBScan Placement/Movement per day
Map - DBScan Placement/Movement per day
58.38
58.37
58.36
centroid_episode_Lat
26.70
26.71
26.72
26.73
centroid_episode_Lng
factor(cluster_day)
1
2
3
4
5

Figure 14 - DBSCAN Episode clustering analyses by day

3.2.5.3 DBSCAN Clustering analyses to all days episodes

The last DBSCAN analyses is the most important because it process scatter which belong to episodes, so they represent placement or movement; and also because the analyses is on all the same day together for all the weeks analyses. This is means that is the optimal input to identify patterns from placement and movement episodes.


TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632
STACC Software Technology and Applications Competence Center
DBScan Analyses to all days episodes
Map - DBScan Analyses to all days episodes
58.370
58.365
58.360
58.355
centroid_episode_Lat
26.700
26.705
26.710
26.715
26.720
centroid_episode_Lng
factor(cluster_day)
● 1
● 2
● 3
● 4

ID User: 1666415

Day: Friday

Distance mts: 300

Minimal Points: 10

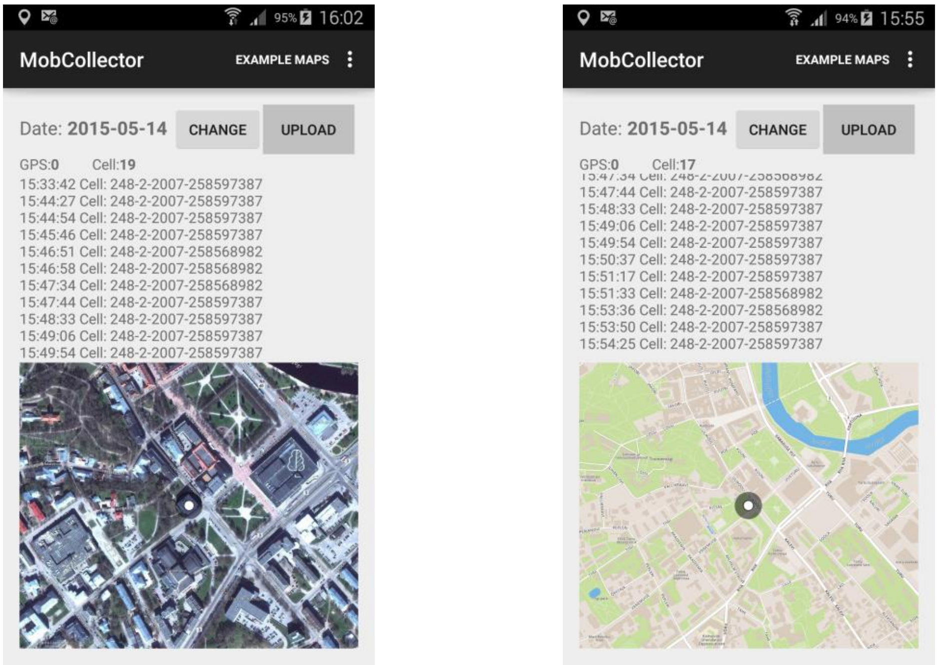
Zoom: 13

Figure 15 - DBSAN Clustering to all days episodes

4. RESULTS AND DISCUSSION

4.1 Polygon map of user trajectory

When starting with this placement and movement episodes detection thesis, one of the concern was the polygons input that represent the coverage area of the cell. However there are some wrong data related with the coverage area and azimuth. The mobile app, developed by the Distributed System Group, shows this situation.



Date: 2015-05-14

GPS:0 Cell:19

```
15:33:42 Cell: 248-2-2007-258597387
15:44:27 Cell: 248-2-2007-258597387
15:44:54 Cell: 248-2-2007-258597387
15:45:46 Cell: 248-2-2007-258597387
15:46:51 Cell: 248-2-2007-258568982
15:46:58 Cell: 248-2-2007-258568982
15:47:34 Cell: 248-2-2007-258568982
15:47:44 Cell: 248-2-2007-258597387
15:48:33 Cell: 248-2-2007-258597387
15:49:06 Cell: 248-2-2007-258597387
15:49:54 Cell: 248-2-2007-258597387
```

Date: 2015-05-14

GPS:0 Cell:17

```
15:47:34 Cell: 248-2-2007-258568982
15:47:44 Cell: 248-2-2007-258597387
15:48:33 Cell: 248-2-2007-258597387
15:49:06 Cell: 248-2-2007-258597387
15:49:54 Cell: 248-2-2007-258597387
15:50:37 Cell: 248-2-2007-258597387
15:51:17 Cell: 248-2-2007-258597387
15:51:33 Cell: 248-2-2007-258568982
15:53:36 Cell: 248-2-2007-258568982
15:53:50 Cell: 248-2-2007-258597387
15:54:25 Cell: 248-2-2007-258597387
```

Figure 16 - MobCollector from Distributed System Group to compare GPS against Cell coverage

The figure 36 shows the result of the comparison between polygon coverage and GPS data. There are two coverage area representation polygons, blue and red. The scatter, red and blue, are the GPS collected data, which should belong to the same Cell. There are some GPS data, which is not inside the polygon coverage area.

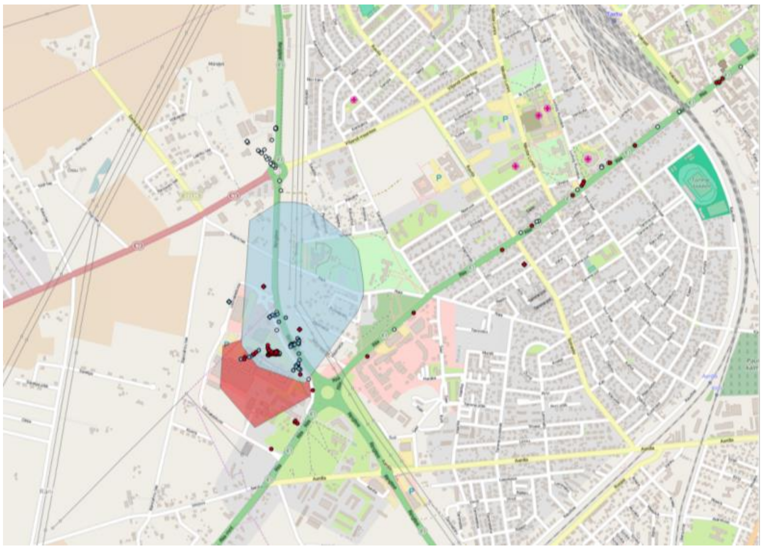
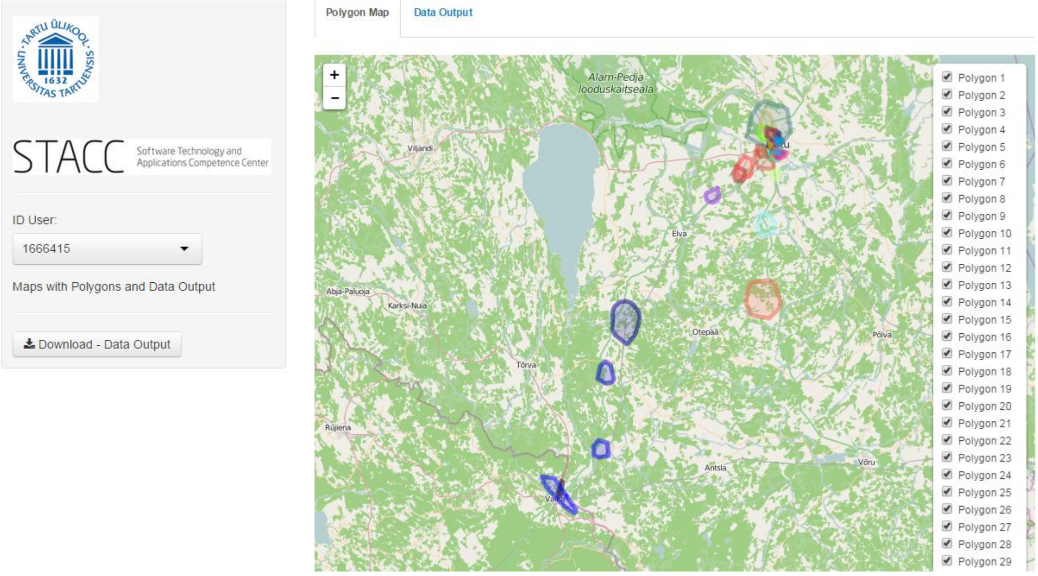


Figure 17 - GPS data against Cell polygons coverage

The overlapping matrix algorithm on the trajectory user assumes the users are inside of the polygon coverage area. A future work could introduce enhancements to the polygon coverage area modelling.

It is important to improve the data understanding to perform the spatial representation of results. It gives an easy comprehension of the events distribution. In figure 9, is possible to see the trajectory polygons of each user for all the weeks.



ID User:

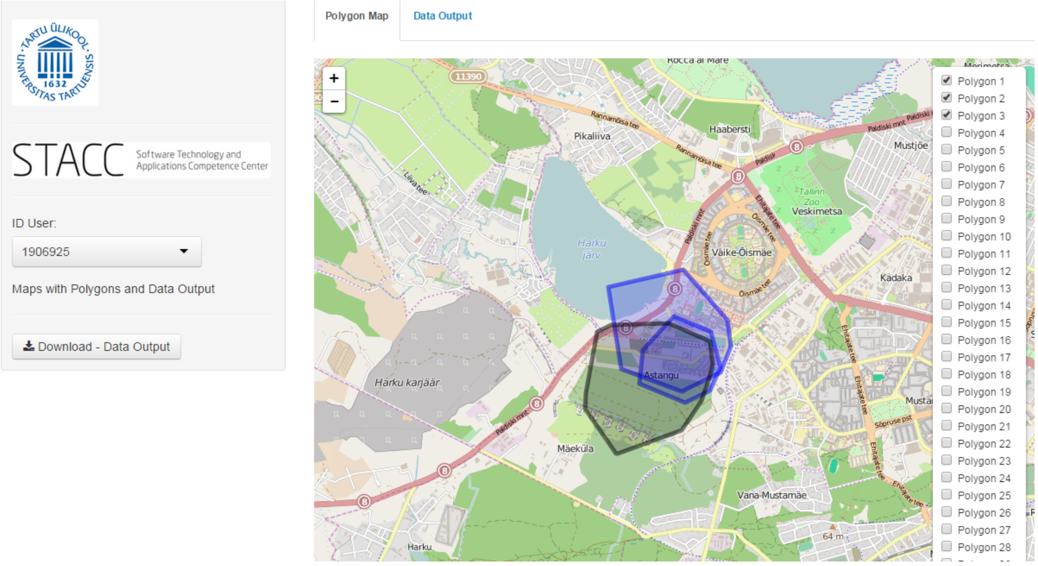
1666415

Maps with Polygons and Data Output

Download - Data Output

Figure 18 - Polygon map of user trajectory

The consecutive trajectory polygons, which have overlapping, are candidates to be part of an episode. This situation is shown in the figure 10.



ID User:

1906925

Maps with Polygons and Data Output

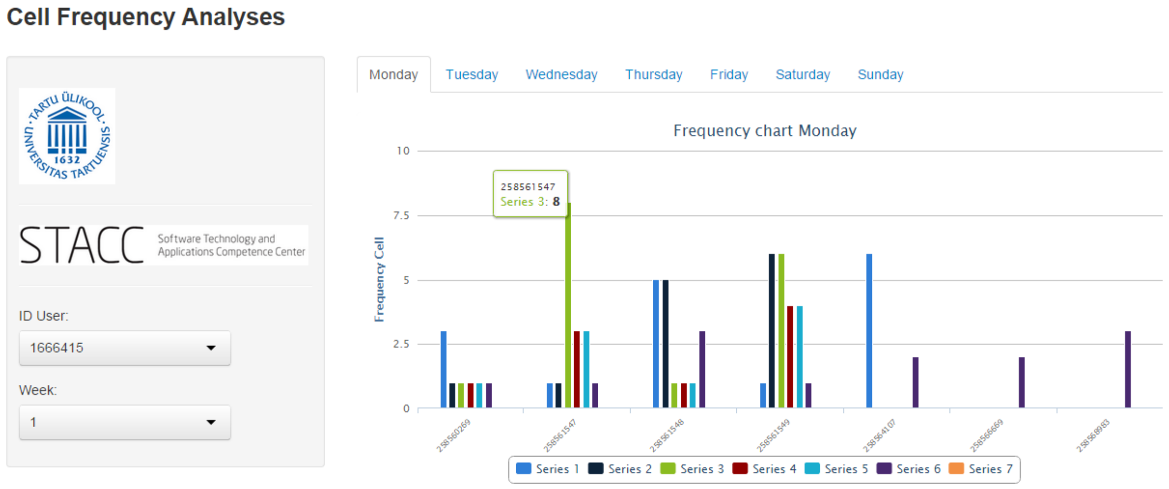
Download - Data Output

Figure 19 - Trajectory polygon overlapping

4.2 Frequency analyses

4.2.1 Most frequent cells visits by day

With this visual approach is easy to identify and classified the most frequent visited cells. Each one of the series is a week. So, if a cell has a high frequent visits over the same day in different weeks could show that there is a pattern in it.



Cell Frequency Analyses

ID User: 1666415

Week: 1

Figure 20 - Most visited cell by day

4.2.2 Most frequent cells visits by week

Another approach from the same data is to analyze the most visited cells by week. In the picture 13 we can see that is possible to compare, the most frequent visited cell by day.

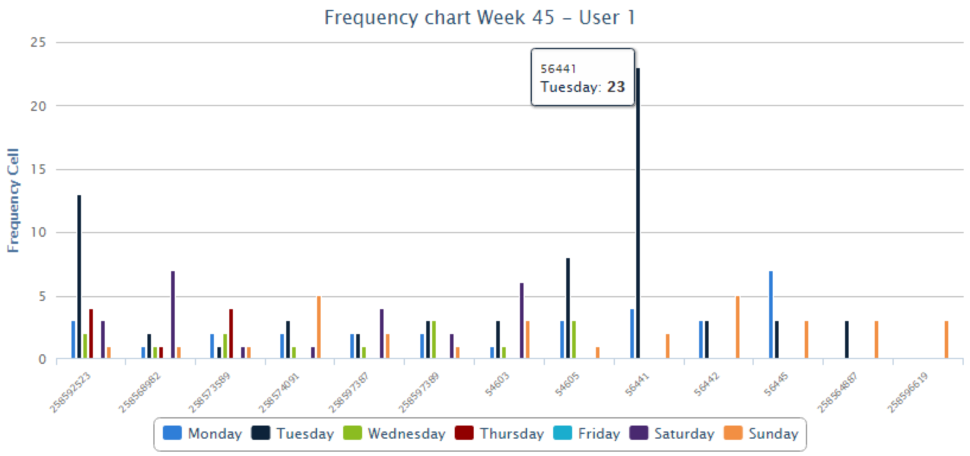


Figure 21 - Most visited cell by week

4.2.3 Frequency visited cells by Cell ID

In order to continue with aggregation of data, is possible to calculate the number of cell visited to all data user. In the figure 14 the most visited cell is the Cell ID number 56445, in posterior analyses we will see the location in map and when is most frequent for the user to stay in this Cell.

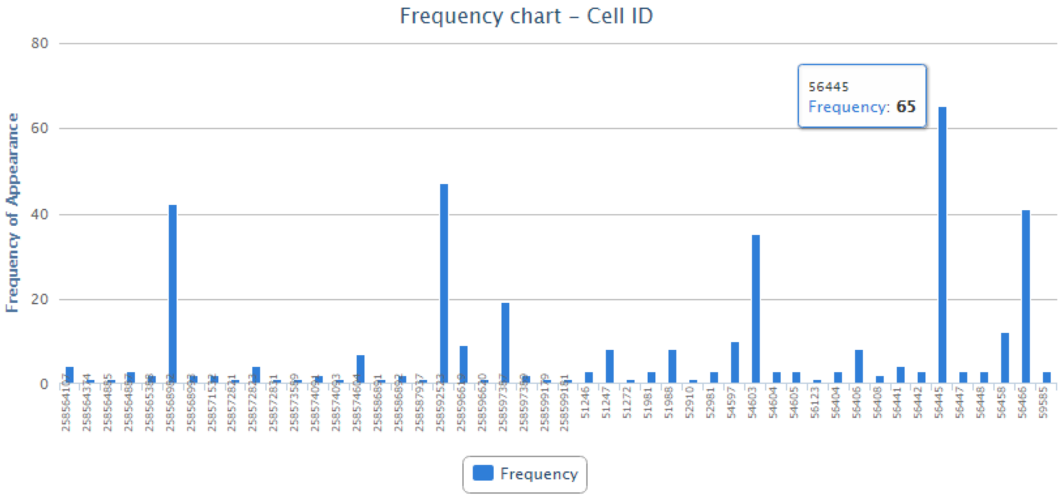


Figure 22 - Visited cells by Cell ID

4.3 Map visits by hour - Centroids

This map is very important in order to understand possible placement and pattern. It combines three elements, most visits by Cell ID, spatial distribution and modality by hour. The figure 15, 16 and 17 help to understand where the most frequent Cell is visited and when it happens.

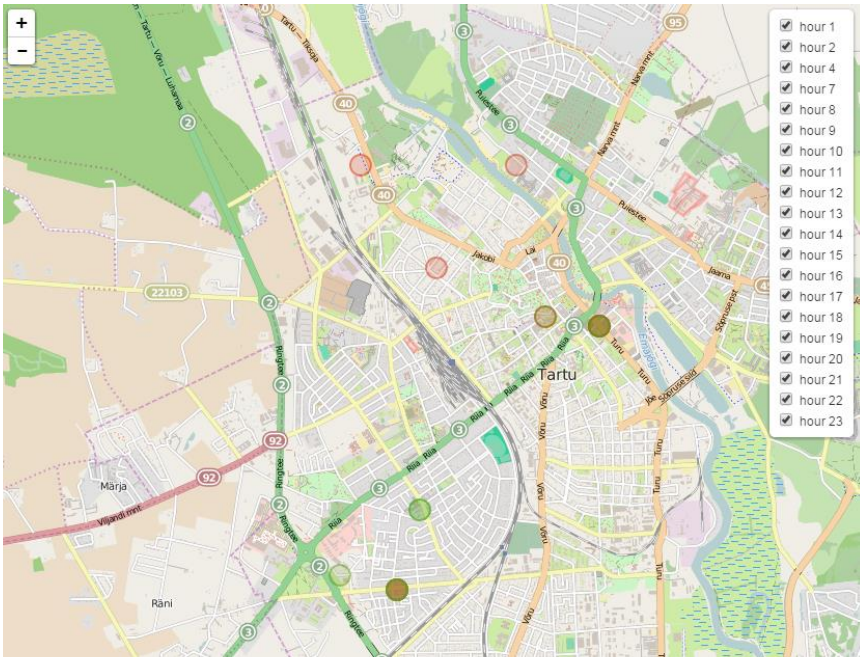
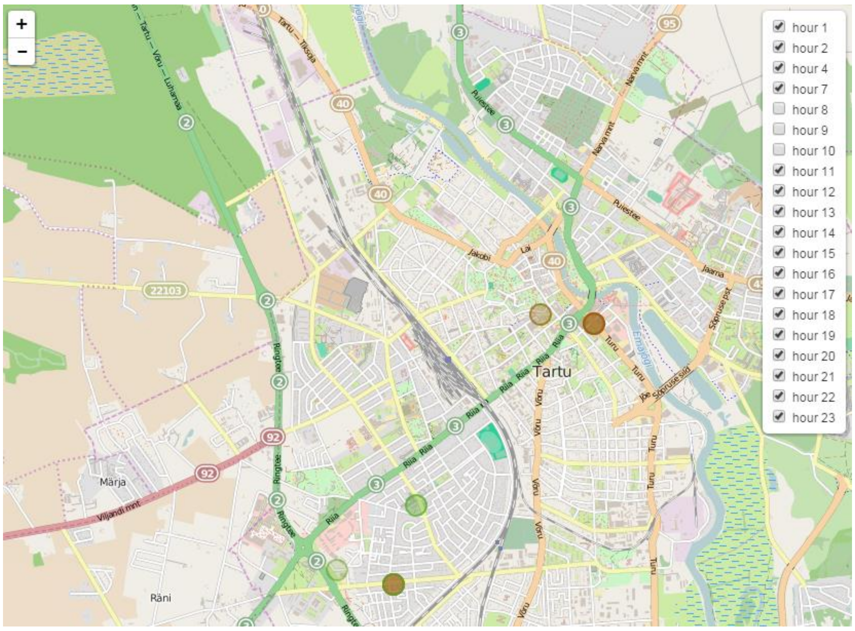


Figure 23 - Map most visited Cell by hour

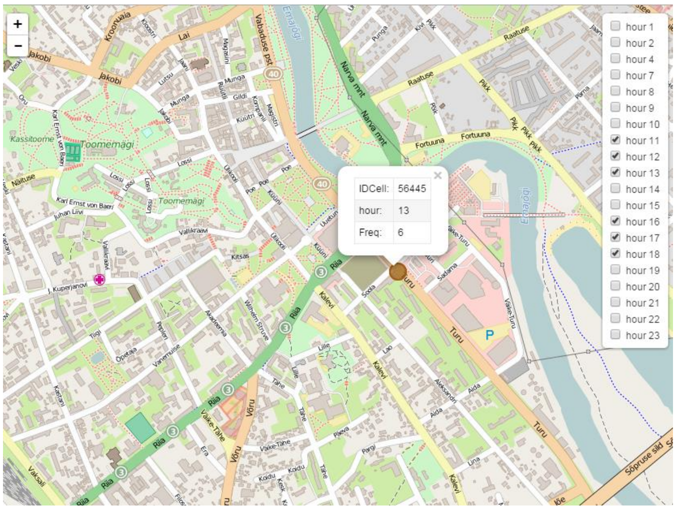
The intensity color starts to define a user pattern. The aim of this thesis is not to label placements however we can see that with the combination of modality and frequency is useful to infer HOME and WORK.



☑ hour 1
☑ hour 2
☑ hour 4
☑ hour 7
□ hour 8
□ hour 9
□ hour 10
☑ hour 11
☑ hour 12
☑ hour 13
☑ hour 14
☑ hour 15
☑ hour 16
☑ hour 17
☑ hour 18
☑ hour 19
☑ hour 20
☑ hour 21
☑ hour 22
☑ hour 23

Figure 24 - Map to show when and where been the most visited cells

As I mention before the ID cell “56445” is the most frequent visited Cell and with figure 17 is possible to locate geographically. There is also possible to discover when it happens.



IDCell: 56445
hour: 13
Freq: 6

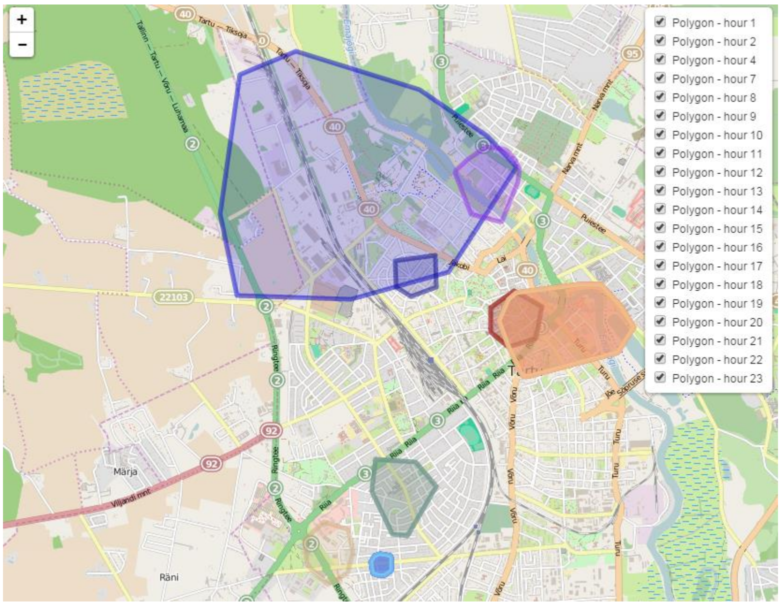
□ hour 1
□ hour 2
□ hour 4
□ hour 7
□ hour 8
□ hour 9
□ hour 10
☑ hour 11
☑ hour 12
☑ hour 13
□ hour 14
□ hour 15
☑ hour 16
☑ hour 17
☑ hour 18
□ hour 19
□ hour 20
□ hour 21
□ hour 22
□ hour 23

Figure 25 - Map - Most Visited cell for a particular user

4.4 Map visits by hour - Polygons

The shape and size of the polygon representation depends on the coverage and type signal. In the figure 18 we can see, as in centroid maps, where are the most frequent polygons and also is possible to observe overlapping situations.

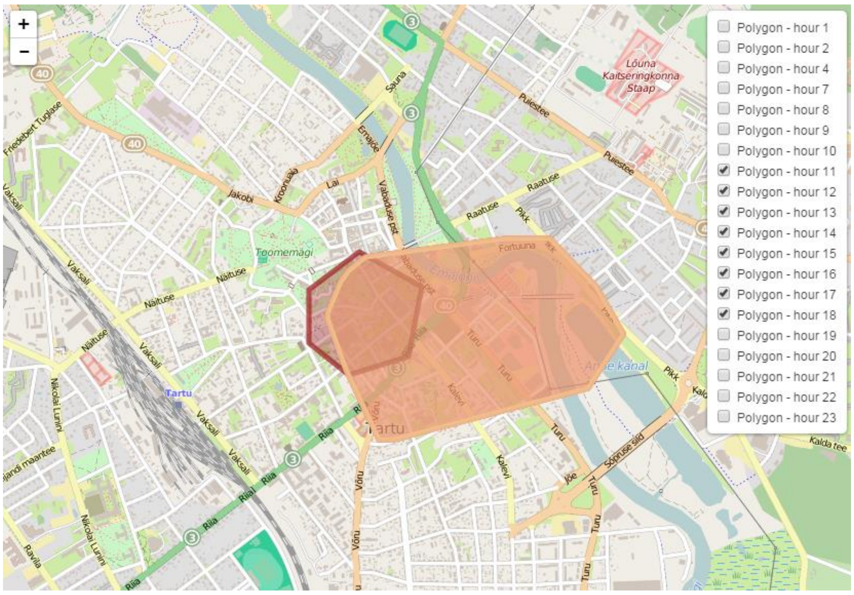
With the “Map visits by hour – Polygons” we have the analyses combination of four elements, most visits by Cell ID, spatial distribution, modality by hour and overlapping.



☑ Polygon - hour 1
☑ Polygon - hour 2
☑ Polygon - hour 4
☑ Polygon - hour 7
☑ Polygon - hour 8
☑ Polygon - hour 9
☑ Polygon - hour 10
☑ Polygon - hour 11
☑ Polygon - hour 12
☑ Polygon - hour 13
☑ Polygon - hour 14
☑ Polygon - hour 15
☑ Polygon - hour 16
☑ Polygon - hour 17
☑ Polygon - hour 18
☑ Polygon - hour 19
☑ Polygon - hour 20
☑ Polygon - hour 21
☑ Polygon - hour 22
☑ Polygon - hour 23

Figure 26 - Map visits by hour - Polygons

With the overlapping polygons, I start to analyze the trajectory in terms of episodes. In the previous analyses, with centroids, we saw two closed centroids in the time interval. There is an episode if the polygons which belong to this centroid intersect, and if the polygons are consecutive in the trajectory. The figure 19 shows this situation with the cell "56445".



☐ Polygon - hour 1
☐ Polygon - hour 2
☐ Polygon - hour 4
☐ Polygon - hour 7
☐ Polygon - hour 8
☐ Polygon - hour 9
☐ Polygon - hour 10
☑ Polygon - hour 11
☑ Polygon - hour 12
☑ Polygon - hour 13
☑ Polygon - hour 14
☑ Polygon - hour 15
☑ Polygon - hour 16
☑ Polygon - hour 17
☑ Polygon - hour 18
☐ Polygon - hour 19
☐ Polygon - hour 20
☐ Polygon - hour 21
☐ Polygon - hour 22
☐ Polygon - hour 23

Figure 27 - Map trajectory polygon intersection

4.5 Movement of users

Another analysis available from the most visited cell approach is to get the pattern of displacement by day of the user with the most frequent cell visits. The figure 20 shows when the user stays in a placement, when he moves and what distance he travel in average.

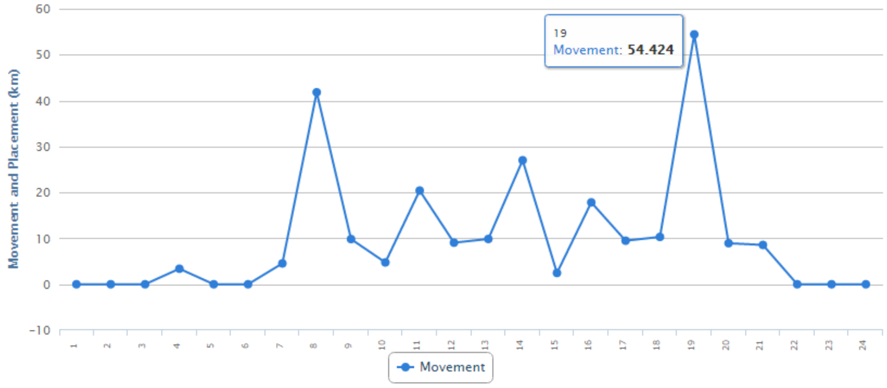


Figure 28 - Movement of user by hour

4.6 Ping-pong Handover

The figure 22 shows a graphical representation of the ping-pong phenomenon. The horizontal lines represent candidates to be ping-pong handover of trajectory A-B-A.

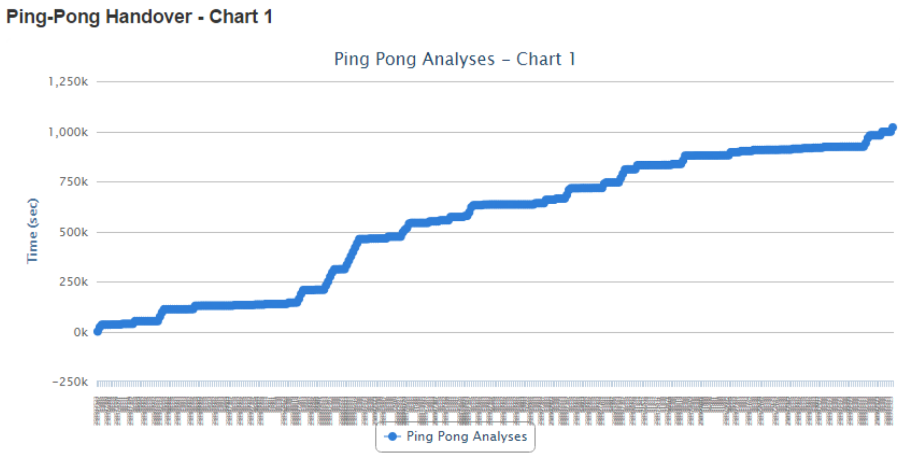


Figure 29 - Trajectory cumulative transition time

The figure 23 shows the transition time of the ping-pong handovers, this graphs is useful for subsequent analyses, with the objective to adjust the ping-pong handover transition time.

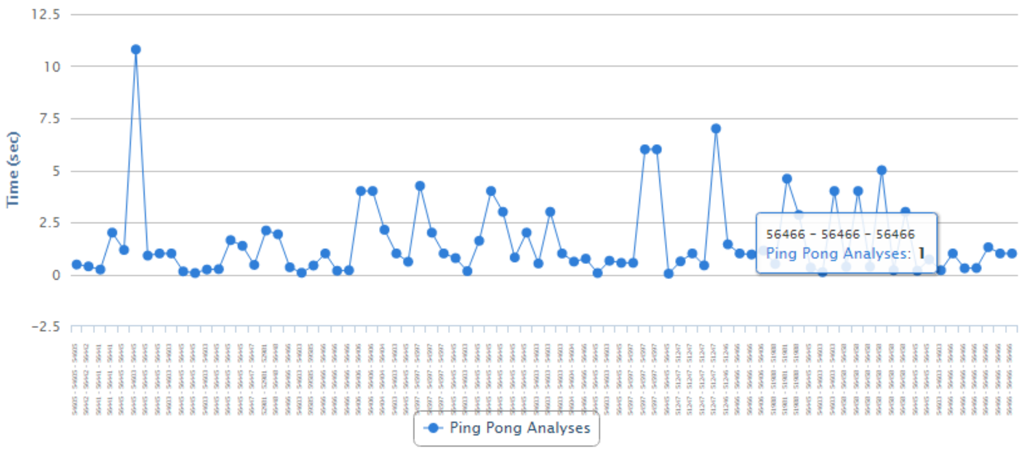


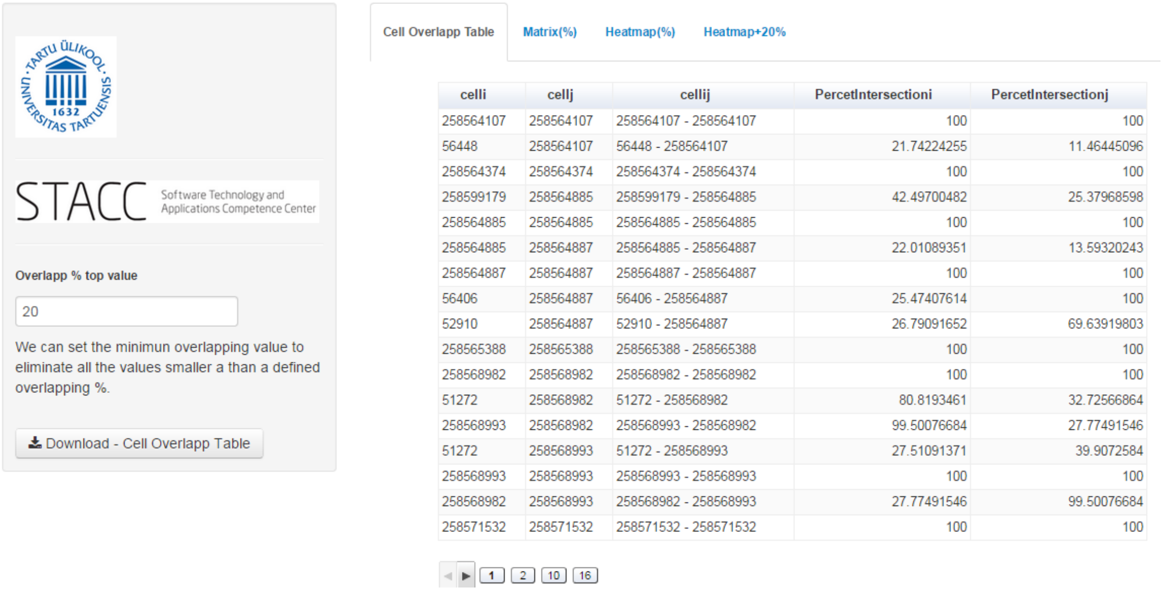
Figure 30 - Ping-pong handover transition time

If the ping-pong handover happens inside the episode, the ping pong handover has no effect on the episode.

For every user more than the 90% of the ping-pong handover cases has the pattern A-A-A. This is means that belongs to the same episode.

4.7 Cells Overlapping Analyses

In order to have a more accurate overlapping results we can a applied a threshold of 20% intersection to the polygon. However the analytics tools can reactive change the intersection value.


STACC Software Technology and Applications Competence Center
Cell Overlapp Table
Matrix(%)
Heatmap(%)
Heatmap+20%
1
2
10
16

Overlapp % top value

20

We can set the minimun overlapping value to eliminate all the values smaller a than a defined overlapping %.

Download - Cell Overlapp Table

celli	cellj	cellij	PercetIntersectioni	PercetIntersectionj
258564107	258564107	258564107 - 258564107	100	100
56448	258564107	56448 - 258564107	21.74224255	11.46445096
258564374	258564374	258564374 - 258564374	100	100
258599179	258564885	258599179 - 258564885	42.49700482	25.37968598
258564885	258564885	258564885 - 258564885	100	100
258564885	258564887	258564885 - 258564887	22.01089351	13.59320243
258564887	258564887	258564887 - 258564887	100	100
56406	258564887	56406 - 258564887	25.47407614	100
52910	258564887	52910 - 258564887	26.79091652	69.63919803
258565388	258565388	258565388 - 258565388	100	100
258568982	258568982	258568982 - 258568982	100	100
51272	258568982	51272 - 258568982	80.8193461	32.72566864
258568993	258568982	258568993 - 258568982	99.50076684	27.77491546
51272	258568993	51272 - 258568993	27.51091371	39.9072584
258568993	258568993	258568993 - 258568993	100	100
258568982	258568993	258568982 - 258568993	27.77491546	99.50076684
258571532	258571532	258571532 - 258571532	100	100

Figure 31 - Cell Overlapp table

To have an easy visualization about two cell intersections in groups of ABC cells I defined an overlapping matrix. In this matrix is possible to see on his two column diagonal the trajectory. The intensity color is related with the overlapping percentage. Any other intersection which is not on the two column diagonal is not important for this analysis because is not happening at the same time.

The first approach is for the entire user data frame, it means all the weeks together. However, the second step in this process is to split the analyses by days of the week.

The Cells Overlapping Matrix and the Trajectory Episode Matrix are the key results of this thesis mainly because it is a new approach to the combination of trajectory events. It also simplifies the display of the user trajectory movement and placements.

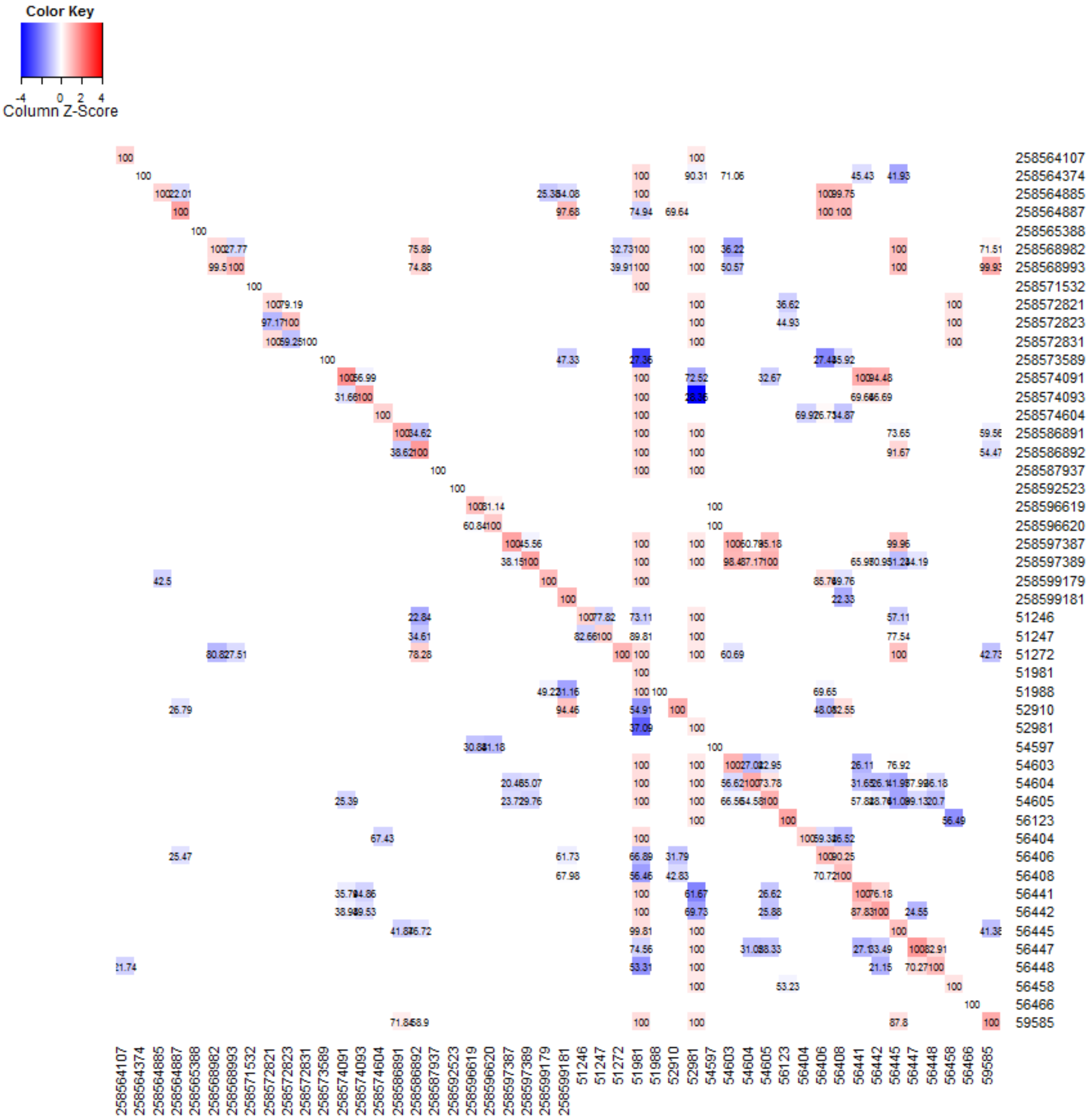
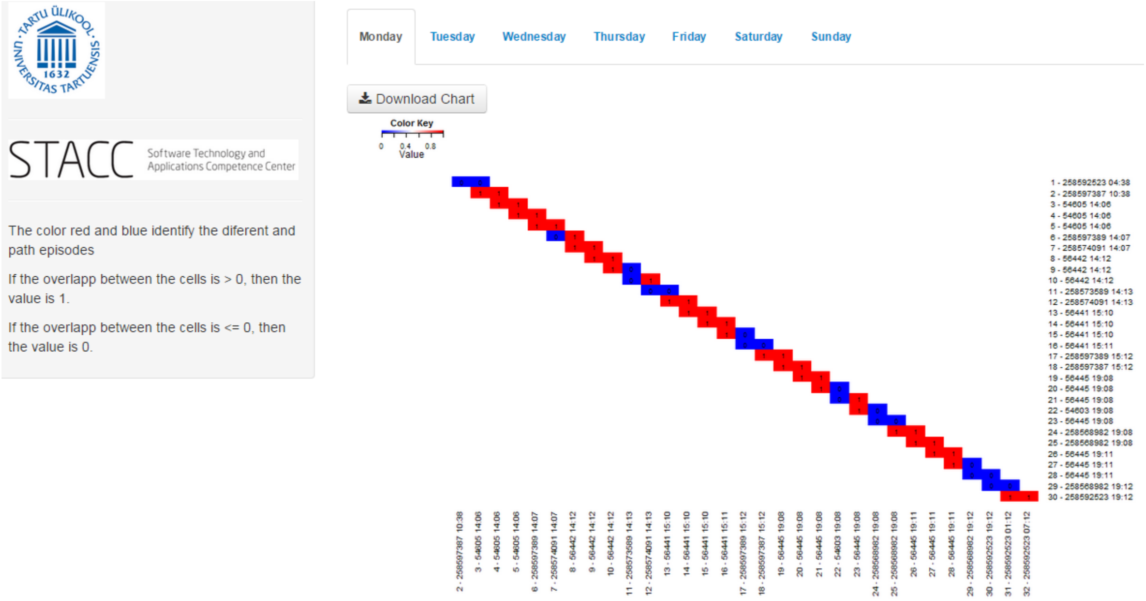


Figure 32 - Overlapping Matrix + 20% intersection

4.8 Trajectory episodes

The objective is to identify clearly the different episodes in the user trajectory, for this reason the transition from one episode to another is represented by the colors red and blue, and also by the numbers “0” and “1”. And it is also available to visualize the overlapping matrix by day.

From this chart (see figure 33) we can extract lot information related with the trajectory and movement/ placement candidates. The episodes in red are candidate to be a placement because are the intersection of several trajectory cells and because initial timestamp of the episode is different than the final timestamp of the episode.



The color red and blue identify the diferent and path episodes

If the overlapp between the cells is > 0, then the value is 1.

If the overlapp between the cells is <= 0, then the value is 0.

Figure 33 - Overlapping Matrix per Day

The overlapping matrix is the input a series of spatio-temporal analysis in order to discover placement and movement patterns.

4.9 Path map by hour with polygons

The best way to have an interpretation of the overlapping matrix and his impact on the analyses is to show the results in a map. The Path map by hour with event polygons has even more information than any other previous report. It is possible to identify the events and the trajectory path of the events.

Comparing the map with the trajectory events and the map with the trajectory episodes helps to understand the progress in the analysis. With this methodology is possible to discover meaningful locations.

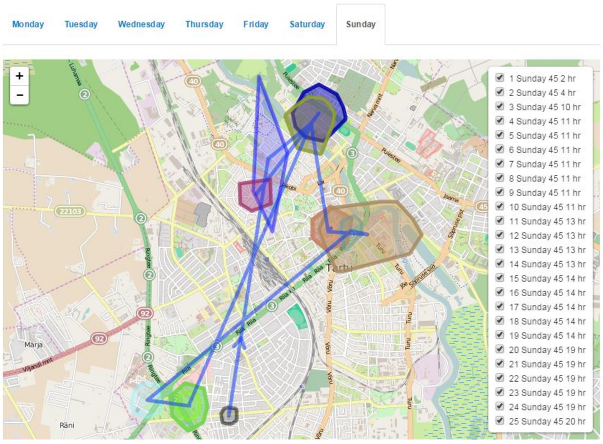


Figure 34 - Trajectory events map by hour

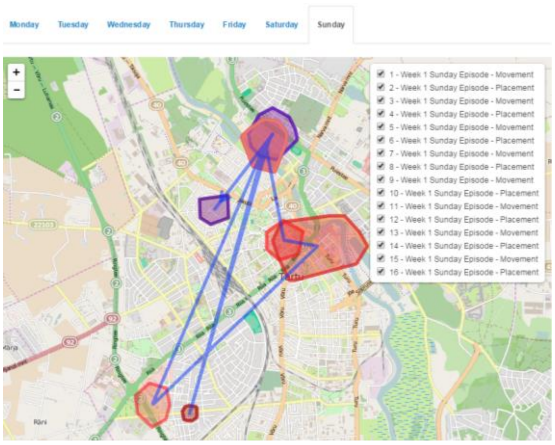

Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
+
-
☑ 1 - Week 1 Sunday Episode - Movement
☑ 2 - Week 1 Sunday Episode - Placement
☑ 3 - Week 1 Sunday Episode - Movement
☑ 4 - Week 1 Sunday Episode - Placement
☑ 5 - Week 1 Sunday Episode - Movement
☑ 6 - Week 1 Sunday Episode - Placement
☑ 7 - Week 1 Sunday Episode - Movement
☑ 8 - Week 1 Sunday Episode - Placement
☑ 9 - Week 1 Sunday Episode - Movement
☑ 10 - Week 1 Sunday Episode - Placement
☑ 11 - Week 1 Sunday Episode - Movement
☑ 12 - Week 1 Sunday Episode - Placement
☑ 13 - Week 1 Sunday Episode - Movement
☑ 14 - Week 1 Sunday Episode - Placement
☑ 15 - Week 1 Sunday Episode - Movement
☑ 16 - Week 1 Sunday Episode - Placement
Räni

Figure 35 - Trajectory episode map by hour

4.10 Cluster Analyses

The last analysis, more accurate and sophisticated, is the classification with density clustering. The DBSCAN clustering analysis fits better to work with spatial data and it help in the aim to search meaningful locations.

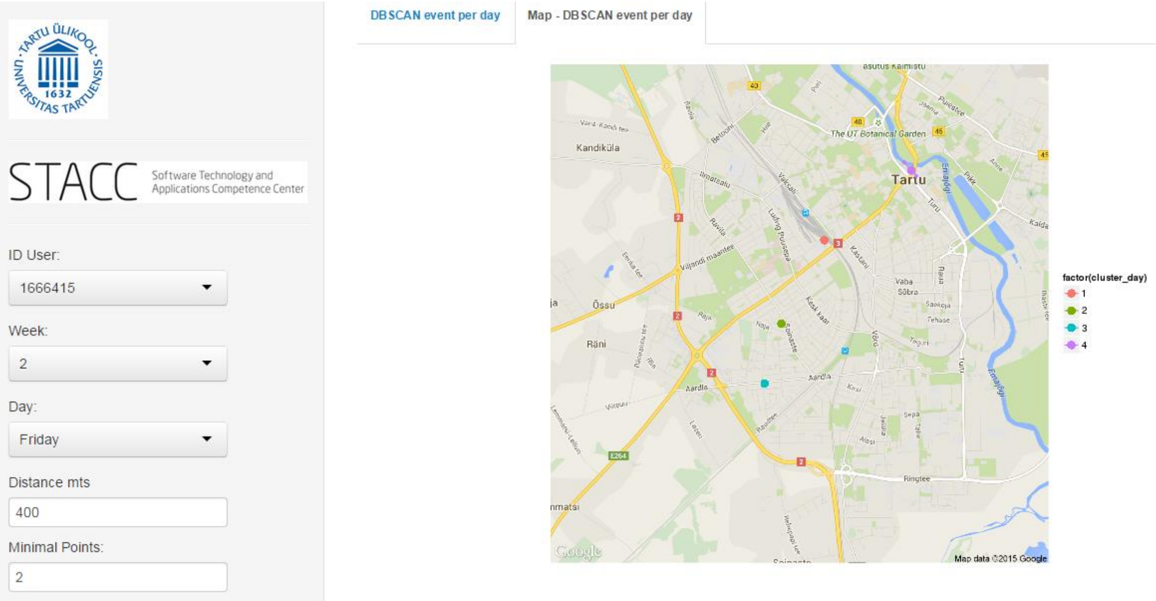
As I mentioned above, in the methodological DBSCAN clustering statement, this analysis has three steps.

4.10.1 DBSCAN Clustering event analyses by day

The first DBSCAN clustering for every user is by day. This clustering is helpful, through the clustering centroid to found meaning full locations. In the output chart and map we can see the clustering centers and the points, which belong to the center. In this representation the outliers scatter are not shown.

The figure 29 and 30 has as input values “Distance = 400 mts” and “Minimal Points=2”.

This first analysis is just for events; however there is a pattern in the scatter distribution.



ID User: 1666415

Week: 2

Day: Friday

Distance mts: 400

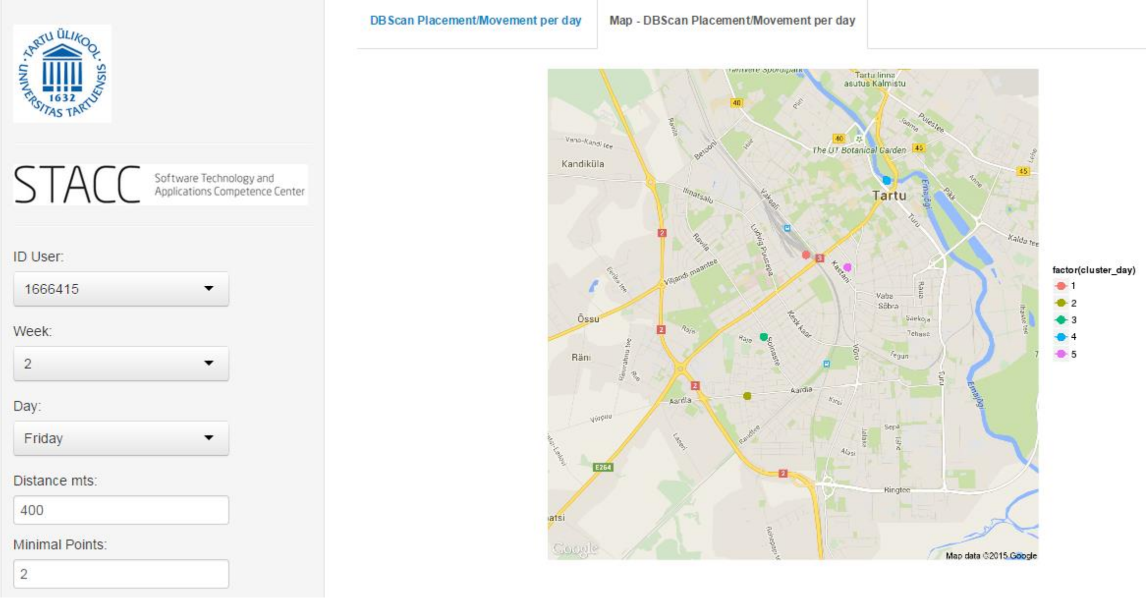
Minimal Points: 2

Figure 36 - Map - DBSCAN Clustering event analyses by day

4.10.2 DBSCAN Episode Movement/Placement clustering analyses by day

The aggrupation of the events per episode should help in the analyses because there is less dispersion on the scatters. However there are fewer points to analyze per day.

From the first two analyses is not possible to see a real improvement in classification meaningful locations. The main reason is that to perform the analyses per each day and week there is not enough data to define patterns.


STACC Software Technology and Applications Competence Center
DBScan Placement/Movement per day
Map - DBScan Placement/Movement per day
Tartu
factor(cluster_day)
1
2
3
4
5
Map data ©2015 Google

ID User: 1666415

Week: 2

Day: Friday

Distance mts: 400

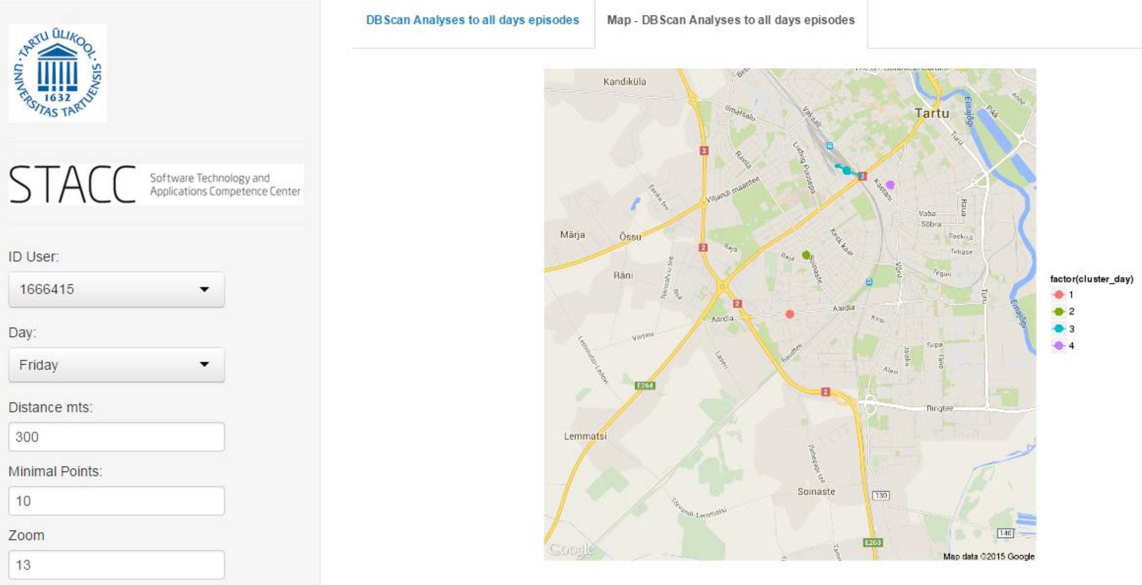
Minimal Points: 2

Figure 37 - Map - DBSCAN Episode clustering analyses by day

4.10.3 DBSCAN Clustering analyses to all days episodes

To solve the problem related with not enough data to define placement and movement patterns, applied the analyses to all the same days together. For instance all the Mondays for every week set together.

These are the results of the analyses. To have an accurate output in the example below I used “Distance = 300 mts” and “Minimal point = 10”. This change in the input against the previous analyses in mainly because big number of points in a small urban area which means the same location.


STACC Software Technology and Applications Competence Center
DBScan Analyses to all days episodes
Map - DBScan Analyses to all days episodes
Tartu
factor(cluster_day)
1
2
3
4
Map data ©2015 Google

ID User: 1666415

Day: Friday

Distance mts: 300

Minimal Points: 10

Zoom 13

Figure 38 - Map - DBSCAN Clustering to all days episodes

The clustering result is related with distribution and number of episodes of the user data. The DBSCAN clustering is sensitive to these variations and is possible to get optimal results after the tuning with ”Distance” and “Minimal Points” inputs.

The visual analytic tool has only implemented the DBSCAN clustering. The spatiotemporal analysis could be improved with the implementation of more classification algorithms. The ST-DBCAN has the geographic and temporal distances. The VDBSCAN can work better with not homogenous scatter densities.

Future works should be oriented on the implementation and validity of classification algorithms for the episode analysis by day.

5. CONCLUSIONS AND FUTURE WORK

The motivation behind the thesis topic lies in providing the knowledge and tools needed to extract meaningful information out of the huge amounts of data from mobile operators.

We worked with mobile location data, which is a combination of CDR and Cell Plan data. Then we performed analysis to find meaningful locations and identify outliers or incorrect data. We created the episode trajectory matrix and classified the episode distribution in order to recognize placement and movement patterns.

The cell frequency analysis was used to discover meaningful locations, to calculate the most visited cells and to display the meaningful event trajectory user on a map. The concept of most visited cells helps to understand which events are candidates to be part of a placement or movement episodes.

This approach shows the patterns of displacement day by day. The displacement can be categorized to stay action or a movement action. Additionally, we can also get information about transition distance.

When the user time-distance is high and cannot represent a real movement, the events is eliminated from the trajectory. The centroid analysis performs this task.

The most important contribution of this thesis is the “overlapping trajectory matrix” and its implementation in the “trajectory episodes matrix”. The “overlapping trajectory matrix” shows the intersection between cells in the trajectory.

The intersection of consecutive trajectory polygons is an episode. The “trajectory episodes matrix” clearly identifies the different episodes in the user trajectory; the transition from one episode to another is highlighted. From this matrix it is possible to determine placement and movement episodes. The overlapping matrix is used to discover placement and movement episode patterns.

The overlapping matrix results are shown in a spatial representation in order to identify episodes trajectories. By comparing the map with the trajectory events and the map with the trajectory episodes it helps to understand the progress of the analysis and find meaningful locations.

The ping-pong handover analysis main result shows that if the handover happens inside the episode, the phenomenon has no effect on the episode.

The cluster classification helps improve to find meaningful locations of placement and movement episodes. The clustering results are related with distribution and number of episodes. The DBSCAN clustering is sensitive to these variations and it is possible to get optimal results after the experimenting with the ”Distance” and “Minimal Points” inputs.

To build upon the work done it is possible to introduce enhancements to the polygon coverage area modelling to have accurate results on the placement and movement episodes detection with the “episodes trajectory matrix”.

The visual analytic tool has only implemented the DBSCAN clustering. The spatiotemporal analysis could add more classification algorithms. The ST-DBCAN has the geographic and temporal distances. The VDBSCAN can work better with heterogeneous scatter densities.

Future works should be oriented towards the implementation and validity of classification algorithms for the episode analysis day by day.

The episode trajectory matrix could be used as an input to define semantics of movement and placement. This data has the dimensions and features needed to achieve this aim. Moreover, we have the intention to continue this work to bring up to the level of a scientific publication.

REFERENCES

[1] Candia, J., González, M., Wang, P., Schoenharl, T., Madey, G., Barabasi, A-L.2008. “Uncovering individual and collective human dynamics from mobile phone records.” Journal of Physics A: Mathematical and Theoretical, Vol. 41 N. 224015.

[2] González, M., Hidalgo, C., Barabási, A. 2008. “Understanding individual human mobility patterns.” Nature, Vol. 453: 779-782.

[3] Järv, O., Ahas, Saluveer, E., Derudder, B., Witlox, F. 2012. Mobile Phones in a Traffic Flow: A Geographical Perspective to Evening Rush Hour Traffic Analysis Using Call Detail Records, PLoS ONE 7(11), http://dx.plos.org/10.1371/journal.pone.0049171

[4] Vivek K. Singh, Laura Freeman, Bruno Lepri, and Alex Pentland (2013) “Predicting Spending Behavior Using Socio-mobile Features” SocialCom, page 174-179. IEEE

[5] Nicholas D. Lane, Emiliano Miluzzo, Hong Lu, Daniel Peebles, Tanzeem Choudhury, and Andrew T. Campbell, Dartmouth College, S. 2010. “A Survey of Mobile Phone Sensing” Communications Magazine, IEEE Volume:48 Issue:9

[6] Jiang, S., Fiore, G., Yang, Y., Ferreira, J., Frazzoli,E., González, M. 2013. “A review of urban computing for mobile phone traces: current methods, challenges and opportunities.” Proceedings of the ACM SIGKDD International Workshop on Urban Computing.

[7] Lee, A., Chen, Y-A., Ip, W-C. 2009. “Mining frequent trajectories patterns in spatial-temporal databases.” Information Sciences, Vol. 179: 2218-2231.

[8] Shan Jiang, Joseph Ferreira, Jr., Marta C. Gonzalez - 2012. “Discovering Urban Spatial-temporal Structure from Human Activity Patterns”. Proceedings of the ACM SIGKDD International Workshop on Urban Computing

[9] Ericson, Definition of Nominal Cell Plan http://learning.ericsson.net/icp/content/content/lesson2/icp213.html -

[10] Noticewala Maitry, Dinesh Vaghela, 2014 - “Survey on Different Density Based Algorithms on Spatial Dataset”

[11] Ran Nathan, Wayne M. Getzb, Eloy Revillac, Marcel Holyoakd, Ronen Kadmona, David Saltze, and Peter E. Smousef, 2008 – “A movement ecology paradigm for unifying organismal movement research” Proceedings of the National Academy of Sciences

[12] Samiul Hasan, Christian M. Schneider, Satish V. Ukkusuri, and Marta C. Gonzalez, 2012 - “Spatiotemporal patterns of urban human mobility” - Journal of Statistical Physics – Pages 304 - 318

[13] Couceiro, M., D. Suarez, D. Manzano, and L. Lafuente. "Data Stream Processing on Real-Time Mobile Advertisement: Ericsson Research Approach." In 2011 12th

IEEE International Conference on Mobile Data Management (MDM), 1:313320, 2011. doi: 10.1109 / MDM. 2011.42.

[14] Rein Ahas , Siiri Silm , Olle Jrv , Erki Saluveer and Margus Tiru (2010), "Using Mobile Positioning Data to Model Locations Meaningful to Users of Mobile Phones," Journal of Urban Technology, 17:1, 3-27, DOI: 10.1080/10630731003597306

[15] Giannotti, F. and Pedreschi, D. – 2008. Mobility, Data Mining and Privacy. Springer-Verlag, Berlin Heidelberg.

[16] R. Ahas, J. Laineste, A. Aasa, and U. Mark, "The Spatial Accuracy of Mobile Positioning: Some experiences with Geographical Studies in Estonia," in Location Based Services and Tele Cartography, P. D. G. Gartner, P. D. W. Cartwright, and P. D. M. P. Peterson, Eds. Springer Berlin Heidelberg, 2007, pp. 445460.

[17] R. Ahas, S. Silm, E. Saluveer, and O. Jarv, "Modelling Home and Work Locations of Populations Using Passive Mobile Positioning Data," in Location Based Services and Tele Cartography II, G. Gartner and K. Rehrl, Eds. Springer Berlin Heidelberg, 2009, pp. 301315

[18] Fosca Giannotti · Mirco Nanni · Dino Pedreschi · Fabio Pinelli · Chiara Renso · Salvatore Rinzivillo · Roberto Trasarti, “Unveiling the complexity of human mobility by querying and mining massive trajectory data”, Received: 29 September 2010 / Revised: 1 June 2011 / Accepted: 29 June 2011 / Published online: 30 July 2011 – Springer - Verlag 2011

[19] Chih-Chieh Hung, Wen-Chih Peng, “A regression-based approach for mining user movement patterns from random sample data” Department of Computer Science, National Chiao Tung University, Taiwan, ROC - 2010

[20] Yihong Yuan, Martin Raubal, “Spatio-temporal knowledge discovery from georeferenced mobile phone data”, Department of Geography, University of California, Santa Barbara, USA, 93106, September 14th, 2010

[21] Zoltán Fehér$_1$, , András Veres$_2$, , Zalán Heszberger, “Ping-pong Reduction using Sub cell Movement Detection”, $_1$HSNLab, Dept. of Telecommunications and Media Informatics Budapest University of Technology and Economics Budapest, Hungary zoltan.feher, $_2$Ericsson Research Ericsson Hungary LTD. Budapest, Hungary

[22] Günther Sagl$_1$, Martin Loidl$_2$ and Euro Beinat$_{2,}$ “A Visual Analytics Approach for Extracting Spatio-Temporal Urban Mobility Information from Mobile Network Traffic”. $_1$Doctoral College Geographic Information Science, University of Salzburg, Schillerstraße 30, A-5020 Salzburg, Austria. $_2$Department of Geoinformatics (Z_GIS), University of Salzburg, Hellbrunnerstraße 34, A-5020 Salzburg, Austria; Published: 2 November 2012

[23] Magdalena Nohrborg – Self-Organization Networks – www.3gpp.org

[24] 3GPP TS 32.500 V12.1.0 (2014-12) Technical Specification - 3rd Generation Partnership Project, Technical Specification Group Services and System Aspects, Telecommunication Management, Self-Organizing Networks (SON), Concepts and requirements (Release 12)

[25] Yiannis Kamarianakis - Poulicos Prastacos, Department of Economics, University of Crete, Rethymnon, Greece, and Regional Analysis Division, Institute of Applied and Computational Mathematics, Foundation for Research and Technology-Hellas Vasilika Vouton, Heraklion-Crete, Greece - “Spatial Time-Series Modeling: A review of the proposed methodologies” – 2006

[26] Ilya Boyandin1, Enrico Bertini2, Peter Bak3 andDenis Lalanne1 – “Flowstrates: An Approach for Visual Exploration of Temporal Origin-Destination Data” - Published online: 28 JUN 2011 DOI: 10.1111/j.1467-8659.2011.01946.x

[27] A. Neubacher, 2009 - “Method for preventing ping-pong handover in mobile radio networks” U.S. Patent WO2009021711A1, issued Februar 19, 2009.

[28] [15] O. Jukiü, Nov. 2012 - “The use of call detail records and data mart dimensioning for telecommunication companies” - Telecommunications Forum (TELFOR), 2012 20th - 20-22

[29] Michael Hahsler, 2014 - Southern Methodist University, Matthew Bolaños - Southern Methodist University, John Forrest – Microsoft, “Introduction to stream: An extensible Framework for Data Stream Clustering Research with R”

[30] Miller, 2009 – “Geographic Data Mining and Knowledge Discovery” - J. P. Wilson and A. S. Fotheringham (eds.) Handbook of Geographic Information Science, in press

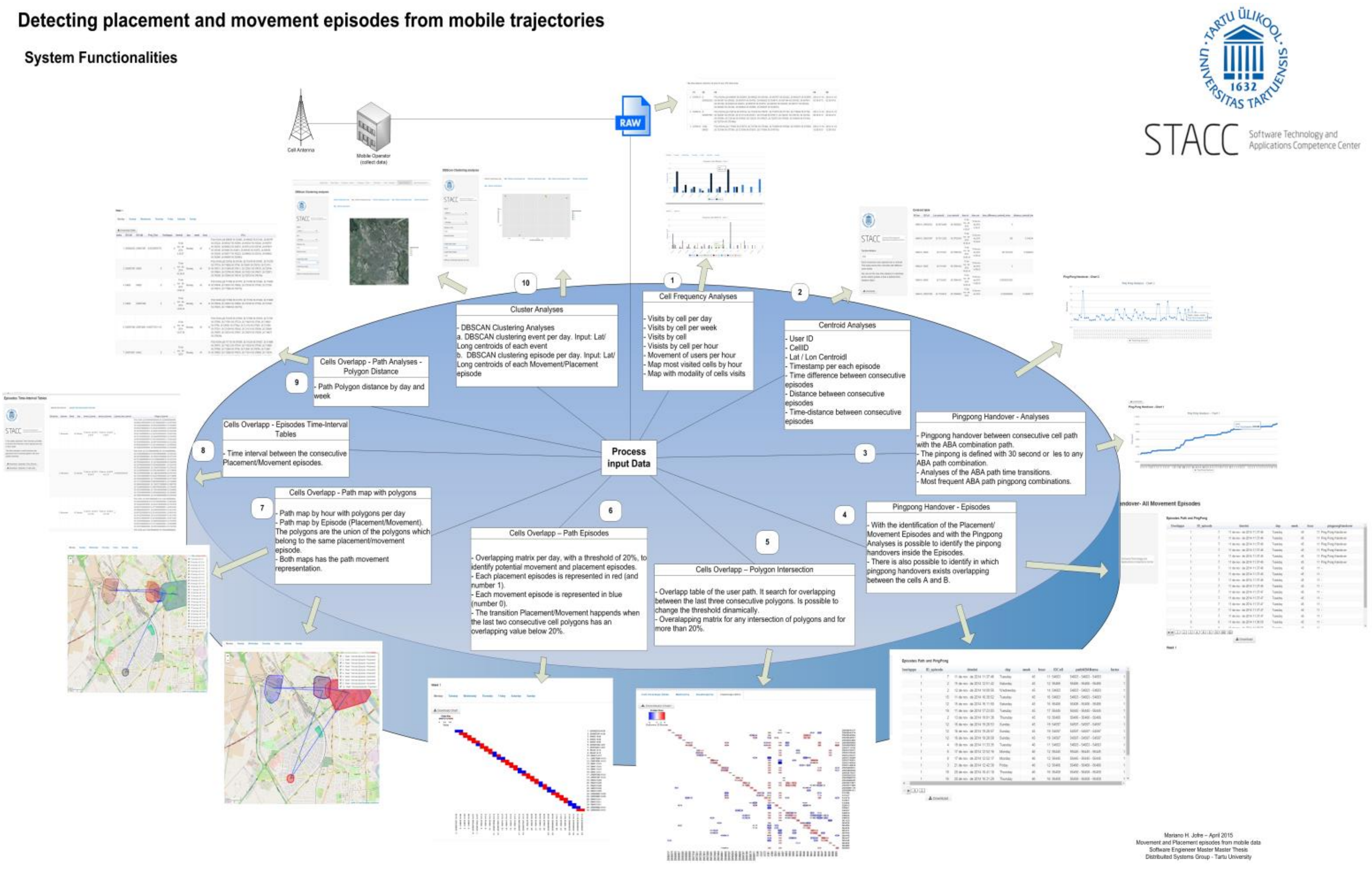
[31] ”Introduction to R.” R-project. 4. April. 2014. http://www.r-project.org/

[32] Introduction to Data Mining – Chapter 8 - Cluster Analysis: Basic Conceptsand Algorithms - Pang-Ning Tan, Michigan State University, Michael Steinbach, University of Minnesota, Vipin Kumar, University of Minnesota - ISBN-10: 0321321367 • ISBN-13: 9780321321367 - 2006 • Pearson • Cloth, 769 – pp Published 05/02/2005

[33] Carlos Roberto Valêncio, Guilherme Priólli Daniel, Camila Alves De Medeiros, Adriano Mauro Cansian, Luiz Carlos Baida, Fernando Ferrari – 2013 – “VDBSCAN+: Performance Optimization Based on GPU Parallelism” - Published in: PDCAT '13 Proceedings of the 2013 International Conference on Parallel and Distributed Computing, Applications and Technologies - Pages 23-28 - IEEE Computer Society Washington, DC, USA ©2013 - ISBN: 978-1-4799-2419-6

Appendix

I. Annex 1: System Functionalities



Detecting placement and movement episodes from mobile trajectories

System Functionalities

Cell Antenna
Mobile Operator (collect data)
RAW
Process input Data
1
2
3
4
5
6
7
8
9
10
andover- All Movement Episodes

1 Cell Frequency Analyses

- Visits by cell per day
- Visits by cell per week
- Visits by cell
- Visists by cell per hour
- Movement of users per hour
- Map most visited cells by hour
- Map with modality of cells visits

2 Centroid Analyses

- User ID
- CellID
- Lat / Lon Centroidl
- Timestamp per each episode
- Time difference between consecutive episodes
- Distance between consecutive episodes
- Time-distance between consecutive episodes

3 Pingpong Handover - Analyses

- Pingpong handover between consecutive cell path with the ABA combination path.
- The pinpong is defined with 30 second or les to any ABA path combination.
- Analyses of the ABA path time transitions.
- Most frequent ABA path pingpong combinations.

4 Pingpong Handover - Episodes

- With the identification of the Placement/ Movement Episodes and with the Pingpong Analyses is possible to identify the pinpong handovers inside the Episodes.
- There is also possible to identify in which pingpong handovers exists overlapping between the cells A and B.

5 Cells Overlapp – Polygon Intersection

- Overlapp table of the user path. It search for overlapping between the last three consecutive polygons. Is possible to change the threshold dinamically.
- Overalapping matrix for any intersection of polygons and for more than 20%.

6 Cells Overlapp – Path Episodes

- Overlapping matrix per day, with a threshold of 20%, to identify potential movement and placement episodes.
- Each placement episodes is represented in red (and number 1).
- Each movement episode is represented in blue (number 0).
- The transition Placement/Movement happends when the last two consecutive cell polygons has an overlapping value below 20%.

7 Cells Overlapp - Path map with polygons

- Path map by hour with polygons per day
- Path map by Episode (Placement/Movement). The polygons are the union of the polygons which belong to the same placement/movement episode.
- Both maps has the path movement representation.

8 Cells Overlapp - Episodes Time-Interval Tables

- Time interval between the consecutive Placement/Movement episodes.

9 Cells Overlapp - Path Analyses - Polygon Distance

- Path Polygon distance by day and week

10 Cluster Analyses

- DBSCAN Clustering Analyses
a. DBSCAN clustering event per day. Input: Lat/ Long centroids of each event
b. DBSCAN clustering episode per day. Input: Lat/ Long centroids of each Movement/Placement episode

Mariano H. Jofre – April 2015
Movement and Placement episodes from mobile data
Software Engineer Master Master Thesis
Distributed Systems Group - Tartu University

Figure 39 - System Functionalities

II. License

Non-exclusive licence to reproduce thesis and make thesis public

I, **Mariano Hedberto Jofre** (date of birth: 08/07/1975),

1. Here with 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,

of my thesis

Placement and Movement Episodes Detection using Mobile Trajectories Data,

supervised by Dr. Amnir Hadachi & Elis Kõivumägi,

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, **26.05.2015**