.

UNIVERSITY OF TARTU

Faculty of Science and Technology

Institute of Computer Science

Computer Science Curriculum

Dmitri Timašjov

Human Mobility Mining Using Spatio-Temporal Data

Master's Thesis (30 ECTS)

Supervisor: Amnir Hadachi, PhD

Tallinn 2016

Human Mobility Mining Using Spatio-Temporal Data

Abstract

Geospatial technologies have become an integral part of our lives. With technological progress and rapid increase of geospatial information and inexpensive positioning technologies, more space-related data is becoming available at any time. Data is collected using multiple sources such as GPS and mobile computer logs, wireless communication devices, location-aware services and other positioning systems. This gives scientists the opportunity to create new innovative platforms for spatio-temporal data analysis and improve methods for mining and visualization for decision support. In order to provide a good decision support systems, it is vital to understand people's movement, mobility behaviour and be able to discover hidden patterns and associations in their daily activities. The aim of this thesis is to analyze and discuss spatial data mining techniques by answering questions like what kinds of patterns can be extracted from spatio-temporal data or which methods are best for predicting human mobility behavior. In this work, we verify existing methodologies and theories about spatio-temporal data mining and propose a sequence of algorithms to achieve good human mobility prediction. We evaluate the results and propose a methodology for adaptive data mining of human mobility behavior.

Keywords: *Spatio-temporal data mining, GPS data, Location prediction, Human mobility, Location analytics*

CERCS: P170

Inimeste aegruumilise käitumise ja mobiilsuse uuring

Resümee

Georuumilised tehnoloogiad on lahutamatu osa meie elust: tehnoloogilise arengu ja positsioneerimise seadmete levikuga on toimunud kiire kasv kättesaadavate georuumiliste andmete mahus. Andmed kogutakse erinevate allikate kaudu nagu GPS ja mobiilseadmete logid, traadita sidevahendid ja asukohapõhised teenused ning teised positsioneerimise süsteemid. Liikumise kohta on võimalik infot koguda suures mõõtkavas ja hea täpsusega - see annab uurijatele võimaluse luua uusi ja innovaatilisi platvorme ja teenuseid georuumilise info analüüsimiseks ning parandada andmete kaevandamise ja visualiseerimise tehnikaid. Selleks, et luua hea nõustussüsteem, on väga oluline saada aru inimeste liikumisharjumustest ja käitumisest ning leida igapäevaste tegevuste varjatud mustrid.

Magistritöö eesmärgiks on analüüsida andmekaevandamise meetodeid, uurides, millised mustrid võivad olla liikumise trajektoorides või milliste algoritmidega saab ennustada inimeste käitumist. Töös kontrollitakse nii olemasolevaid metoodikad ja teooriad ruumilise andmekaevandamise valdkonnas kui ka pakutakse arendatud algoritmide jada inimeste liikumise ennustamiseks. Me hindame ja võrdleme tulemusi omavahel ning töötame välja metoodika inimeste liikumiskäitumise adaptiivseks andmekaevandamiseks.

Märksõnad: *aegruumiline analüüs, GPS andmed, asukoha ennustamine, inimeste mobiilsus, aegruumilised liikumised*

CERCS: P170

Acknowledgements

First, I wish to express my sincere gratitude to my supervisor Dr. Amnir Hadachi for guidance and faith. For a period of two years he has been a continuous source of knowledge and wisdom.

I also would like to express my sincere appreciation to my mother Irina for constant reminders as well as my wife Olga for her unending support and help. I wouldn't have finish this thesis without you!

Table of contents

Abbreviations and Acronyms

This section clarifies some terms used in the paper.

DBSCAN

Density-based spatial clustering of applications with noise

GIS

Geographic Information System

GPS

Global Positioning System

LHS

Left-hand side of an equation

OSM

OpenStreetMap

POI

Point of Interest

RHS

Right-hand side of an equation

1. Introduction

Geospatial technologies affect almost every aspect of life. A number of modern geopositioning technologies is progressing rapidly and more geospatial information is becoming available. New advancements and developments in the field continue to take place. Nowadays mobility data in the form of spatially referenced time series is collected on a very large scale and with a good precision. Data is collected from different sources: positioning systems, network traffic controllers, geo-tagged photos and geo-referenced datasets, mobile computer logs, location-aware and wireless communication devices and much more. The number of such sources and size of their datasets are growing rapidly [30], therefore, real-time location information are commonly part of our everyday lives. Widespread availability of low cost GPS devices also did not play the least role in this expansion.

This contributes scientists, geoinformatic and telecommunication specialists to create new innovative platforms for spatio-temporal data research and analysis. Such platforms are designed to improve methods for mining, visualization of moving objects and discovering of hidden patterns. Development of this research area depends on different social, commercial and technological aspects. Spatio-temporal data can be used for many various purposes and in many different scientific and applied sciences as well as in designing and management of cities to make them more sustainable. Associations in spatio-temporal data can greatly help with understanding and predicting our environment, for example customers mobility, weather forecasting, mobile marketing and targeted advertising, personalization of contents and services or even monitoring epidemics and predicting spread of the disease. Those platforms are aimed primarily for making better and faster decisions.

In order to make a decision, we must be able to get raw spatio-temporal data, process this data and extract useful information from it. Due to the fact that the amount of available space-related data is growing blazingly fast, it becomes challenging to distinguish useful information, because it requires new and efficient computational analysis methods, which must be able to handle large amounts of data with ease as well as new representation methods and ways of storing the data. Such methods must be able to use all available information gathered over the years as well as personalised information and data from every single user correspondingly. Therefore, sustainable data mining techniques must exist in order to provide high quality results.

1.1 Problem statement

If we think about our daily movement, it is obvious that our location points do not spread uniformly, but they tend to gather in few limited areas, where we stay for a longer period of time [41]. Those geographic areas carry some semantic meaning and are called significant places. An example of a significant place can be a workplace, friend's house, shopping center, office building, bar and supermarket, restaurant or any other place that capture user's interest. Multiple statistical studies have shown that most people have regular daily routines of traveling [19] and visiting the same locations. Given that we have observed people for a long enough time and collected sufficient amount of observations, mining of those areas can greatly assist in extracting useful information that can be used for prediction of the next possible location. This, in turn, provides a new ways to understand human mobility and activity patterns, opens new chances for location-based services as well as introduce new issues in performing data mining and analysis in today's pervasive computing environments.

Although a lot of research have been done in the field of location prediction, we found a very few studies on the topic of the combination of different mobility prediction methods capturing various aspects of human movements, such as semantic or temporal information.

Towards this end, thesis focuses on the following research questions:

1. Can combination of different location prediction methods result in achieving better prediction success rate than utilization of only one location prediction method?
2. If so, which circumstances contributed to the increase in prediction success rate?

1.2 Contributions

Methodology of this work is based on two major principles of data analysis: understanding and predicting. This thesis targets to achieve following objectives:

- Identify the main reasons that drive people to change their location.
- Investigate methods and techniques for spatio-temporal data mining.
- Analyze people's movements and detect and classify geographic areas that carry some semantic meaning and capture their interest.
- Derive a framework for predicting people's next geographic location by capturing the sequential relations between places visited in a given time period by all individuals [29]. Based on the derived statistical patterns we are focusing on predicting future locations to be visited. Specifically, we propose a hybrid method based on [37, 40, 41].

1.3 Road map

The rest of the thesis dissertation is organized as follows.

Chapter 2: Presents an overview of related literature and possibilities of geographical data mining and talks about the constituents used in the thesis. Chapter introduces basic definitions and their properties. Also, semantics in geographic data and importance of adding it were discussed.

Chapter 3: Describes the origin and specification of the data as well as software technologies used in this thesis.

Chapter 4: First and foremost, data preprocessing techniques were discussed. Next, we propose a model for analysis and prediction of human mobility. Each integral component of the model is overviewed and backed by examples.

Chapter 5: Presents the results achieved when applying proposed model to real life geographic data. Also, the pros and cons of the model together with problems encountered during the implementation were covered in details.

Chapter 6: Concludes the results as well as presents future research perspectives.

2. Background and Related Work

Before we proceed with a framework for prediction of human mobility, it is important to have an understanding of the existing widely used models and theories. Knowledge of basic conceptions will help towards discovering additional features of mobility analysis and prediction that may be of relevance. This chapter provides an overview of the relevant literature, definitions and their properties used in the paper while developing the framework.

2.1 State of the art

Analysis of spatial data and human mobility have been a hot topic for a long time and was addressed and studied in many papers. There is no uniform opinion among scientists about this topic - a certain group of scientist believes that movements of people follow some random regulation [20, 31], whereas the other ones believe that human trajectories follow common patterns and show a high degree of temporal regularity [16, 24].

There are a huge amount of methods for analyzing human mobility and location prediction, but in general they can be classified into three major categories:

1. Data-mining techniques
2. Space-state models
3. Semantic analysis techniques.

First and most widely used method for analyzing human mobility is by applying data mining techniques for exploration of hidden patterns and mining of association rules. In a nutshell, this includes analysis of previous occurrences by clustering, aggregation and extracting patterns from time series data. This method also heavily uses the notion of spatial analysis and heuristic algorithms to make a decision - this means finding longest common subsequences, analyzing route dynamics and similarity indexes or any other algorithms for distance analysis. Such approach was used, for instance, in [1,6,41]. A diverse variety of different models were created, for example authors of [23] proposed a model called “M-Model” for mining and querying of complex trajectory data by combining common behavior of groups of objects.

The second type of approaches, which is space-state models, use sequence models and probabilistic automation for mining location history [19, 25, 29, 36]. One of the most popular and cited representatives of this class is Hidden Markov Model (HMM). In HMM the system is considered to be a Markov process with hidden states and the main goal is to analyze the data that is not immediately observable by training the model: location history in our case. Algorithm has lots of advantages, like being able to capture dependencies between measurements and representing variance through probability distribution, however, as with all machine learning techniques, final result is not fixed and depends on amount of training and visible states. Thus, this results in a very different prediction accuracy: [29] report an accuracy of 13.85% when using HMM, while authors of [5] in their work get an accuracy of 45% with HMM. Other models can also be used for mobility prediction: conditional random fields, for instance [8].

Third approach is semantic analysis, which deals mainly with template matching and considers semantics as a main criteria when analyzing movement history. It analyzes location history and produces a so-called ”semantic space”, which consists of semantic links, that play a key role in decision making. It analyzes social aspects of human mobility as well as points of interest (POIs). For example [38] used this approach to discover regions of different function in a city. Usually those methods do not include any perception of spatial analysis, thought, can use temporal data for creating necessary semantic links. However, complex semantic analysis processes are not yet fully automated and often need help of people as described in [3]. This happens due to the fact that often data simply lack necessary semantic links, which cannot be interpreted by computers, but can be easily read, understood, and if necessary, restored by humans.

2.2 GPS data

First step is to record all needed information to get digital track of people's movements. The most common way of getting positioning information is using GPS (Global Positioning System). GPS is space-based radio-navigation system developed by the US Department of Defense that uses the notion of satellites to provide location and time information. The idea is based on the fact that it is possible to determine the location on the Earth by knowing the exact time, speed and location of the satellite. Nowadays there are 31 satellites used for positioning services circulating at 14000 km/hr about 20000 km above the Earth's surface. Microwave radio signals travelling at the speed of light from at least three satellites are used by the receiver's built-in computer to calculate its position, altitude and velocity. Determination of the exact location is measured by the reception timings from the navigation satellites to the receiver antennas.

GPS navigation is freely accessible for using with any GPS receiver, providing GPS data. Stored GPS recordings are also called GPS logs.

Definition 1. ***GPS log:*** *a collection of GPS points* $P = \{p1, p2, ..., p_n\}$*, where each point* $p \in P$ *contains latitude* $(p.Lat)$*, longitude* $(p.Lon)$*, timestamp* $(p.T)$*, altitude* $(p.A)$*, velocity* $(p.V)$ *and other information.*

GPS positioning has its own advantages and disadvantages. Probably the most attractive feature of the GPS is that it covers 100% of the planet and can operate in almost all weather conditions and on any surface. Also, GPS greatly facilitates navigation as it can report the direction and the angle of the movement. GPS receivers costs very low and are easily integratable into computers and mobile devices when comparing with other navigation systems. Nevertheless, GPS is not infallible and might be not very accurate in some cases. The main problem comes from inaccurate time-keeping by the receiver's device clock - the time when receiver's computer got the signal and the time used by the whole global positioning system for synchronization might be slightly different. Those tiny discrepancies may lead to the fact that calculated distance can drift, which means that accuracy of location positioning will not be fully accurate. Furthermore, the quality of the GPS signal depends on the landscape where it is received. Radio signals may easily be distorted as they are unable to pass through solid structures like tall buildings, underground, deep forest or underwater.

GPS greatly contributed to the creation of location-based social networks and services such as FourSquare[1], Rally Up[2] or Runtastic[3]. Nowadays they are being increasingly used as means to track GPS traces, store and share human location histories. For instance, Flickr[4] allows geotagging photos, Twitter[5] maps tweets and interests, while Facebook[6] allows sharing and tagging locations representing particular interest.

When carefully processed, this data can provide important information for urban planning and management, vehicle tracking, monitoring and other tasks. Determining trajectories representing people's location histories and extracting people's most frequently visited locations from raw data can provide valuable information about human mobility patterns.

Next, we clarify the meaning of related terms.

Definition 2. ***GPS trajectory:*** *On a two dimensional plane, it is possible to sequentially connect raw GPS points into a curve based on time serials, and split this curve into GPS trajectories* (Tr) *if the time interval between consecutive GPS points exceeds a certain threshold* ΔT [40]. *Thus,* $Tr = p_1 \rightarrow p_2 \rightarrow ... \rightarrow p_n$*, where* $p_i \in P$*,* $p_{i+1}.T > p_i.T$ *and* $p_{i+1}.T - p_i.T < \Delta T (1 \leq i < n)$ [40].

The notion of trajectories and spatio-temporal data allows to build elementary human mobility models, for example, to understand classical *work-to-home* sequence by checking starting times of the trajectories. When

[1] https://www.foursquare.com/

[2] http://www.getupandrally.com/

[3] https://www.runtastic.com/

[4] https://www.flickr.com/

[5] https://www.twitter.com/

[6] https://www.facebook.com

analyzing large amounts of spatial data, it is often essential to preprocess and classify spatial data into groups, so that points within the same group are more similar to each other than those in disparate groups.

Definition 3. ***Geo-location:*** *A geo-location g stands for a geographic region where user stayed over a certain time interval and which carries some semantic meaning for the user. The extraction of geo-locations depends on two parameters: distance threshold (D_{threh}) and time threshold (T_{threh}). A group of consecutive GPS points $P \in \{p_m, p_{m+1}, \ldots, p_n\}$, where $\forall m < i \leq n, D(p_m, p_i) \leq D_{threh}$ and $|p_n.T - p_m.T| \geq T_{threh}$. With P, D_{threh}, T_{threh} a geo-location is defined as $g = (Lat, Lon, arvT, levT)$, where*

$$g.Lat = \sum_{i=m}^{n} p_i.Lat/|P|$$

$$g.Lon = \sum_{i=m}^{n} p_i.Lon/|P|$$

are average latitude and longitude of the collection P, $g.arvT = p_m.T$ is user's arrival time, $g.levT = p_n.T$ is user's leaving time and D is distance between GPS points [40].

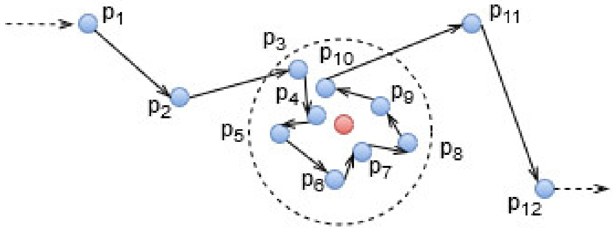


Figure 1: *Example of a geo-location.*

Geo-location (*Figure 1*) is nothing more than a sufficiently large group of non-randomly distributed GPS points that have accumulated in some place. We will use clustering techniques to discover those homogeneous groups in the data. There exist a countless number of different clustering algorithms and their variations, but in this work we will use density based clustering methods and their the most famous representative - DBSCAN algorithm [18]. Its applicability and ability to work with GPS data was also reviewed in [33].

Definition 4. ***DBSCAN:*** *Density-based spatial clustering of applications with noise algorithm uses notion of density reachability to discover clusters. Algorithm identifies all point p neighbours which are within distance ε. If number of such neighbours is greater than minimum predefined number minPts, points are considered as a part of a cluster, otherwise p is considered as a noise* [33]. *Algorithm terminates when all points have been visited. Average complexity of the algorithms is $\mathcal{O}(n^2)$.*

Algorithm usually uses Euclidean distance as metric for calculating distance between points, however, other distance metrics can also be used. DBSCAN algorithm does not specify the upper limit of how many objects may form a cluster and therefore detected clusters have wide variation in local density. Density based clustering algorithms are perfect for spatial data clustering given its distinctive features:

1. The ability to detect non-spherical clusters of arbitrary shape. Other clustering methods like hierarchical clustering or k-means algorithms fail in this regard.

2. The ability to discover noise and being robust to outliers. Algorithm required input parameters can be chosen in the way that sparsely distributed points will not be included in any cluster.

3. Speed and complexity - in worst case DBSCAN algorithm has $\mathcal{O}(n^2)$ time complexity. Furthermore, $\mathcal{O}(n \log n)$ complexity can be obtained by using indexed data structure. Numerous other clustering algorithms have considerably higher complexity.

Although, it should be noted that right now neither trajectories nor geo-locations carry any semantic value. We will enrich them with semantic meaning - it will provide us with better insights and open new possibilities for human mobility analysis.

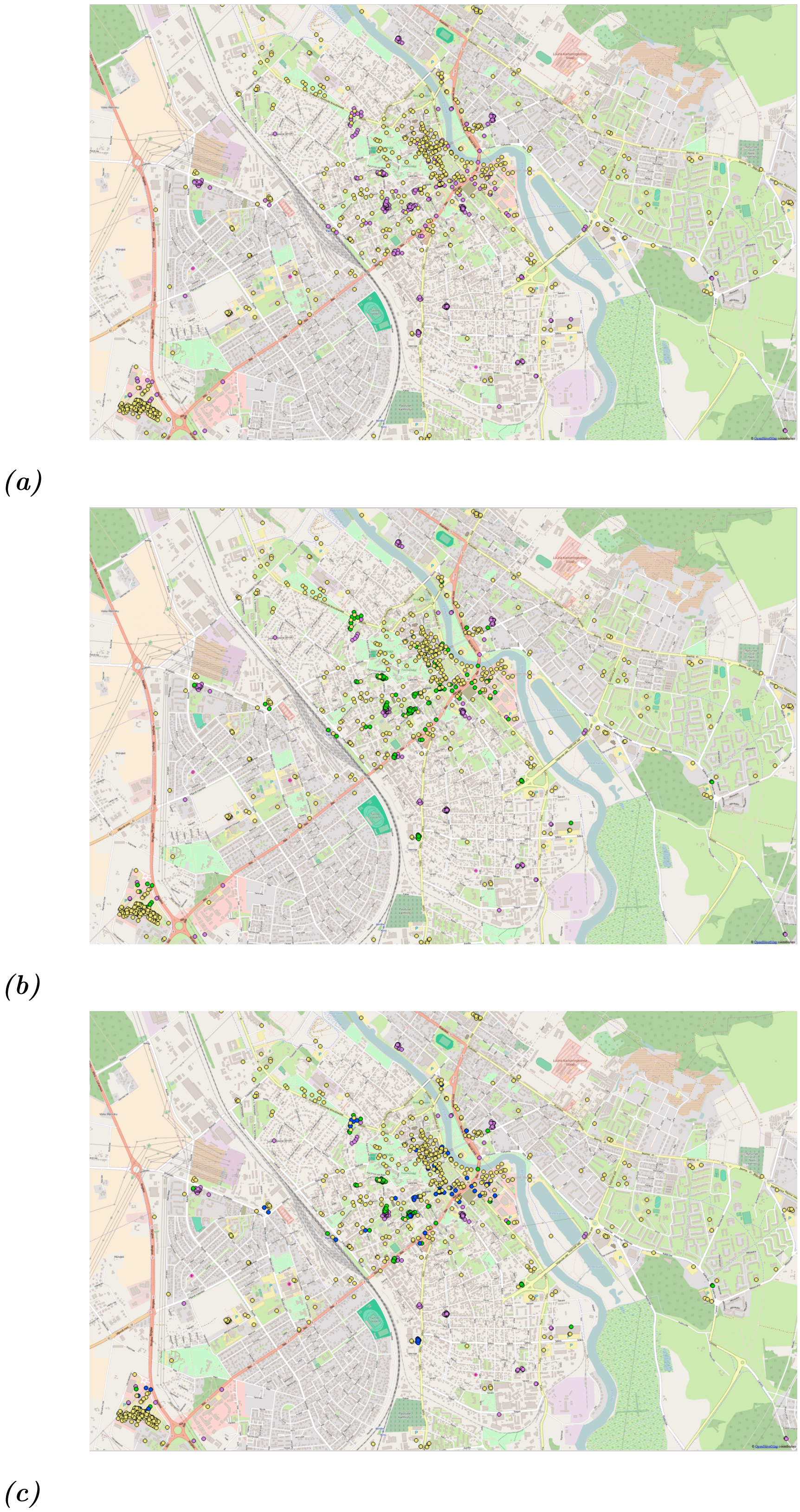
2.3 Semantic analysis

After extracting geo-locations we are going to recognize activities associated with those places and information about the types of businesses close to the location. Before doing this, we need to find a list of POIs (Points Of Interest) and public amenities located in study area. Such information can be extracted from different databases that store semantic category of POIs - we used OpenStreetMap database[7]. We will use a buffer around geo-location centroid to classify geo-location according to amenities falling into the buffer. This means, that a geo-location will be labeled with a semantic tag and associated with some activity. However, occasionally it is not possible to determine the type of the geo-location unambiguously as frequently many amenities are located next to each other. For example, when multiple restaurants are located inside a shopping mall or when public transport stops are in close proximity to post offices. POIs and classified geo-locations are depicted on *Figure 2*.

Second part of semantic analysis is to determine which locations are significant for the user. Significance can be indicated by time spent in a place [4] and our approach relies on measuring the time periods a person stays at each place and uses time threshold to distinguish significant and insignificant places. Determination of the correct thresholds is critical as we should be able to find out significant places, such as commonly frequented public areas like restaurants, sport centers, cinemas, etc., while ignoring places without semantic meaning, like waiting for traffic lights or being stuck in a traffic jam. When using smaller time threshold it becomes possible to extract more geo-locations representing small pauses, for example, less than 5 minutes, which are transit-locations between start point and destination [41].

All this implies that there is a relation between a spatial description and the social context of the human movement. Various pattern mining algorithms and methods, for instance Apriori algorithm [2] or FP-Growth algorithm [39], can be applied for exploring the relationship between geographic and semantic properties and as a result obtaining frequent semantic patterns of behaviours

[7] http://wiki.openstreetmap.org/wiki/Downloading_data



(a)

(b)

(c)

Figure 2: *POIs and classified geo-locations. (a) - POIs (yellow), all geo-locations (purple). (b) - POIs (yellow), all geo-locations (purple), classified geo-locations (green), (c) - POIs (yellow), all geo-locations (purple), classified geo-locations (green), geo-locations with more than one POI nearby (blue).*

of people. Proposed conception allows to capture sociological aspects of human movements - it becomes possible to build more complicated models and comprehend why people have chosen that particular path and decided to make a stop in that particular place. For example, it becomes possible to mine and understand the classical *landmark-to-bar* travel sequence: an individual would be more likely to go to a bar after visiting a cultural landmark than they would before [40]. Another example is illustrated on *Figure 3*.

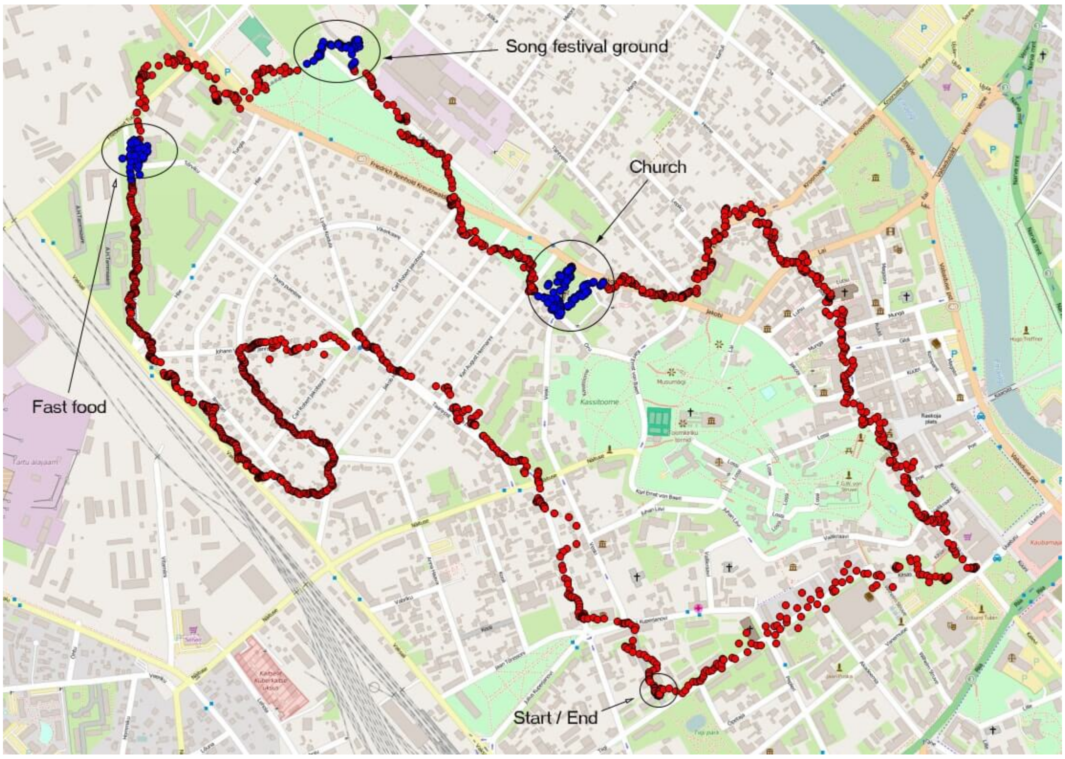


Figure 3: *Path containing 3 classified geo-locations.*

3. Design and Technology

This section describes the data and software technologies used in the thesis.

3.1 Used technologies

Almost all code is written in Groovy[8] - a modern dynamic language for the Java platform. Gradle system[9] was used for building and running the code. Small scripting tasks like data import were done in Python[10], statistics and data analysis were done in R[11]. PostgreSQL[12] - open source database server, was used to store the data. Also, PostgreSQL was extended with PostGIS[13] extension - software that adds support for geographical objects and allows to perform aggregation functions over them. Visualizations was done using QGIS[14] software - cross-platform and open-source desktop GIS application for geographical data viewing and analysis.

3.2 Data source

In this paper we will use the data collected by “MobCollector” - mobile application created by Distributed Systems Group of University of Tartu[15]. Main goal of the application is to record GPS and mobile data: basic location information (user identificator, timestamp, latitude, longitude, speed, quality of signal, strength of signal) and mobile identificator (mobile country code, mobile network code, location area code, cell ID, network type). User interface

[8] http://www.groovy-lang.org/

[9] http://www.gradle.org/

[10] https://www.python.org/

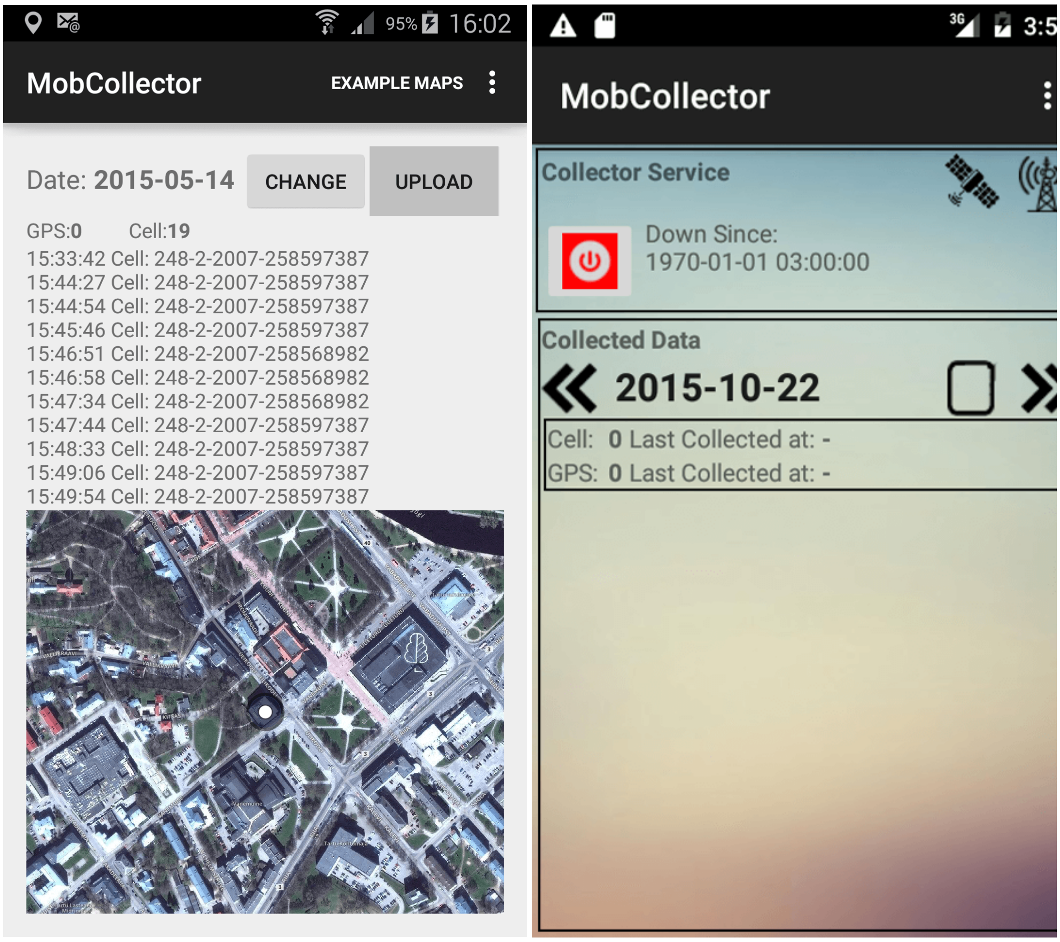
[11] https://www.r-project.org/

[12] http://www.postgresql.org/

[13] http://www.postgis.net/

[14] http://www.qgis.org/en/site/

[15] http://www.ds.cs.ut.ee/



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
```

Down Since:
1970-01-01 03:00:00

2015-10-22

Cell: 0 Last Collected at: -
GPS: 0 Last Collected at: -

Figure 4: *User interface of "MobCollector" application*

of the application can be viewed on *Figure 4*. Application was installed on mobile phones of 13 users and worked in background mode. Data was collected for a period of 6 months from March to September 2015. During the data collection period people used different transportation modes, such as walking on foot, riding a bicycle or driving a car. Collected data was a high-sampling-rate data, which means that time granularity for every GPS point is around 3-10 seconds. Temporal spacing of the records is irregular. Different representations of used GPS data and created trajectories can be viewed in *Figure 6*.

We are using real world data to demonstrate effectiveness of our approach. As observed in [27] real world mobility models are statistically different from those generated from commonly used synthetic mobility models such as random waypoint [11] and Brownian motion [12].

ID	User ID	Date	Time	Lat	Lon	Trajectory ID
1	1	2015-03-21	14:50:54	58.37430482	26.71254817	1
2	1	2015-03-21	14:51:01	58.37385347	26.71122877	1
3	2	2015-06-15	20:17:45	58.3774068	26.6853793	2
4	2	2015-06-15	20:17:48	58.37768334	26.68471776	3

***Table 1:** Example GPS log*

Initial dataset contains 273 625 GPS points (*Table 1*), which we store in PostgreSQL relational database. Each GPS point record has variety of different properties, however, in this work we will concentrate only on GPS data and on following properties: latitude and longitude coordinates in EPSG:4326 coordinate system, timestamp and user ID. Database schema showing fundamental structure is shown in *Figure 5*. Notwithstanding, we were using plenty of other tables for holding intermediate results, doing analytics and prediction.

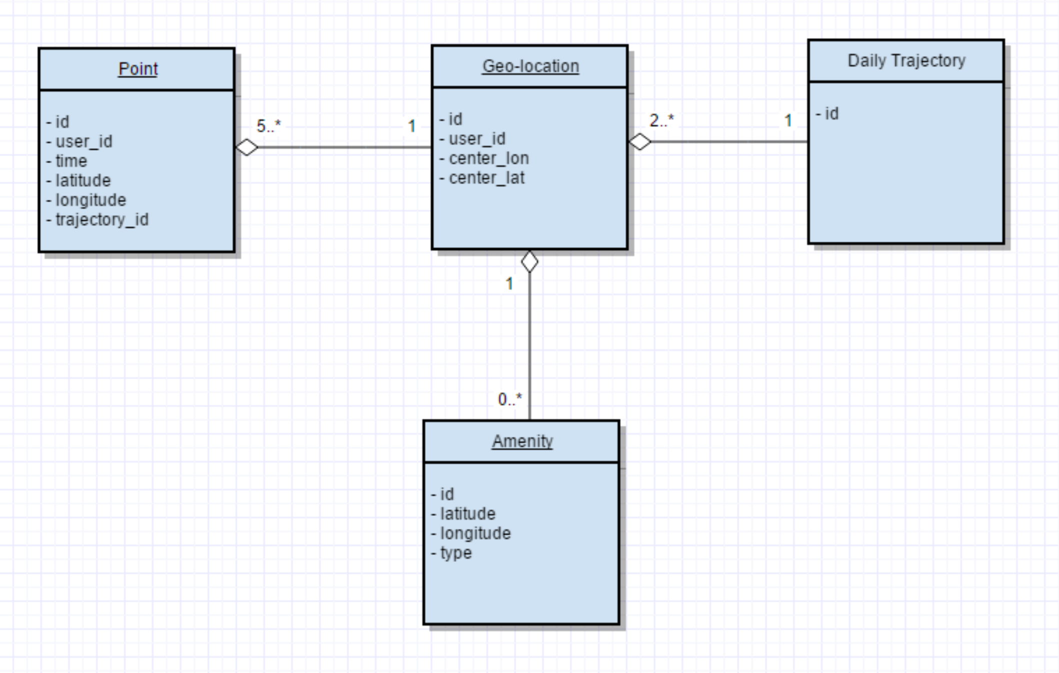
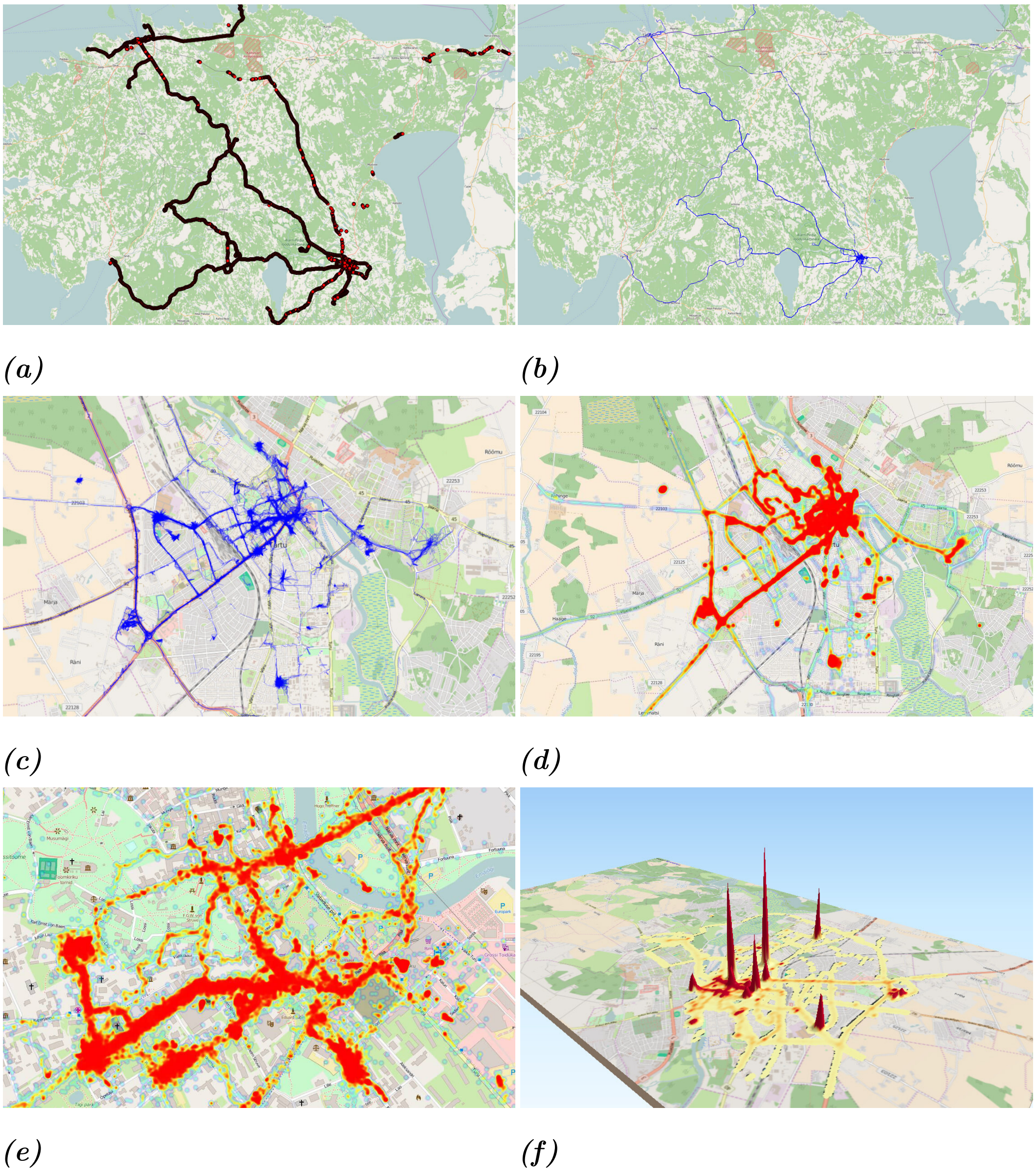


***Figure 5:** Internal data model*

Most parts of the data were collected predominantly in Tartu, Estonia. *Figure 6* depicts the distribution of the GPS data used in the experiment. Considering the privacy issues, we use all the data anonymously.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 6: *Representation of used GPS data and computed trajectories. (a) - All GPS points, small scale. (b) - All computed trajectories, small scale. (c) - Computed trajectories in Tartu city, large scale. (d), (e), (f) - Heatmap of all GPS points in Tartu city, small scale.*

4. Methodology

This chapter describes a framework for analysis and prediction of human mobility. The workflow is shown on *Figure 7* and is as follows: first, we extract trajectories from raw GPS logs of all users, then extract geo-locations from trajectories and enrich them with semantic and temporal tags. As a final step before starting with prediction, we unite geo-locations into daily trajectories.

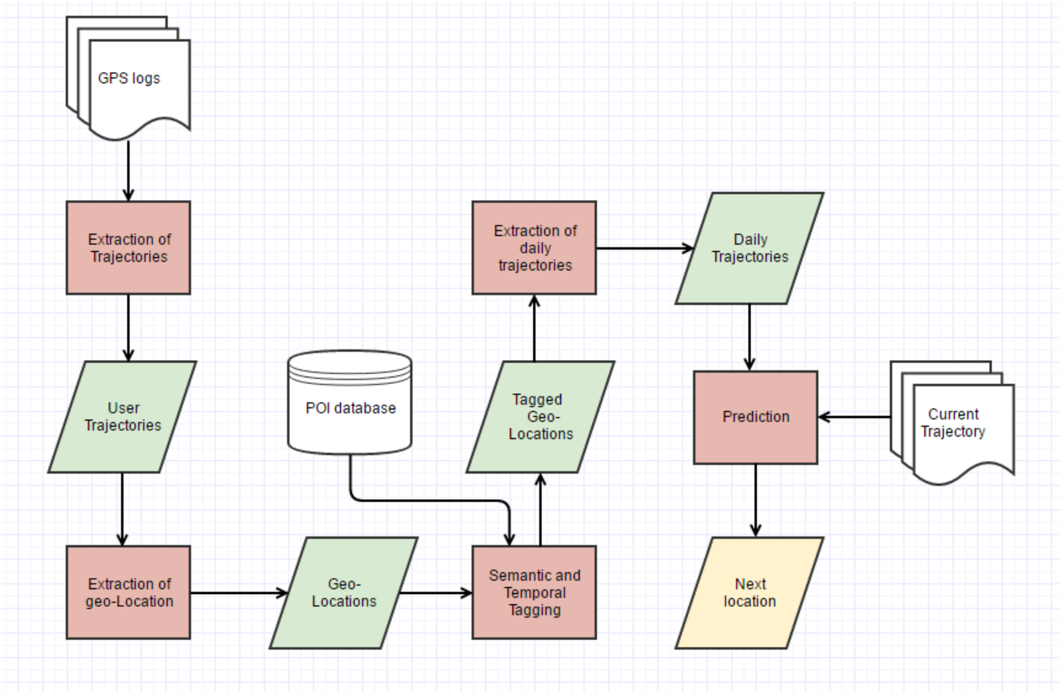


Figure 7: *Prediction workflow*

4.1 Data preprocessing

Foremost, initial raw GPS dataset was preprocessed and cleaned. Cleaning data is a process used to determine and improve inaccurate, incomplete and unreasonable raw data [14]. There are some degree of errors and omissions in any GPS data, because there are many factors that contribute to the accuracy of GPS recordings. It is necessary to understand causes of the errors in data to successfully clean and improve raw GPS point locations [14].

One of the most common GPS measurement errors is related to the GPS jumping around and thus showing incorrect location. We encountered lots of errors preventing us from plotting the data and performing proper analysis. The issue was fixed in two steps:

1. Finding and deleting users duplicate GPS points - we compared longitude and latitude of the GPS points and removed duplicates from the dataset.

2. Adding constraints to raw GPS data when extracting trajectories - two consecutive points belong to the same trajectory if only the distance and time between them is respectively less than 200 meters and 1 minute. The distance between points from their longitudes and latitudes was calculated using Haversine formula. **Definition 5.** ***Haversine distance:***

   $$d = 2r\sin^{-1}(\sqrt{\sin^2(\frac{\varphi_2 - \varphi_1}{2}) + \cos(\varphi_1)\cdot\cos(\varphi_2)\cdot\sin^2(\frac{\psi_2 - \psi_1}{2})}),$$

   where r is sphere radius (6371 km), ϕ is latitude and λ is longitude.

   After that we delete all GPS points that belong to trajectories containing less that 5 GPS points as we consider them uninformative and not providing any value for location prediction - on average their duration is ≈20 seconds, distance less than 100 meters and they do not present any meaningful movement activity.

First, we preprocessed raw GPS logs and extracted 273 625 GPS points belonging to 13 unique users. Next, we applied above mentioned data cleaning techniques: 14 153 GPS points were deleted and in total there left 259 472 GPS points, which formed 2548 trajectories, on average 102 GPS points in the trajectory. After the whole dataset was processed we did not add any additional links, fields or relations between GPS points. Furthermore, we did not apply any map matching algorithms as we concentrated on places where people spend significant amount of time and those places might not always be located near roads or other mapped paths that can be found in free and open datasets.

4.2 Location prediction

To start with, we describe what is location prediction. It can be defined as an approach for identification of the next location user is most likely to visit. In a nutshell, the process is very similar to the process of recommending next location using some kind of recommender system [22], such as Teleport[16], for instance. However, there exist one important difference between those approaches: recommender methods do not take current location of the user and movement dynamics into account, when mobility prediction method do.

Human mobility prediction is very interesting topic as the criterias people use to choose next location are very different - rational and irrational, subjective and objective. Decision can be influenced by many factors, since every individual has different cost functions [41]. This implies, that usually there exist a reason, other than interestingness, why individual decides to visit some particular location. These reasons can be very different, starting from sport activities and ending with social intentions, but according to [37] all they can be categorized into three classes:

1. Geographic-triggered intentions
2. Semantic-triggered intentions
3. Temporal-triggered intentions

This means that movement of the individual can be considered as a behaviour driven by at least one of the enumerated intentions. However, in practice, several intentions act as a trigger to change a location. That is why we decided to predict mobility behavior by taking all geographic, semantic and temporal properties into account. We believe that simultaneous consideration of all three properties will result in efficient model as all they have a direct impact on the prediction task and cannot be omitted.

[16] https://www.teleport.org/

Definition 6. ***Location prediction:*** *Given a set of users U and a set of locations L, the problem of location prediction can be formulated as an estimation of the probability of a given user visiting a given location based on one's current movement* [37].

$$f(l|u,t) \rightarrow [0,1],$$

where $u \in U$, $l \in L$ and t is u's current movement.

There are a variety of ways and algorithms for that, each with its own advantages and drawbacks. Some example algorithms and approaches can be found in [7,13,17,25]. However, very often they are either bounded to a specific case or to a specific dataset, like in [1], for instance.

One of the patterns that we observed during analysis of the dataset, is that people did not track GPS permanently and turned tracking device on only while some activity, for instance, when walking from or to somewhere. Therefore, ordinary trajectory usually consists of one or two geo-locations, which in turn does not provide a full picture of user's movement when analyzing it in isolation. Taking into account the fact that user trajectories are often linked to each other, for example path from home to work in the morning and visiting grocery store when going back home from the work in the evening, we decided not to analyze user's trajectories separately, but combine them into one day time intervals. Such separation provides more natural overview of the movement as well as a full picture of daily activities.

Definition 7. ***Daily trajectory:***

$$Dtr = g_1 \rightarrow g_2 \rightarrow \cdots \rightarrow g_n, \text{ where } g_i \in G,\ g_{i+1}.T > g_i.T\ (1 \leq i < n),$$

where G is a set of geo-locations, T is one day period from 0:00 to 23:59 and $g.T$ is user's geo-location arrival time.

We strongly believe that combining different location prediction algorithms covering various aspects of human mobility can be more efficient, which will result in less error prone and more unified prediction model. We will split the prediction into two parts and then combine the results.

Following techniques will be applied:

1. Predicting position of the next geo-location location on the map.

    - We find the approximate distance to the next geo-location by calculation intra-distances between geo-locations of the daily trajectory.

    - We find approximate direction of the movement by calculating route similarity index between daily trajectories and choosing the most similar ones.

      Also, we will check if following techniques increase prediction probability:

        - We choose only those daily trajectories that have the similar starting area as in examined daily trajectory.
        - We choose only those daily trajectories that intersect with the ending area of the examined daily trajectory.

2. Predicting the type of the next geo-location by analyzing semantic patterns. We will apply first and second order Bayesian inference and analyze which one gives higher prediction probability.

3. We examine how temporal aspect affects prediction accuracy. We will add temporal information about geo-locations to our prediction model to determine the mathematical relationship between the variables. For instance, authors of [36] managed to improve the prediction by 9% by considering temporal-social ties in their model.

During the prediction phase, we will also investigate how short transitions between geo-locations affect prediction success rate. Usually, those transitions are related to GPS measurement errors and denote that person stays on the same place.

Definition 8. ***Short (insignificant) transition:*** *A transition between two geo-locations g_1 and g_2 of a daily trajectory Dtr, where $d(g_1, g_2) <$ 50 meters.*

Definition 9. ***Long (significant) transition:*** *A transition between two geo-locations g_1 and g_2 of a daily trajectory Dtr, where $d(g_1, g_2) >= 50$ meters.*

4.2.1 Predicting next geo-location on the map

This section focuses on mobility prediction techniques driven by geographically-triggered intentions, which study the sequences of visited geographic areas. Public transport is a great example - it follows particular predefined routes and given stops A and B, we can predict B as a next location for the user who is currently at A.

4.2.1.1 Average distance between geo-locations

The first component that we will analyze is the distance between geo-locations as authors of [41] found, that the choice of the next geo-location is greatly influenced by distances between previous geo-location transitions.

Definition 10. ***Geo-location inter-distance:*** *Geo-location inter-distance d is defined as the the length of shortest path between two sequential geo-location centroids c_1 and c_2.*

We will calculate the distance using Haversine formula, see *Definition 5*. According to [41] inter-distance distribution follows an upper-truncated Pareto distribution, which implies that humans generally prefer short paths between geo-locations and take long jumps less frequently.

We will try out different approaches and see which one gives better results:

1. Calculate the average distance between all transitions.

2. Calculate the average distance between all significant transitions.

3. Calculate the average distance between all transitions that fit into the interval between first and third quartiles.

4. Calculate the average distance between significant transitions that fit into the interval between first and third quartiles.

5. Calculate the distance of the last significant transition.

4.2.1.2 Route similarity index and direction of the movement

Among measuring geo-location intra-distances, we have chosen route similarity index as a second component in location prediction. We analyze and predict the behaviour of a user in accordance with the akin behaviour of other users, meaning that they tend to follow the same paths and do stops in the same places. Similar trajectories coincide in space, have similar shape and dynamic behaviour. However, they do not necessarily coincide in time - for example, moms on maternity leave often visit playgrounds and children's stores in different time.

There are many different algorithms for finding how similar trajectories are, most popular ones are described in [26, 35] and use different variations of spatio-temporal filtering and spatio-temporal distance. We decided to use Hausdorff distance algorithm to measure how far trajectories X and Y are from each other.

Definition 11. ***Hausdorff distance:*** *Hausdorff distance $d_H(X,Y)$ is defined by*

$$d_H(X,Y) = max\{\sup_{x\in X}\inf_{y\in Y} d(x,y), \sup_{y\in Y}\inf_{x\in X} d(x,y)\},$$

where X and Y are two non-empty subsets of a metric space (M,d).

Informally speaking, [9] defines Hausdorff distance as a longest distance you can be forced to travel by an adversary who chooses a point in one of the two sets from where you then must travel to the other set. In our case it is the greatest of all the distances from a point in the daily trajectory Dtr_1 to the closest point in the daily trajectory Dtr_2. This means that every point of either trajectory is close to some other point in the other trajectory. We applied PostGIS implementation of Hausdorff distance, where result units are in the units of spatial reference system of the trajectory geometries.

Next step is to calculate the movement direction of the examined daily trajectory Dtr_e based on the direction of the most similar daily trajectory Dtr_s. In order to find a Dtr_s, we calculate Hausdorff distance between Dtr_e

and all other trajectories and pick one with the lowest metric. It should be also noted that Hausdorff distance is the same when moving from A → B and from B → A, thus we additionally compared distances between first and last points of Dtr_e with first and last points of Dtr_s. This allowed us to get a movement direction. Thereafter, we calculated the bearing[17](*Figure 8*).

Definition 12. ***Bearing:*** *An angle between the north-south line of Earth or meridian and the line connecting the target and the reference point. Formula:*

$$\theta = atan2(sin\Delta\lambda \cdot cos\phi_2, cos\phi_1 \cdot sin\phi_2 - sin\phi_1 \cdot cos\phi_2 \cdot cos\Delta\lambda),$$

where $\phi_1\lambda_1$ is the start point, $\phi_2\lambda_2$ the end point, $\Delta\lambda$ is the difference in latitude.

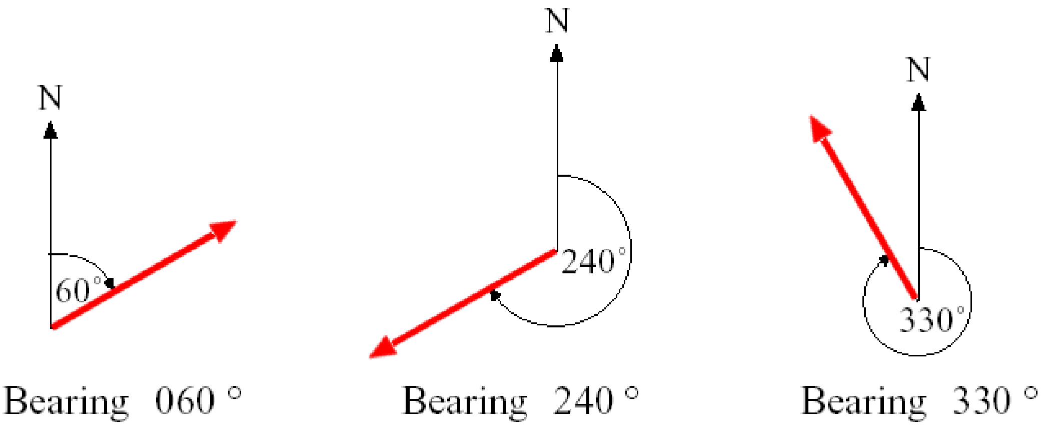


Figure 8: *Example of different bearings*[18]

We will try different approaches for calculation of bearing and see which one gives better results:

1. Calculate the average bearing between all geo-location transitions of Dtr_s
2. Calculate the bearing between penultimate and the last geo-location transition of Dtr_s.
3. Calculate the bearing of last significant transition of Dtr_s.

[17] http://www.movable-type.co.uk/scripts/latlong.html

[18] http://www.cimt.plymouth.ac.uk/projects/mepres/book8/bk8i11/bk8_11i3.htm

4.2.1.3 Similar starting area

In this subsection we analyze the third component that we will consider when predicting human movements - sharing the similar starting area. According to [3] spatially close trajectories have similar start and end areas, very often they are even identical to each other, see *Figure 9* for example. A good example might be paths starting from work or home such as visits to the gym or to the restaurant. In order to calculate the starting area, we cannot just simply take the first point of the examined trajectory as due to the GPS measurement errors trajectories rarely will have exactly the same starting point. Instead, we apply a buffer around the starting area and find all trajectories whose starting point is in the buffer.

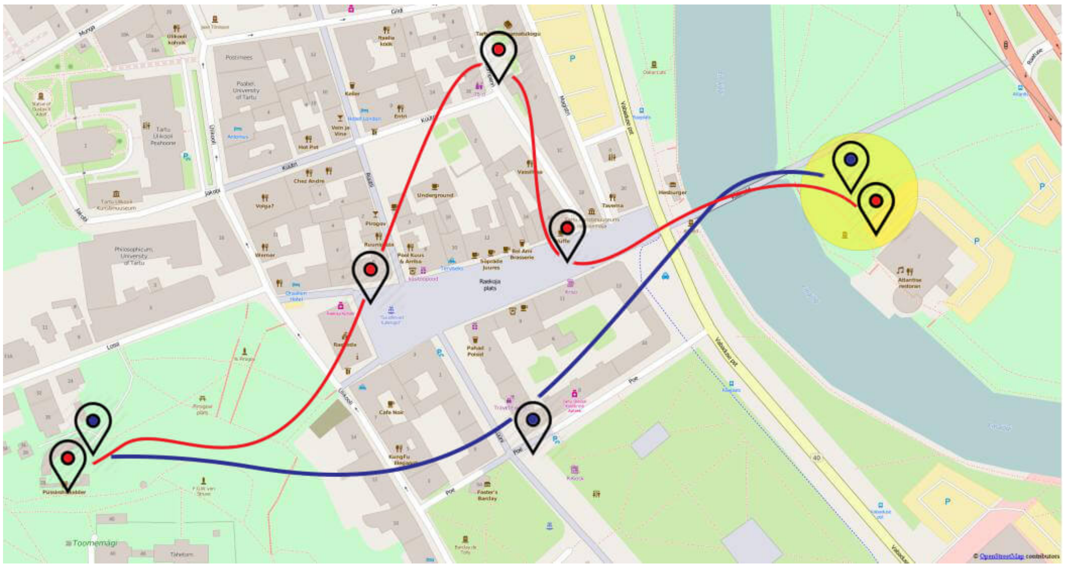


Figure 9: *Similar daily trajectories with the similar starting area.*

4.2.1.4 Intersection with the ending area

Fourth component, that we will concern as influencing factor, is the intersection with the ending area of the examined trajectory. In other words, we will find all trajectories that go through ending area of the trajectory. It may appear that it is completely useless factor that does not add any value to the prediction of the next location, but we believe that it is not - tourism area is a vivid confirmation of this. Tourist routes that go through city culturally important places or commonly frequented public areas have very similar dynamics, for example, well known *“Trafalgar Square-to-Big Ben-to-Westminster Abbey”* sightseeing route. Countless number of people follow this route and given a tourist who started his path in the hotel, visited Trafalgar Square and reached Big Ben, with a high probability we can expect that user’s next geo-location will be a Westminster Abbey. *Figure 10* illustrates similar example.

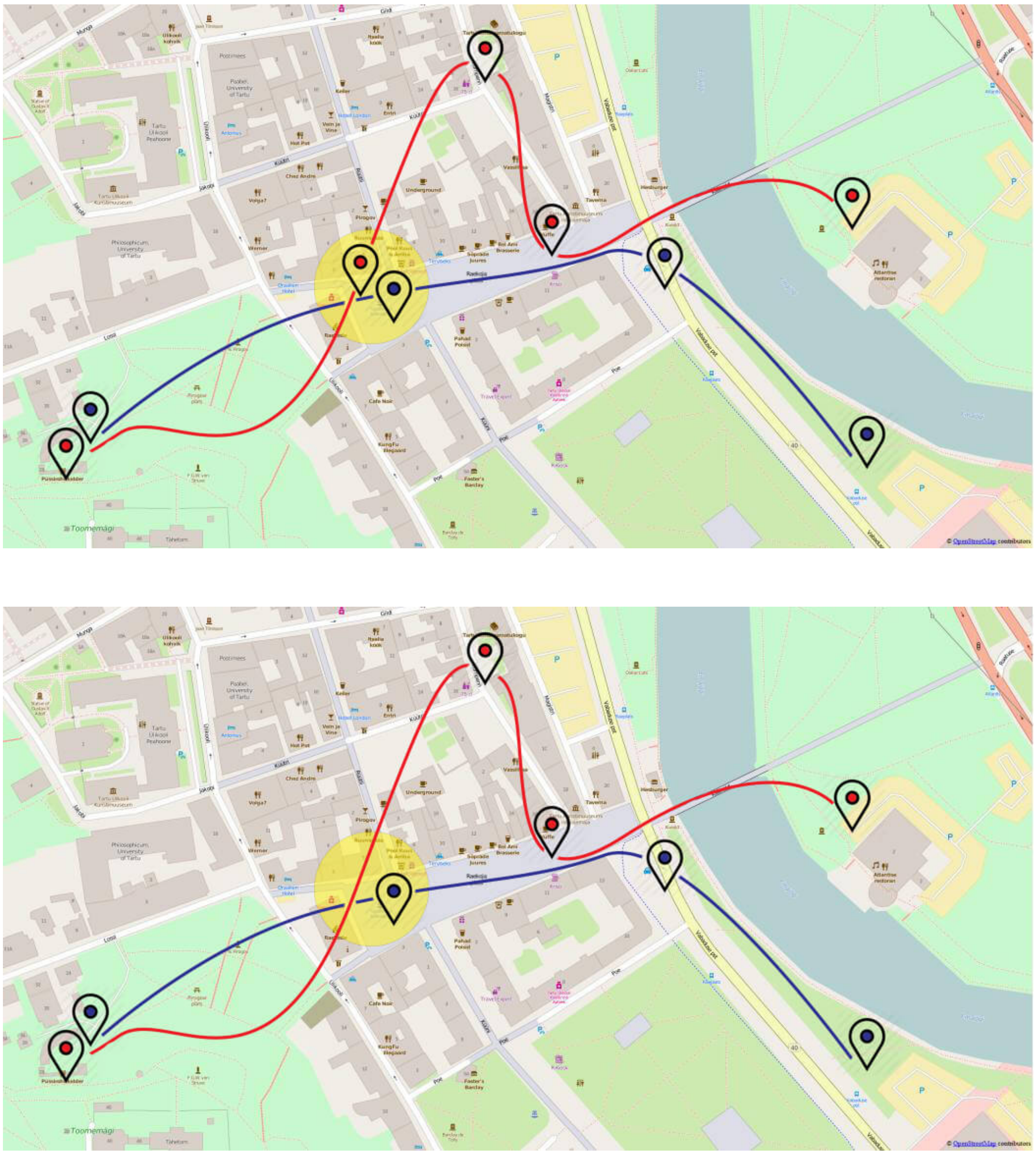


Figure 10: *Similar daily trajectories with intersecting ending areas.*

4.2.2 Predicting the type of the next geo-location

Next component of the proposed model is semantic analysis and its role in mobility prediction. Semantic-triggered intentions reflect and reveal the reasons why people visit some specific locations preceded by some other locations. For example, going from home to work is very common sequence while dining out twice in a row is rare [28]. As another example, we can consider people working in the office and leaving for a lunch - we can predict that ensuing geographical region will contain many shops, eateries or restaurants. We are analyzing movements of real people, thus, capturing sociological aspects can provide very good insights and be very promising for predictions.

In our work we pay attention to static phases of movement as they characterize some interest to the particular place and can be used to form a sociological portrait of the person. Such approach allows us to create a map of points of interest, that includes significant places, such as home, workplace, shopping centers, meeting places as well as important routes used to get from one place to another [28]. Such map can be either personal and applied for recognition of individual's behavior and location prediction or aggregated for all users. In our work we will use the latter approach and create a map that will contain semantic information of all users.

For predicting the type of the next geo-location we will use first and second order Bayesian inference. When applying second order Bayesian inference, the probability of the next geo-location g_{n+1} type depends on both current g_n and previous g_{n-1} geo-location types. In case of the first order Bayesian inference, only current geo-location g_n is taken into account.

Definition 13. ***First order Bayesian probability:***

$$P(g_{n+1}|g_n) = \frac{P(g_n|g_{n+1}) \cdot P(g_{n+1})}{P(g_n)},$$

where $P(g_{n+1})$ and $P(g_n)$ is a relative number of occurrences of g_{n+1} and g_n geo-location types in the past and $P(g_n|g_{n+1})$ is a relative number of transitions from g_{n+1} to g_n.

Definition 14. ***Second order Bayesian probability:***

$$P(g_{n+1}|g_{n-1}, g_n) = \frac{P(g_{n-1}, g_n|g_{n+1}) \cdot P(g_{n+1})}{P(g_n|g_{n-1}) \cdot P(g_{n-1})},$$

where $P(g_{n-1}, g_n|g_{n+1})$ is a relative number of transitions from g_n to g_{n+1} knowing that the user was in g_{n-1} before in the past, $P(g_{n+1})$ is the relative number of occurrences of g_{n+1} geo-location types in the past and $P(g_n|g_{n-1})$ is a relative number of transitions from g_{n-1} to g_n.

The ability to use the n-th order Bayesian inference raises the question of what order model will result in an increase of a predictive power [21]. However, in practice with higher order models the quantity of the data is a limiting factor - transitions required for higher order Bayesian inference may not exist in the training database and their probability will be zero. Furthermore, computational cost of building the model increases as it requires more time and resources to train and store the prediction model. For this reason, we decided to limit ourselves with a second order model.

4.2.3 Adding temporal aspect

As final component of our model we will consider the relationship between locations, activities and temporal information. As observed in [15], human movement exhibits strong temporal cyclic patterns in terms of the hour of the day and the day of the week [36]. These kinds of intentions reflect the reasons why users visit and leave locations at a certain time [37]. It was shown in [15], that temporal-triggered intentions and periodic behavior explains about 50% to 70% of all human movements. For instance, it is very common that person leaves home in the morning time, works whole day in the office and comes back home in the evening. Moreover, people are used to lead a quiet life during the weekdays and do social and family activities during the weekend. Such temporal information can help us to identify more common mobility patterns and establish links between them. To the purpose of better understanding and improving movement prediction, we calculate the probability of the next location considering both spatial and semantic information combined with daily timestamp.

Next, we enriched geo-locations with temporal information. We split a day into 3 periods:

1. Night (00:00-08:00)
2. Daytime (08:00-17:00)
3. Evening (17:00-00:00)

Afterwards, we classified geo-location arrival time according to those periods. Example of classified subset of geo-location of 3 users can be viewed on *Figure 11*.

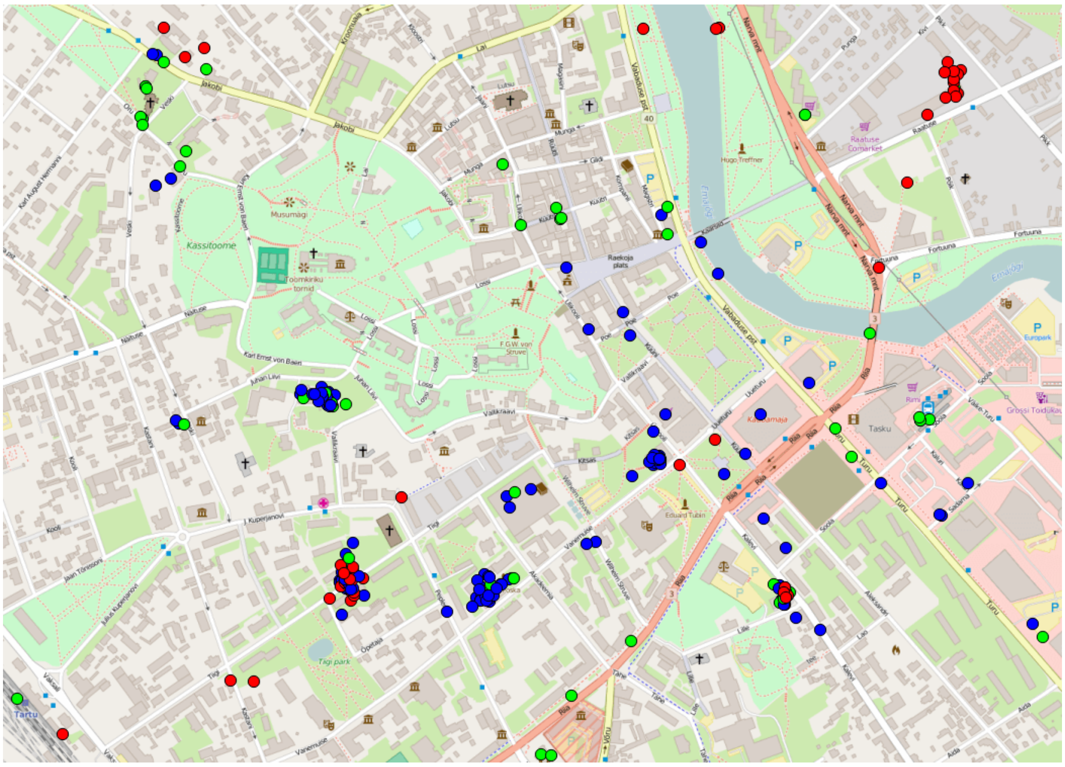


Figure 11: *Subset of geo-locations of 3 users classified by time. Colors: red - night activity (00:00-08:00), blue - daytime activity (08:00-17:00), green - evening activity (17:00-00:00).*

5. Experimental Results and Analyses

The intent of this chapter is to present the experiments and the results obtained by analyzing and predicting human mobility patterns. We reveal the implementation details and carry out the case study to answer the questions and problems declared in *Chapter 1*.

5.1 Data overview

Before doing human mobility prediction, lets look at the data we have. High quality data is a key to success, thus, first of all, we preprocessed and cleaned the data (see *Chapter 4.1*). During that process we removed about 5.1% of all GPS points. More detailed statistics about cleaned GPS data and constructed trajectories can be found in *Table 2* and *Figure 12*.

The total distance of all GPS trajectories exceed 4729.3 kilometers. The longest trajectory is a nearly complete Tallinn - Tartu car trip (138 km), the most durable trajectory (21.86 hours) represents human movements inside an apartment during a weekend. Human transportation mode is also detectable from *Figure 12*. There are two large and dense accumulations of points, which correspond to two different travel modes: by car (lower one) and on foot (in the middle).

Metric	**Value**
Average number of points in trajectory	112
Mean distance of trajectory	1622 meters
Max distance of trajectory	138 700 meters
Mean time of trajectory	22.9 minutes
Max time of trajectory	21.86 hours

Table 2: *Detailed statistics about GPS trajectories.*

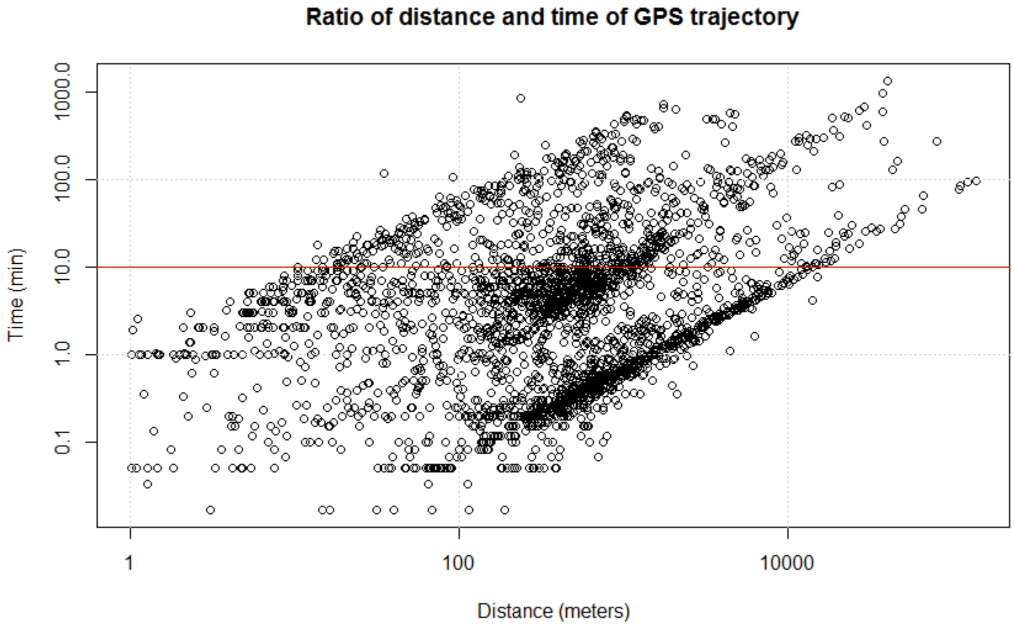


Figure 12: *Ratio of distance and time of GPS trajectory. Both axes are logarithmic. Red line indicates 10 minutes threshold used for geo-location extraction.*

5.2 Extracting geo-locations and daily trajectories

Next task is to find and extract geo-locations. Foremost, we find spatially close geographical areas by clustering GPS points and detect geo-locations from those areas. In this experiment we are using DBSCAN clustering algorithm and set *minPts* to 10 points and ε to 30 meters. This means, that the cluster will be created, if there will be at least 10 consecutive GPS points at a distance of 30 meters from each other. Also, we set *Dthreh* to 300 meters and *Tthreh* to 10 minutes. In other words, cluster is a geo-location if an individual stays over 10 minutes within a distance of 300 meters. These two parameters enable us to find significant places, such as restaurants and shopping malls, etc., while ignoring the geo-regions without semantic meaning, like the places where people wait for traffic lights or meet congestion [40].

In total 786 unique geo-locations were extracted from the dataset, which means that on average single trajectory contains 0.3 geo-locations. On the other side, 182 036 or 70.1% of all GPS points fall into geo-locations. This leads us to the conclusion that the data was collected by the people not inclined to the active movement and leading a quiet life. On average each geo-location contains 232 GPS points. Movement activity of three users was very low and thus we did not manage to extract any geo-locations from their movements. An example of a GPS trajectory containing three geo-locations can be viewed on *Figure 3*.

Next step is to classify extracted geo-locations and recognize activities associated with those places. First of all, we parsed OpenStreetMap database and extracted 683 unique POIs located in Tartu city. Then we classified them into seven different categories:

1. Public buildings: police, post office, hotel, etc.
2. Food: cafe, restaurant, etc
3. Transportation: gas station, parking lot, bicycle parking, etc.
4. Entertainment: museum, nightclub, gallery ,etc.
5. Education: library, university, school, etc.
6. Shopping: shop, shoemaker, tailor, etc.
7. Residential buildings

Type	**Amount**
public buildings	213
food	104
transportation	79
entertainment	55
education	44
shopping	96
residential building	516

Table 3: *Number of geo-locations in classified POI groups.*

The main reason we did not use a native classification of OpenStreetMap POI system is that it consists of 71 different categories[19], most of which in our case will not contain any POIs. We used a 75 meters buffer around geo-location centroid and checked the intersection of the buffer with a POI. Geo-location can belong to multiple POI classes as there might be multiple POIs inside a buffer. If no POIs were located inside a geo-location buffer, we classified geo-location as residential building. The division of all geo-locations into groups is presented in *Table 3*.

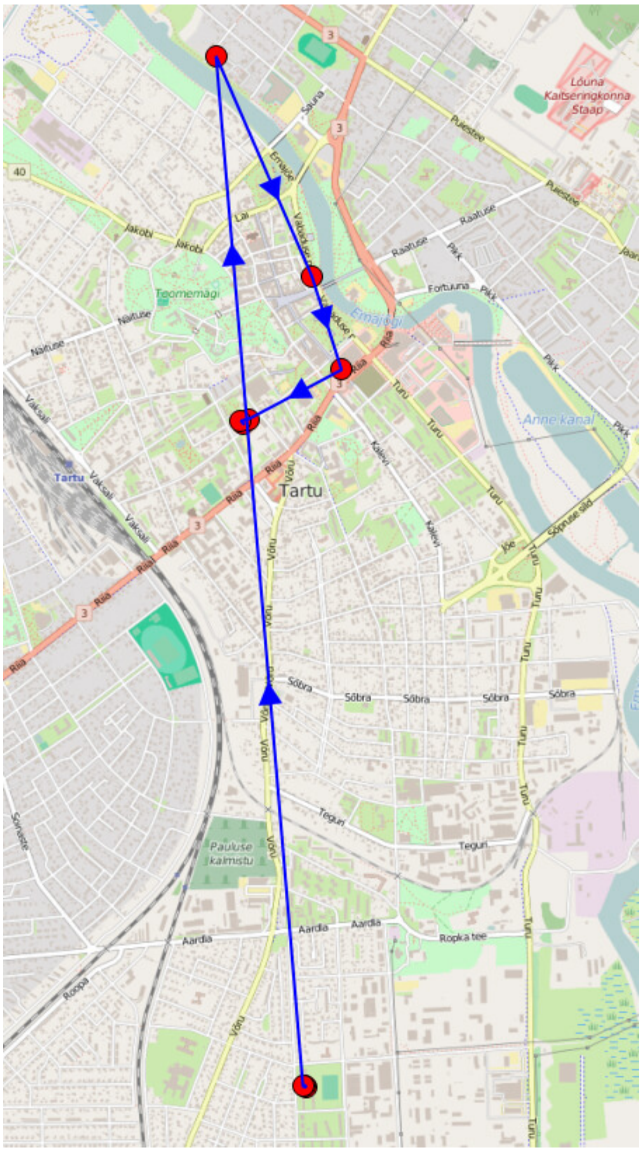


Figure 13: *Example of a daily trajectory.*

[19] http://wiki.openstreetmap.org/wiki/Map_Feature

As a final step we extracted daily trajectories: in total we got 136 unique daily trajectories containing at least two geo-locations. An example of daily trajectory is illustrated on *Figure 13*. On average, each daily trajectory consists of 6 geo-locations, what means approximately of 1138 GPS points. On maximum, there were 17 geo-locations and on minimum two geo-locations in a daily trajectory.

5.3 Evaluation criterias

Before discussing the results we present evaluation criterias which we use to explore the effectiveness and performance of our prediction model. We will use following metrics:

1. Percentage of correct predictions:

$$\text{\% of correct predictions} = \frac{\text{number of correct predictions}}{\text{total number of predictions}} \cdot 100$$

2. Percentage of wrong predictions:

$$\text{\% of wrong predictions} = \frac{\text{number of wrong predictions}}{\text{total number of predictions}} \cdot 100$$

3. Percentage of failures to make a prediction:

$$\text{\% of failures} = \frac{\text{number of failures}}{\text{total number of predictions}} \cdot 100$$

To verify the location we apply 200 meters buffer around the probationary geo-location and check if predicted geo-location is inside the buffer. We are checking against the buffer due to the fact that each GPS point of the geo-location may have its own measurement error, hence, geo-location position cannot be accurate enough. For this reason we decided to define the position of the geo-location by the position of its centroid. Taking that into consideration as well as the size of the Tartu city, we decided that 200 meters is appropriate buffer size for our test.

We will use 80/20 principle for training and testing the model. Due to the fact that we are operating with a relatively small dataset, data cross-validation will be applied: we partitioned data to training subset (80%) and validated our tests on testing/validation set (20%). To reduce variability we perform such analysis 20 times for each test and use averaged results.

We are considering our model to be able to predict next geo-location of the user when one appeared for the first time in the system.

5.4 Results

This chapter presents the results achieved after applying proposed model to our dataset. We discuss the advantages and disadvantages of the model as well as reveal the problems encountered during the implementation.

Our workflow is as follows: first of all we try to predict the next location by concentrating only on geographic-triggered intentions - this means we will take only geographic properties into account. As a next step, we will analyze the prediction potential of semantically-triggered intentions and add them to our model. As a final step, temporal aspects will be considered when predicting a location.

5.4.1 Predicting next geo-location on the map

Given a geo-location g_1, prediction of the next geo-location g_2 consists of two core components: (i) distance and (ii) bearing. If at least one of the components is predicted incorrectly, the whole prediction is also incorrect. Thus, we try to find methods with the highest prediction success rate separately for each component. For predicting the next geo-location we took the most successful methods and combined them together.

To start with, we concentrated on prediction of distance and bearing to the next geo-location. We tried out different approaches - their detailed description can be found in *Chapter 4.2.1.1* and *Chapter 4.2.1.2*. As we wanted to get the highest success rate for each component separately, in our tests we assumed that in the prediction equation other needed component is known except for searched one: we used correct bearing for distance prediction and correct distance for bearing prediction. To validate the correctness of the prediction,

Method	Correct	Wrong
Average distance between all transitions	24%	76%
Average distance between significant transitions	29%	71%
Average distance between all transitions that fit into the interval between first and third quartiles	45%	55%
Average distance between significant transitions that fit into the interval between first and third quartiles	48%	52%
Length of last transition	24%	76%
Length of last significant transition	57%	43%

Table 4: *Comparison of methods for predicting distance to the next geo-location.*

we used a 200 meters buffer around a geo-location. Results are presented in *Table 4* and *Table 5*.

The most successful result is achieved by taking the length of the last significant transition - we are able to predict the distance in more than half of the cases. As we can see, the results achieved by calculating averages show the lowest success rate - 24% in both cases. This is primarily conditioned by the fact that common daily trajectory consists of numerous significant and insignificant transitions and thus their average might not be always rational measure. Presence of many insignificant transitions in dataset is also the reason why taking the distance of last transition results in such a low percent of correct predictions. We thought, that success rate could be improved by not considering outliers and thus took the average distance between transitions that fit into the interval between first and third quartiles. We also found the average separately for significant and insignificant transitions. Achieved results are much better, however, still accordingly 9% and 12% worser that the best result. Taking into account that distance prediction is only one step of a prediction, result of more than 50% correct answers is promising.

As for the second component, *Table 5* shows, that the best result is obtained by taking the bearing of the last significant transition. This is driven by the fact that in the majority of cases the last significant transition is a good metric for

Method	Correct	Wrong
Bearing of the last transition	69%	31%
Bearing of the last significant transition	72%	28%
Average bearing of all transitions	65%	35%

Table 5: *Comparison of methods for prediction of bearing to the next geo-location*

showing the overall movement direction. Results achieved by taking the bearing of all last transitions and only significant ones do not differs significantly (3% difference), but as our observations show, it is more sustainable to take only significant transitions into account. Such minor difference is due to the fact, that all significant and insignificant trajectories have their headings and directions with accordance of the main intended course to the destination. As with distance computations, finding the average bearing did not give good outcome - it is 7% worser than the best result.

As a final step, we combined two best approaches and started predicting the next location. For the calculation of next distance and bearing we used methods with the highest prediction success rate (see *Table 4* and *Table 5*) - distance and bearing of the last significant transition. Results can be viewed in *Table 6*.

Method	Correct	Wrong
Bearing and length of last significant transition	46%	54%

Table 6: *Aggregated result for prediction of location of the next geo-location.*

We can observe that next location can be predicted with the probability of 46%. Combining both methods gives smaller success rate than each method separately, because now both parameters must be correct. Thus we can conclude that consideration of only geographic-triggered intentions is not enough for a successful prediction.

5.4.2 Similar starting area and intersection with ending area

This chapter presents the results of the techniques that potentially can improve location prediction success rate by selecting only those trajectories that share similar starting area or intersect with the ending area of the examined trajectory. Reasons why we believe that this approach might improve prediction success rate are described in *Chapter 4.2.1.3* and *Chapter 4.2.1.4*. We followed the same approach regarding prediction validation and used a 200 meters buffer around a geo-location. Results of the experiment are presented in *Table 7*.

Method	**Correct**	**Wrong**
Common approach	46%	54%
Similar starting area	25%	75%
Intersection with ending area	37%	63%

Table 7: *Comparison of different location prediction approaches.*

Common prediction method does not intentionally take similar starting area or intersection with ending area into account, however, the possibility that the most similar daily trajectory will have those properties exist. Unfortunately we can observe a decrease of a predictive power for both experiments - achieved results are accordingly 21% and 9% worser. We believe that there are two main reasons why considered approaches did not improve prediction success rate:

1. Size of the dataset - there were too few daily trajectories that fall under above mentioned conditions. On average in the tests there were only 18 daily trajectories with similar starting area and 20 daily trajectories that intersect with ending area of examined daily trajectory.

2. Temporal aspect - the fact that we are concentrating on daily trajectories and analyzing human activity throughout the day. Having visited the same locations in the morning does not imply that further actions and

visited locations will also coincide. Especially this concern our first test where we compared similar starting areas.

Also, it should be noted that there exist one more important aspect that affects the results of our tests - direction of the movement. This means that we are not considering trajectories with the perpendicular movement direction, even if they coincide in space. Considering the fact, that geographically such trajectories have different common attributes (either different starting areas or their ending areas do not intersect), we do not add those trajectories to the list from where the most similar trajectory is picked from. All this leads to the decrease of the training dataset size and loss of valuable historic, but suitable for analysis, data.

5.4.3 Predicting the type of the next geo-location

As a first step, we figured out which Bayesian order works better with our dataset. We calculated transition probabilities between geo-location types across all daily trajectories and started predicting only the type of the next geo-location. When predicting using Bayesian first order inference we took only current geo-location type into account, while with Bayesian second order inference we used both current and penultimate geo-location types. However, as we mentioned in *Chapter 2.3*, it is not always possible to determine the type of the geo-location ubiquitously as it might be located in the immediate vicinity of several POIs. In such cases there may be three options and we proceeded as follows:

1. Correct geo-location is associated with multiple types - if predicted type match at least one of the correct geo-location types, we mark the prediction as correct.

2. One or more geo-location used for prediction are associated with multiple types - we separately calculate probabilities for all types, find an average and pick the result with the highest probability. If predicted type match the type of the geo-location, we mark the prediction as correct.

3. One or more geo-location used for prediction are associated with multiple types and correct geo-location is associated with multiple types. Prediction phase is the same as in 2: if predicted type match at least one of the correct geo-location types, then we mark the prediction as correct.

After all transition probabilities between geo-location types have been computed, we pick one with the highest probability. Prediction is marked as failed when there is no transitions required for the prediction in the training database. Prediction results can be viewed in *Table 8*.

Method	**Correct**	**Wrong**	**Fail**
Bayesian first order	86%	14%	0%
Bayesian second order	71%	29%	0%

Table 8: *Comparison of first and second order Bayesian inferences for predicting the type of the next geo-location.*

As we can see, first order Bayesian inference gives us better results and is 15% better than Bayesian second order inference. This was a comparatively unexpected result, because theoretically higher order models give better probability values, for instance [21] observes a 20% increase of a predictive power when using $n >= 2$. However, in our case the size of the dataset was a key limiting factor as not all sequence groups of three transitions were presented in sufficient quantity. An example is illustrated below:

P(residential building, shopping | public building) = 0

P(residential building, shopping | food) = 0.4375

P(residential building, shopping | transportation) = 0

P(residential building, shopping | entertainment) = 0

P(residential building, shopping | education) = 0

P(residential building, shopping | shopping) = 0.5625

P(residential building, shopping | residential building) = 0

We can see that only two groups of three transitions were presented in the training database, hence only those will be taken into account when doing a prediction. However, all groups of two transitions were present in the training dataset.

P(shopping | food) = 0.1053

P(shopping | entertainment) = 0.0263

P(shopping | shopping) = 0.1053

P(shopping | public building) = 0.1842

P(shopping | transportation) = 0.0789

P(shopping | education) = 0.0526

P(shopping | residential building) = 0.4474

This means the second order model overfits the data in this particular case, hence result is poorer comparing with 1st order model. Fail rate for both methods is 0%, which means that at least one transition type combination was present in the training database for each transition group.

As a result, we decided to use first order Bayesian inference for predicting the type of the next geo-location. Whole algorithm of next location prediction is as follows:

1. We predict the possible region of the next geo-location. The process of prediction and results are described in *Chapter 4.2.1* and *Chapter 5.4.1*. As a result, we get a region with a diameter of 200 meters.

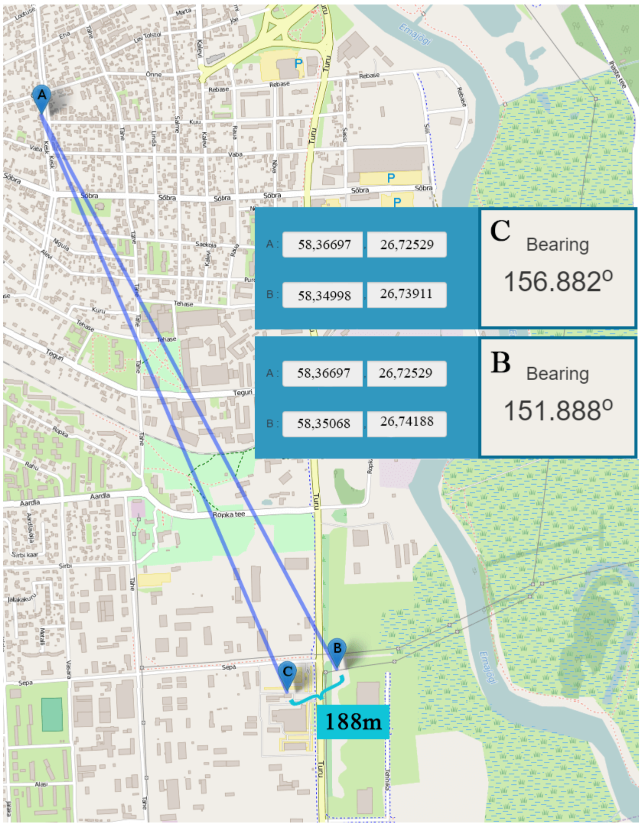
2. We predict the type t of the next geo-location.

3. We check if any of the POIs of type t are located in predicted region. If so, we finish the prediction. Otherwise, we start looking for POIs of type t located close to the predicted region.

We try out different approaches for adjusting predicted region (*Table 9*):

1. Adjustment of distance and/or bearing - we start with correcting distance and bearing and search for such area, which would intersect with established buffer size around the amenity of needed type. We adjust distance and bearing values in both directions and select the least possible deviation values that satisfy the above mentioned condition. Distance was adjusted for up to 500 meters and bearing for up to 30 degrees. In case of t = *residential building* we were looking for a region, where POIs area and region area ratio is less than 30%. The example of bearing adjustment can be viewed on *Figure 14*.

2. We adjust the region in such a way that the nearest POI of type t falls into the region. We choose the region to be our prediction area, in this case adjusting both distance and heading. In case of t = *residential building* we look for a region where POIs area and region area ratio is less than 30%.



C Bearing 156.882°

A : 58,36697 , 26,72529

B : 58,34998 , 26,73911

B Bearing 151.888°

A : 58,36697 , 26,72529

B : 58,35068 , 26,74188

Figure 14: *Bearing adjustment.*

Method	Correct	Wrong
Distance adjustment	48%	52%
Bearing adjustment	50%	50%
Nearest POI	53%	47%

Table 9: *Comparison of location adjustment methods.*

As results indicate, all three methods improve overall prediction success rate. Bearing and distance adjustment results do not differ significantly from each other, however, the fact that bearing adjustment is better, is indeed interesting, especially taking into account the fact that it was observed in *Chapter 5.4.1* that bearing prediction failure rate is lower. Nevertheless, looking for a nearest POI of predicted type overperforms both above mentioned methods and in total allows to make a correct prediction in more than half of the cases. Thus, we decided to use this method in our prediction model. Results can be observed in *Table 10*.

Method	Correct	Wrong
Prediction without taking geo-location type into account	46%	54%
Prediction with taking geo-location type into account	53%	47%

Table 10: *Comparison of the impact of semantic-triggered intentions.*

According to the results, we can conclude that prediction success rate can be improved up to 7% by considering semantic-triggered intentions when building a prediction model.

5.4.4 Taking temporal information into account

As a next step, we started taking temporal information into account when doing a prediction. Foremost, we examined the effect of temporal aspect when predicting the type of the next geo-location. As we figured out in *Chapter 4.2.2*, first order Bayesian inference gives better results with our dataset, thus we concentrated only on it. Results are presented in *Table 11*.

Method	Correct	Wrong
Geo-location type prediction without temporal aspect	86%	14%
Geo-location type prediction with temporal aspect	87%	13%

Table 11: *Impact of temporal aspect when predicting type of the next geo-location.*

The correctness of geo-location type prediction almost did not change, we see a minor improvement of 1%. Our dataset contained very clear and obvious geo-location type patterns (for example, staying at home at night and going to work in the morning) and there were so many of those, that temporal aspect almost did not add any value to the prediction. And if the pattern was out of general track, then in the vast majority of cases it was a unique and no repetitive behaviour. Since the results differ very insignificantly and incorporation of temporal aspect brings additional level of complexity to the prediction algorithm, we decided not to use temporal aspect when predicting the type of the geo-location.

As a next step, we mined association rules from our database. We wanted to discover the relations and regularities between the temporal aspect and geo-location type. For this purpose we used Apriori algorithm with *support=0.02, confidence=0.5* parameters. Detailed results can be observed in *Table 12*.

#	Rule LHS (current location)	Rule LHS (time period)	Rule RHS (next location)	Support	Confidence	Lift
1	public building	daytime	food	0.02	0.82	1.64
2	education	daytime	residential building	0.02	0.78	1.56
3	food	daytime	residential building	0.03	0.76	1.52
4	residential building	evening	residential building	0.08	0.81	1.38
5	residential building	night	residential building	0.19	0.80	1.38
6	shopping	daytime	residential building	0.02	0.60	1.19
7	residential building	daytime	public building	0.03	0.53	1.07

Table 12: *Association rules sorted by lift.*

As results indicate, majority of presented association rules are related to residential buildings. For example, rule 2 indicates that during a daytime person goes home from studies and rule 1 shows that person does a lunch during a working day. Confidence of rules 4 and 5 is very high, which means that rules are correct for 81% of transactions in the evening and night periods, when person is going from residential building to residential building. Taking into account the fact, that usually person is associated only with one residential building, with very high probability we can conclude that person stays on the same place. Moreover, it is logical, that residential areas are more populated at nights and city centre at the daytime.

All rules have *lift > 1*, which shows that RHS and LHS occurrences are dependent on each other, which makes rules potentially useful for predicting next location. We added the support for those rules to our model and rerun the experiment. Results can be viewed in *Table 13*.

Method	Correct	Wrong
Location prediction without semantic and temporal aspect	46%	54%
Location prediction with semantic and without temporal aspect	53%	47%
Location prediction with semantic and temporal aspect	65%	35%

Table 13: *Comparison of the impact of temporal-triggered intentions.*

We can see that results improved by 12% when comparing with the prediction without taking temporal patterns into account. Such an increase happened mainly due to the fact that temporal aspect helped identifying the problem of a person staying on the same place.

6. Discussion and Perspectives

This chapter presents our conclusions and discussions about the topic as well as perspectives for future work, analysis and research opportunities in order to expand the boundaries of the framework to different fields.

6.1 Conclusion

This thesis investigates human mobility behaviour and prediction of the next location. We concentrated on predicting user's daily movements and next significant locations, hence grouped spatio-temporal data into one day time intervals as such separation provides more natural overview of the movement as well as gives a full picture of daily activities. In this work we analyzed real-world GPS data, extracted user trajectories, detected significant geo-locations as well as recognized user activities associated with those geo-locations. We identified three main components that drive people to change their location and proposed a human mobility prediction model that considers them all. The model establishes the relations between geographic, semantic and temporal information captured from human movements.

A systematic evaluation of the model was carried out and, according to the results, model was able to predict a correct location in 65% of cases. Results showed that geographic-triggered intentions cannot solely explain the movement behaviour and be sufficient for a successful location prediction - additional variables such as semantic and temporal aspects should be taken into account. Consideration of semantic links helped to reduce the size of the prediction area, while identification of temporal associations between locations contributed to solving the problem of a person staying in the same region.

6.2 Impact of the dataset size

Our research was conducted on a relatively small dataset collected for the purpose of validation of detection of mobility episodes [10] done by Distributed Systems Group of University of Tartu. Used data can be described as a moderately uniform data not representing high diversity or having many unpredictable cases.

Proposed prediction model is likely to be used on large real-world datasets, which grow at a rapid rate nowadays. Such a great difference between training set sizes can lead to model learning performance being fundamentally different from what we achieved in the research. Furthermore, quality and diversity of examples in training sample also have a direct impact on the performance and efficiency of the model. Our prediction model uses machine learning techniques, however, working with machine learning models that have learning and training phase requires size of training set to be larger. Optimal size of dataset depends on many factors including the complexity of used prediction model, noise ratio as well as quality of the original data. For a better validation of our research, proposed model should be applied to a bigger dataset.

On the whole we consider our dataset to be limiting factor in the research.

6.3 Future work

Developing this subject in the future, the main emphasis should be placed on improving prediction accuracy. There are several directions to be carried out:

- More attention could be paid to the various transportation modes (walking, car, bus, bicycle, etc). Analyzing movement and finding geo-locations, which are specific only to some certain transportation mode can help with further understanding of human mobility patterns.

- The process of finding geo-locations may be improved as well. We were using density-based clustering algorithms to extract significant places, but it is acknowledged, that considered class of algorithms do not perform well when data is not not sampled continuously. This is a very important aspect as GPS signals may be corrupted or be completely missing for a certain period of time and, as a result, extracted geo-locations will get a completely different semantic meaning. Authors of [32] show that interpolation techniques will help to solve the problem and fill in the data gaps.

- Temporal cycles can be analyzed more thoroughly, especially in the applications where movements are linked to daily, weekly or seasonal cycles.
- More emphasis should be put on sociological aspects when doing semantic tagging of geo-locations. The assumption that geo-location can be classified as residential building in case of absence of POIs in close proximity may not hold with non-urban data.
- In order to provide an enhanced positioning output, map matching algorithms can be applied to align inaccurate locational data with road network.

References

1. Agamennoni, G., Nieto, J. and Nebot, E., 2009. Mining GPS data for extracting significant places. In *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on* (pp. 855-862). IEEE.

2. Agrawal, R. and Srikant, R., 1994. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB* (Vol. 1215, pp. 487-499).

3. Andrienko, G., Andrienko, N. and Wrobel, S., 2007. Visual analytics tools for analysis of movement data. *ACM SIGKDD Explorations Newsletter*, 9(2), pp.38-46.

4. Andrienko, G., Andrienko, N., Rinzivillo, S., Nanni, M., Pedreschi, D. and Giannotti, F., 2009. Interactive visual clustering of large collections of trajectories. In *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on* (pp. 3-10). IEEE.

5. Asahara, A., Maruyama, K., Sato, A. and Seto, K., 2011. Pedestrian-movement prediction based on mixed Markov-chain model. In *Proceedings of the 19th ACM SIGSPATIAL international conference on advances in geographic information systems* (pp. 25-33). ACM.

6. Ashbrook, D. and Starner, T., 2002. Learning significant locations and predicting user movement with GPS. In *Wearable Computers, 2002.(ISWC 2002). Proceedings. Sixth International Symposium on* (pp. 101-108). IEEE.

7. Ashbrook, D. and Starner, T., 2003. Using GPS to learn significant locations and predict movement across multiple users. *Personal and Ubiquitous Computing*, 7(5), pp.275-286.

8. Assam, R. and Seidl, T., 2014. Context-based location clustering and prediction using conditional random fields. In *Proceedings of the 13th*

International Conference on Mobile and Ubiquitous Multimedia (pp. 1-10). ACM.

9. Avola, D., Conde, C., de Diego, I.M., Cabello, E., Maghari, A.Y.A., Liao, I.Y., Sharif, M.H.U., Uyaver, S., Sharif, M.H., Marcon, M. and Frigerio, E., Computational Modelling of Objects Represented in Images III Fundamentals, Methods and Applications.

10. Batrashev, O., Hadachi, A., Lind, A. and Vainikko, E, 2015. Mobility Episode Detection from CDR's Data using Switching Kalman Filter. In *Proceedings of the 4th ACM SIGSPATIAL International Workshop on Mobile Geographic Information Systems* (pp. 63-69). ACM.

11. Bettstetter, C., Resta, G. and Santi, P., 2003. The node distribution of the random waypoint mobility model for wireless ad hoc networks. *Mobile Computing, IEEE Transactions on*, 2(3), pp.257-269.

12. Camp, T., Boleng, J. and Davies, V., 2002. A survey of mobility models for ad hoc network research. *Wireless communications and mobile computing*, 2(5), pp.483-502.

13. Castro, P.S., Zhang, D. and Li, S., 2012. Urban traffic modelling and prediction using large scale taxi GPS traces. In *Pervasive Computing* (pp. 57-72). Springer Berlin Heidelberg.

14. Chapman, A.D., 2005. *Principles of data quality.* GBIF.

15. Cho, E., Myers, S.A. and Leskovec, J., 2011. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1082-1090). ACM.

16. Chon, Y., Shin, H., Talipov, E. and Cha, H., 2012. Evaluating mobility models for temporal prediction with high-granularity mobility data. In *Pervasive Computing and Communications (PerCom), 2012 IEEE International Conference on* (pp. 206-212). IEEE.

17. Doss, R.C., Jennings, A. and Shenoy, N., 2004. A review of current mobility prediction techniques for ad hoc networks. In *The Fourth IASTED International Multi-Conference, Banff, Canada* (pp. 536-542).

18. Ester, M., Kriegel, H.P., Sander, J. and Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd* (Vol. 96, No. 34, pp. 226-231).

19. Firouzi, H., Liu, Y. and Sadrpour, A., Mobility Pattern Prediction Using Cell-phone Data logs.

20. Fülöp, P., Szabó, S. and Szálka, T., 2007. Accuracy of random walk and markovian mobility models in location prediction methods. In *Software, Telecommunications and Computer Networks, 2007. SoftCOM 2007. 15th International Conference on* (pp. 1-5). IEEE.

21. Gambs, S., Killijian, M.O. and del Prado Cortez, M.N., 2012. Next place prediction using mobility markov chains. In Proceedings of the First Workshop on Measurement, Privacy, and Mobility (p. 3). ACM.

22. Ge, Y., Liu, Q., Xiong, H., Tuzhilin, A. and Chen, J., 2011. Cost-aware travel tour recommendation. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 983-991). ACM.

23. Giannotti, F., Nanni, M., Pedreschi, D., Pinelli, F., Renso, C., Rinzivillo, S. and Trasarti, R., 2011. Unveiling the complexity of human mobility by querying and mining massive trajectory data. *The VLDB Journal—The International Journal on Very Large Data Bases*, 20(5), pp.695-719.

24. Gonzalez, M.C., Hidalgo, C.A. and Barabasi, A.L., 2008. Understanding individual human mobility patterns. *Nature*, 453(7196), pp.779-782.

25. Hadachi, A., Batrashev, O., Lind, A., Singer, G. and Vainikko, E., 2014. Cell phone subscribers mobility prediction using enhanced Markov Chain algorithm. In *Intelligent Vehicles Symposium Proceedings, 2014 IEEE* (pp. 1049-1054). IEEE.

26. Hwang, J.R., Kang, H.Y. and Li, K.J., 2006. Searching for similar trajectories on road networks using spatio-temporal similarity. In *Advances in Databases and Information Systems* (pp. 282-295). Springer Berlin Heidelberg.

27. Lee, K., Hong, S., Kim, S.J., Rhee, I. and Chong, S., 2008. Demystifying levy walk patterns in human walks. *North Carolina State University, Tech. Rep.*

28. Liao, L., Patterson, D.J., Fox, D. and Kautz, H., 2006. Building personal maps from GPS data. *Annals of the New York Academy of Sciences*, 1093(1), pp.249-265.

29. Mathew, W. and Martins, B., 2012. A comparison of first-and second-order HMMs in the task of predicting the next locations of mobile individuals. In *Proceedings of the First ACM SIGSPATIAL International Workshop on Mobile Geographic Information Systems* (pp. 73-79). ACM.

30. Nanni, M. and Pedreschi, D., 2006. Time-focused clustering of trajectories of moving objects. *Journal of Intelligent Information Systems*, 27(3), pp.267-289.

31. Song, C., Qu, Z., Blumm, N. and Barabási, A.L., 2010. Limits of predictability in human mobility. *Science*, 327(5968), pp.1018-1021.

32. Thierry, B., Chaix, B. and Kestens, Y., 2013. Detecting activity locations from raw GPS data: a novel kernel-based algorithm. *International journal of health geographics*, 12(1), p.1.

33. Timašjov, D., 2014. Evaluating Clustering Techniques [WWW] http://ds.cs.ut.ee/Members/hadachi/dss-fall-2014/Dmitri-Timasjov-final-report.pdf (15.05.2016)

34. Veness, C., Calculate distance, bearing and more between Latitude/Longitude points [WWW] http://www.movable-type.co.uk/scripts/latlong.html (15.05.2016)

35. Vlachos, M., Kollios, G. and Gunopulos, D., 2002. Discovering similar multidimensional trajectories. In *Data Engineering, 2002. Proceedings. 18th International Conference on* (pp. 673-684). IEEE.

36. Wen, L., Shi-xiong, X., Feng, L. and Lei, Z., 2014. Improving location prediction by exploring spatial-temporal-social ties. *Mathematical Problems in Engineering*, 2014.

37. Ying, J.J.C., Lee, W.C. and Tseng, V.S., 2013. Mining geographic-temporal-semantic patterns in trajectories for location prediction. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(1), p.2.

38. Yuan, J., Zheng, Y. and Xie, X., 2012. Discovering regions of different functions in a city using human mobility and POIs. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 186-194). ACM.

39. Zaki, M. J., Parthasarathy, S., Ogihara, M., & Li, W. (1997). New Algorithms for Fast Discovery of Association Rules. In *KDD* (Vol. 97, pp. 283-286).

40. Zheng, Y., Zhang, L., Xie, X. and Ma, W.Y., 2009. Mining interesting locations and travel sequences from GPS trajectories. In *Proceedings of the 18th international conference on World wide web* (pp. 791-800). ACM.

41. Zignani, M. and Gaito, S., 2010. Extracting human mobility patterns from GPS-based traces. In *Wireless Days (WD), 2010 IFIP* (pp. 1-5). IEEE.

Non-exclusive licence to reproduce thesis and make thesis public

I, ... Dmiti Timašjov ....

(author's name)

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

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

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

   .... Human Mobility Mining ....
   .... Using Spatio-Temporal Data ....
   ....,

   (title of thesis)

   supervised by .... Amnir Hadachi, PhD ....

   (supervisor's name)

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, **19.05.2016**