UNIVERSITY OF TARTU
Faculty of Science and Technology
Institute of Mathematics and Statistics
Mathematics and Statistics Curriculum
Speciality Mathematical Statistics

Brigitta Rebane

Detection of meaningful locations from passive mobile positioning data using location profiling

Master's Thesis (30 ECTS)

Supervisor(s): Märt Möls, PhD
Kaisa Vent, MSc

Tartu 2022

Detection of meaningful locations from passive mobile positioning data using location profiling

Abstract:
Mobile positioning data is a promising source for investigating people's activity patterns. People regularly visit locations that have different functions to them. Locations with similar activity patterns can be distinguished from the data based on people's calling activities. The problem with assigning meaning to these locations in the data is limited information about the person and access to ground truth data. The thesis proposes a method to profile locations and assign meanings to differently behaving location groups. In the course of the work, various features are added to the location points by means of which they are classified. Additionally, an expert's opinion was considered to provide input for the classes.

Keywords: mobile positioning, anchor point model, principal component analysis, cluster analysis

CERCS: P160 Statistics, operations research, programming, financial and actuarial mathematics; S230 Social geography

Oluliste asukohapunktide leidmine passiivsetest mobiiliandmetest kasutades asukohapunkti profileerimist

Lühikokkuvõte:
Mobiilpositsioneerimisandmed on paljulubav andmeallikas inimeste aktiivsusmustrite uurimiseks. Inimesed külastavad regulaarselt asukohti, mis täidavad nende elus kindlat funktsiooni. Andmete põhjal on võimalik sarnase aktiivsusmustriga asukohti klassifitseerida. Probleemiks asukohtade funktsiooni määramisel on limiteeritud informatsioon kasutajate kohta ja õigete asukohaklasside siltide puudumine. Antud töös esitatakse asukohtade profileerimise meetodit, mis leiab asukohad, kus inimeste käitumismuster on sarnane. Töö käigus lisatakse asukohapunktidele erinevaid tunnuseid, mille abil neid klassifitseeritakse. Lisaks andmetele kasutatakse asukohtade funktsioonide määramisel ekspertide teadmisi.

Võtmesõnad: mobiilpositsioneerimine, ankurpunktide mudel, peakomponentanalüüs, klasteranalüüs

CERCS: P160 Statistika, operatsioonianalüüs, programmeerimine, finants- ja kindlustusmatemaatika; S230 Sotsiaalne geograafia

Contents

Introduction

New innovative data sources have been introduced to people with the rapid speed of evolving technology. Moreover, national statistical offices have investigated such data sources to add new products to the statistical portfolio and increase the timeliness and accuracy of the gathered statistics while reducing cost and respondent burden. One of such new data sources is mobile positioning data (MPD), which is already used by several governments, such as Estonia and Indonesia, to produce official statistics. For example, MPD is used by Estonian Bank to produce outbound and inbound tourism statistics [Pan22], and in Indonesia, it is used to produce statistical indicators for cross-border tourism [LRS+18].

Mobile positioning data is a collection of location points that enables to estimate of peoples' presence at a location at any given time. Multiple domains can benefit from such knowledge, such as population statistics, mobility studies and the tourism sector. However, using MPD in any field of study requires the algorithmic extraction of the information from the initial dataset.

One of the fundamental algorithms, the anchor model, deals with detecting important locations – places that people visit regularly and play a role in their day-to-day lives. Anchor models primarily focus on home and sometimes also work detection, leaving out other types of locations. This is mainly because extracting the semantics of such locations is not a straightforward task - MPD only records peoples' presence in time and space, and any additional information needs to be mined from the data.

The paper aims to mine the MPD and perform quantitative analysis to detect and assign semantics for meaningful locations by creating interpretable and reproducible clustering of meaningful locations using the approach of profiling. Profiling a location is a way to assign certain distinguishable qualities to the location. So far, meaningful locations have not been detected and classified through profiling. Furthermore, past studies have mainly focused on identifying homes and workplaces, such as in [ASJ+10]. However, the methodology introduced in this paper allows creating of various profiles and classifying locations into these profiles. This approach allows the researcher to specify the type of visitation pattern of a location they wish to explore further.

The first section describes mobile positioning data. A brief overview is given on the types, accessibility and significance of the research. The second section is about the methodology that was used to classify meaningful locations and differences from traditional approaches. The third section thoroughly describes the data and its preparation. It additionally discusses implementing the described method, validation and results on another dataset where the ground truth is known.

1 Mobile positioning data

The data that is used in this thesis is mobile positioning data in Estonia. Hereinafter a short form MPD is used for this kind of data. MPD can be considered big data- the whole data set has hundreds of thousands of users and, therefore, billions of data records.

The mobile positioning data can be challenging to access. Since the field has active competition between operators, the operators conceal their data to avoid leakage of business secrets. Leaks can damage the company's image in public and lose the trust of its clients. Therefore companies have strict regulations on sharing the data [AAR+08].

In addition to operators' regulations, the individuals' mobile positioning data is protected by law. There are two main principles for using MPD: identifiable MPD can be processed if the subscriber has given consent or the law has ordered to perform an official task. Secondly, fully anonymous MPD can be processed and used without restrictions. This requires subscribers to be non-identifiable directly or indirectly [Com14].

This thesis develops the methodology based on the fully anonymous mobile positioning data. All records have anonymised identification codes that cannot be linked to a specific person. Therefore, when analysing the data results, the accuracy cannot be verified because there is no information about the subscriber.

1.1 Types of MPD

Mobile positioning data can be collected with two main methods: active and passive positioning. Active positioning is done by making a specific targeted request to locate the mobile device [Tir14, 5]. A special environment and a permit from the user are required to position the device [AAR+08, 470]. Consequently, enough data for analysis would be extremely expensive to collect, and the data investigated in this thesis is collected by passive positioning.

Passive positioning relies on the fact that historical data of people's call activities are stored in databases by mobile network operators (MNO). Call activity events are called Call Detail Records (CDR). CDR can be an incoming or outgoing call, a short message service (SMS) or a multimedia message service (MMS). [Tir14] In the data used in this thesis, there is no distinction if the activity was originally a call, SMS or MMS. Any activity of sort is called a call activity event.

Some operators provide data detail records (DDR) additionally. In that case, the data set also contains records for every incoming or outgoing web request, i.e. internet traffic between the mobile device, and the network [Tir14, 8]. CDR usually has the following attributes: a timestamp, caller ID, recipient ID, cellular tower code and duration. A cellular tower code is a unique code for each tower used to locate it precisely. Still, the attributes vary by country and MNO [BBG+18, 15], but in order to use MPD for statistical production, at least caller ID, timestamp and location information is required. In this thesis, only timestamp, caller ID and cellular tower code are to be used.

1.1.1 Temporal aspects of MPD

Mobile positioning data is not balanced, the density of the records is influenced by individuals' habits. Therefore, the quality of the data depends on the frequency of phone users' call activities [BBG+18]. There are several solutions for dealing with this aspect. For example, individuals and locations with extremely low activity can be removed. However, this should be done with caution not to remove too much information. Moreover, the low density of data can be algorithmically improved, for example, using interpolation techniques.

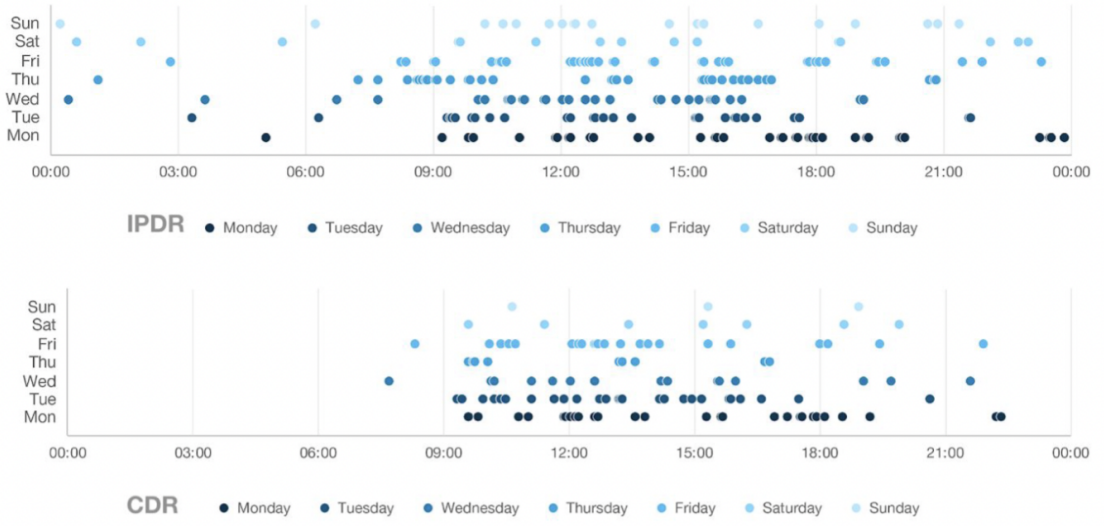

Figure 1. Diurnal distribution of records generated by one subscriber: IPDR (upper) vs CDR (lower). Source: Positium.

The figure 1 illustrates the difference between CDR and IPDR data. IPDR is more evenly distributed during the day because, at night, passive data usage is still generating data. Passive data usage denotes activities that happen in the background and do not require the phone user to be actively engaged with the phone– mobile phone apps refresh themselves automatically (e.g. downloading new emails). On the contrary, CDR captures active engagement from the user. Hence the data is concentrated in the daytime time interval. Therefore, the density of data highly depends on the person's habits as well as whether the operator provides internet usage data.

1.1.2 Spatial aspects of MPD

To record a call activity event, an activity must be made from a mobile phone. Then the antenna ID, to which the mobile device was connected during the initiation of the call,

is saved. Therefore a record of the call activity event consists of the user initiating the activity, event time, and the antenna ID [AAR+08].

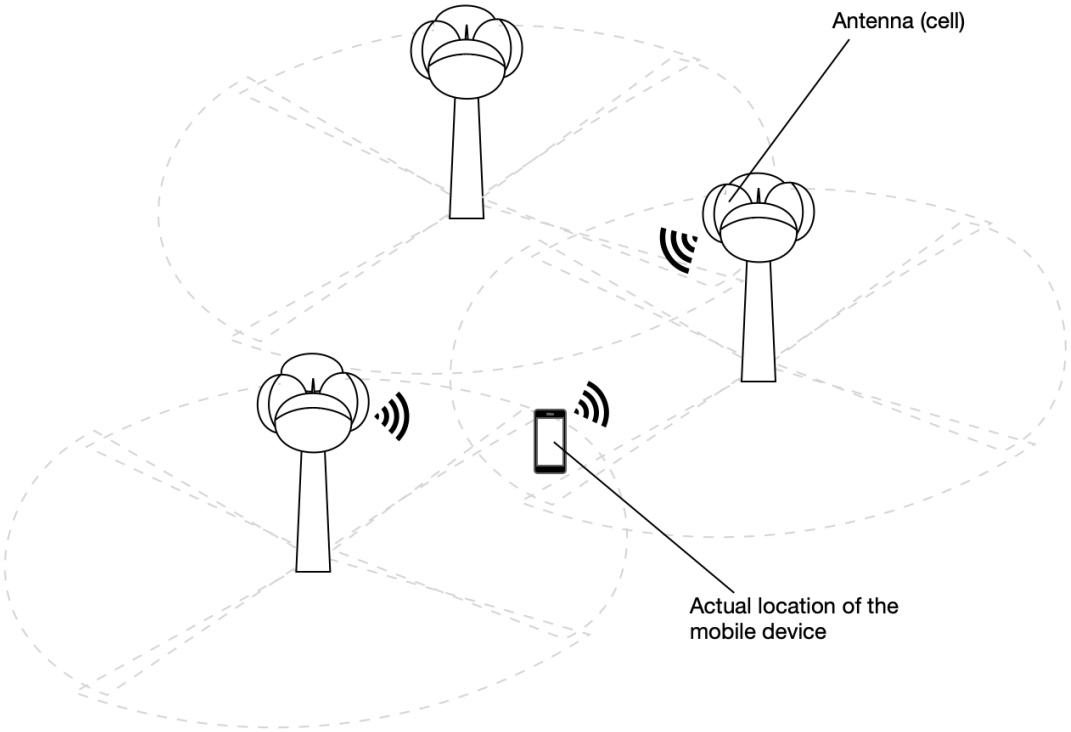

Figure 2. Illustration of antennae coverage areas in relation to a cell signal.

A tower usually has multiple antennae attached to it, which all have their coverage areas; see figure 2. In the records, the identification of the antenna that caught the signal is recorded with its coordinates [AAR+08]. Knowing the antennae's coordinates, the calling records can be generalised to a coverage area accuracy. It means that the mobile device's actual location is not known – it can be anywhere inside the antenna coverage area.

When identifying meaningful locations, overlapping tower reception areas should be taken into consideration. Usually, the mobile device switches to the closest antenna or the one with stronger radio coverage. It might happen that the network is crowded, and the device switches to another neighbouring antenna [AAR+08]. All the neighbouring coverage areas must be considered to obtain all information and get the precise location. Positium has developed a method to calculate theoretical coverage areas and overlaps between them. This thesis relies on Positium's method in the aspect of overlapping coverage areas.

Therefore another generalisation in this thesis is made for identifying important places. The cell towers are grouped based on their overlapping reception area, and MPD records have the spatial accuracy of the overlapping area union.

In remote areas, the cell towers are distributed sparsely because the area is less inhabited, and therefore the network load is low. The accuracy of mobile positioning can vary from 1.5 to 20 km in rural areas. This can lead to large errors in meaningful location detection. However, in urban areas, the cell towers are more densely distributed - in larger Estonia's cities, the accuracy of positioning can be 100 - 1000 meters, and in suburban areas, 450 m to 2 km [AAR+08].

1.2 Anchor model

To better understand the work, it is essential to clarify the key elements of Positium's current anchor model. The anchor model is part of Positium's core methodology that identifies meaningful locations (anchor points) based on MPD using a specific chain of rules. The current model identifies home and work locations. All other detected locations are labelled as "other regularly visited places". Anchor points are calculated with the accuracy of a month. The research done in this thesis provides additional semantics to the anchor model, allowing to better classify "other regularly visited places" and providing input to enhance the detection of meaningful locations.

The first step of the model is considering the theoretical calculated coverage areas. For each person, the antennae that have overlapping coverage are grouped. This includes multiple antennae that are attached to one tower. It can be assumed that the overlapping antennae serve the same location.

The next step of the model is determining if a location point is regular or random. A point is regular if a respondent has made calls on at least two separate days in the month. Next, for each person, the number of calls is calculated. If the person were active on less than seven days in the month, they would not participate in further analysis. Additionally, people with over 5000 calls a month are removed because they are potentially not regular users. This step is done during the importing of the data for the anchor model.

The locations for each person that have the highest number of call days (three highest-ranking locations) are allocated for home and work-time anchor point detection. Other regular points are considered "other regular anchor points". Additional semantics are not analysed through the "other regular anchor points" in the current model.

([ASJ+10])

2 Methodology

The next chapter describes the location classification problem more in-depth and the methodology used to solve it. An overview of traditional approaches is given in addition to proposed modifications. The methodology concentrates on the principal component analysis and classification methods.

2.1 Principal component analysis

Principal component analysis (PCA) is a technique of multivariate data analysis. The main idea is to transform multiple possibly correlated variables into a smaller subset of variables with retaining the most variation in the data set. The new variables are called principal components [MST+17]. Such variables are essentially linear combinations of the columns of a matrix under consideration [JC16]. The steps of finding suitable coefficients of the combinations are discussed in the following chapter.

The first step of the analysis, as in all methods, is to process and clean the data. It is important that all variables describe the measured object. Identification of an object or a group must not be included in computing the principal components. Additionally, all the variables need to be analysed. Let the data be a matrix X with dimensions $n \times p$ and $\overrightarrow{\overline{X}}$ it's mean values' vector.

In the data, there might be variables of different measurement units, or it is not clear whether they have different scales. It might be needed to standardise the variables. Standardisation means both centring and dividing each variable by its standard deviation. It must be noted that centring the data is a common approach to defining the PCs. If there occurs a difference between the scales of some variables, then they are also divided by the standard deviation [JC16]. The new values z_{ij} replace the original data values:

$$z_{ij} = \frac{x_{ij} - \overline{x}_j}{s_j}, \tag{1}$$

where $\overline{x}_j \; j = 1 \dots p$ is the mean value of column vector $\vec{X}_j$ and $s_j \; j = 1 \dots p$ is the standard deviation of column vector $\vec{X}_j$ of the matrix X. In further actions the standardised data matrix Z is used in place of the original data matrix X.

Next step is to calculate the covariance matrix of Z [MST+17]. It can be noted, that with two standardised random variables A, B:

$$\text{cor}(A, B) = \frac{\text{cov}(A, B)}{\sqrt{\text{var}(A)}\sqrt{\text{var}(B)}} \implies$$

$$\implies \text{cov}(A, B) = \text{cor}(A, B)\sqrt{\text{var}(A)}\sqrt{\text{var}(B)} = \text{cor}(A, B).$$

The last equation holds because after standardisation, the variable's variance, therefore standard deviation, is 1. This also supports the fact that PCA on standardised data is called a correlation matrix PCA [JC16]. Let's denote R as a correlation matrix of Z.

Next step is to find eigenvalues and eigenvectors of the matrix R [MST+17]. Denote $\lambda = (\lambda_1, \ldots, \lambda_p)$ as the eigenvalues in descending order and A as an eigenvector matrix, where columns $\vec{a_1} \ldots \vec{a_p}$ are the corresponding eigenvectors of R.

The eigenvectors are the coefficients of linear combinations of Z that determine the principal components. The jth PC score of a row $Z = z$ is $PC_{jz} = a_j^T z = (a_{j1} \ldots a_{jp}) \cdot (z_1 \ldots z_p)^T$ [JWHT13]. Therefore it is a linear combination, where each variable of the row $Z = z$ is multiplied by certain element from the eigenvector a_j.

Based on the properties of matrix calculation, the result can be generalised to a calculation of the jth PC score of all rows. This follows as: $PC_j = Z \cdot a_j$. The result PC_j is a $n \times 1$ matrix, where each row is a jth PC score of each row of Z.

Consequently, all the principal components' scores make a $n \times p$ matrix $PC := Z \cdot A$, where each column is called a principal component.

2.1.1 Reversing components

The principal components are linear combinations and cannot be interpreted in the original data scales. For visual aid in certain situations, it is essential to compare obtained components to original data. This subsection shows how the components are reverted to original data scales.

Firstly, the number of components k used must be chosen. It is the dimension into which the data is desired to be converted. It is important to note that if $k = p$, then all components are used in reversion and the original data set is reconstructed precisely. However, the purpose of PCA is to reduce dimensionality. Therefore it is meaningful to choose $k < p$ in a way that the k components describe the data set well enough.

Furthermore, using matrices' properties and multiplying the equation by A^T from the right, the following result can be obtained:

$$PC = Z \cdot A \implies PC \cdot A^T = Z \cdot A \cdot A^T = \hat{Z} \implies \hat{Z} = \hat{PC} \cdot \hat{A}^T,$$

where $\hat{PC}$ is an $n \times k$ matrix with a chosen number of k components. Analogously $\hat{A}$ is $n \times k$, therefore being a matrix of k eigenvectors as columns. It is important to note that if the chosen number of components k is equal to the number of columns p, then $AA^T = I_p$, by the property of eigenvectors. Therefore the $ZAA^T = ZI_p = Z = \hat{Z}$ and the original data set is precisely reverted. However, if $k < p$, then the reverted matrix $\hat{Z}$ is approximate to the original Z. The fewer components k used, the more approximate $\hat{Z}$ is.

Furthermore, to achieve the original scales, the reverse of standardisation must be additionally performed. Firstly, all values must be multiplied by the standard deviation

of the original variable. Secondly, the mean of the original variable must be added to obtain the original scale. The reverted $\hat{x}_{ij}$ are obtained by reversing the standardisation formula 1:

$$\hat{x}_{ij} = \hat{z}_{ij} \cdot s_j + \overline{x}_j. \tag{2}$$

2.1.2 Standardisation

Showing that standardisation is necessary for the data in this thesis. Let's denote B a subset of the original data set.

Table 1. Subset B

(a) Original scales.

	Monday	***Workhours***	***N***	***Weeks_of_all***
1	0.143	1.0	7	0.8
2	0.115	0.485	165	1.0
3	0.263	0	19	0.8
4	0.161	0.465	409	1
5	0.125	0.375	8	0.8

(b) Standardised scales.

	Monday	***Workhours***	***N***	***Weeks_of_all***
1	−0.313	1.497	−0.659	−0.73
2	−0.779	0.056	0.249	1.095
3	1.708	−1.301	−0.590	−0.730
4	−0.002	−0.001	1.652	1.095
5	−0.614	−0.251	−0.653	−0.730

Denoting the covariance matrix $cov(B) = S$. It appears as:

$$S = \begin{pmatrix} 0.004 & -0.014 & -1.549 & -0.002 \\ -0.014 & 0.127 & -0.662 & 0.001 \\ -1.549 & -0.662 & 30261.8 & 16.54 \\ -0.002 & 0.001 & 16.54 & 0.012 \end{pmatrix}, S_{sc} = \begin{pmatrix} 1.000 & -0.645 & -0.150 & -0.357 \\ -0.645 & 1.000 & -0.011 & 0.025 \\ -0.150 & -0.011 & 1.000 & 0.868 \\ -0.357 & 0.025 & 0.868 & 1.000 \end{pmatrix}$$

It is evident that the big variance in variable N affects covariance matrix S. The variance of N is over 30,000, whereas most of the other values are under 1. PCA, by nature, is very dependent on the covariance matrix since the eigenvectors of the matrix are the loadings of the principal components.

Calculate the eigenvectors and denote the matrix of eigenvectors A:

$$A = \begin{pmatrix} 0.000 & -0.109 & -0.566 & 0.817 \\ 0.000 & 0.994 & -0.072 & 0.083 \\ -1.000 & 0.000 & 0.000 & 0.000 \\ -0.001 & 0.012 & 0.821 & 0.571 \end{pmatrix}, A_{sc} = \begin{pmatrix} 0.460 & -0.522 & 0.648 & 0.310 \\ -0.287 & 0.673 & 0.654 & 0.193 \\ -0.566 & -0.416 & 0.360 & -0.614 \\ -0.621 & -0.319 & -0.151 & 0.700 \end{pmatrix}$$

The loadings matrix shows that the original third column N overtakes the first component completely, as the third element in the first row is -1.0 and others 0 or near 0. As stated in the previous subsection, an original data set can be reconstructed using the principal components. Showing what happens when using $k = 1$ principal components to reconstruct the data set with original scales and standardised scales. Selected components $k = 1$ is relevant in this thesis, which is the reason for not selecting more.

Table 2. Reconstructed subset B

(a) Original scales.

	Monday	***Workhours***	***N***	***Weeks_of_all***
1	0.003	0.010	7.004	0.022
2	0.078	0.237	165.002	0.509
3	0.009	0.027	19.002	0.059
4	0.193	0.588	409.999	1.261
5	0.004	0.012	8.003	0.025

(b) Standardised scales.

	Monday	***Workhours***	***N***	***Weeks_of_all***
1	0.168	0.439	96.652	0.863
2	0.129	0.588	239.388	0.961
3	0.215	0.265	−70.164	0.748
4	0.117	0.631	280.791	0.990
5	0.178	0.402	61.333	0.838

As expected, it appears from table 2a, that the reconstruction is not very accurate. However, the first principal component without standardising claimed to describe nearly 100% of the total variation. The column N is reconstructed almost ideally, but other columns do not correspond to the original data. The values are, in most cases, too small, which was expected due to the first eigenvector.

Meanwhile, the PCA on standardised data claimed to explain 52% of total variation with the first component. It appears from table 2b that variables 'Monday', 'Workhours', 'Weeks_of_all' have been reconstructed fairly similar to the original values. However, the variable 'N' does not correspond to the original values. The values still exhibit a similar pattern. For example, the first and fifth values are still the smallest in the column; the second and fourth rows have the highest value of all, as in the original data.

2.2 Classifying method

One aspect of MPD is not knowing the ground truth. There are the calling activities without any information about the person. It is not known where the person's actual home or workplace is, how they spend their free time or their work type. Therefore, to distinguish meaningful locations, similarly behaving locations must be clustered while maintaining the meaning of each cluster.

Traditional classification algorithms need the correct class to train the classification process. However, the methods to mine information from MPD must be conducted so that there is no need for true classes. This thesis uses the help of an expert in the field to classify objects. The expert describes an input for the classification in place of the ground truth. However, the expert's assessment needs adjusting to the data set. This thesis proposes a method based on elements of cluster analysis to adjust the expert's constructed input.

2.2.1 Base algorithm

The methodology of grouping objects under consideration into types of meaningful locations is based on the idea of the k-medoids algorithm. The algorithm is introduced in more detail in this chapter. Hereinafter the changes in some metrics are outlined and described.

K-medoids is a spatial clustering method. The process is to group objects into clusters, so the clusters' objects are highly similar to one another but dissimilar to other clusters' objects. The number of desired final clusters k is set before the beginning of the algorithm [SP17].

Moreover, it is an iterative method. Firstly, random k objects are selected as initial medoids. Each point is then clustered with its closest medoid with any distance metric. Since it is an iterative method, the method iterates through all clusters and tries to replace each medoid with a more suitable one. During each iteration, the most centrally located object is used as a new cluster centre. The algorithm is terminated if no better replacement is available for each medoid. [HKT01].

K-means is a similar method to k-medoid. The difference is that k-means clusters' centres are calculated means, not actual data objects. The centres are called centroids. Secondly, in k-means, generally, Euclidean distance is used, whereas, in k-medoid, any dissimilarity measure can be used [Ize13]. Therefore medoid is a cluster centre that is an actual object, and a centroid is a calculated centre.

The proposed method in this thesis uses components from both k-means and k-medoid methods. Firstly, the expert's constructed objects are the initial centroids. There are several reasons for this. Firstly, experts in the field are familiar with the nature of the data and are aware of the human geography processes. Secondly, this makes the method reproducible. In the future new centroids can be added to acquire more detailed

information. Concerning reproducibility, this also ensures labelled interpretable clusters each time. Using random initial medoids might lead to similar results, but the groups cannot be automatically labelled and are in a different order.

2.2.2 Distance measure

As stated in a previous chapter, MPD data is not balanced. The density of data varies plenty by individual and cell. Therefore, the distance measure should consider measuring a similar shape of behaviour instead of emphasising a small Euclidean distance.

A tool to measure the proximity of one item to another is called a dissimilarity. Dissimilarities usually satisfy three properties. Proof of the properties of the chosen measure is shown below. When additionally the triangle inequality holds, the dissimilarity measure is also a metric [Ize13].

The aim is to measure similarities between objects that can be far apart but behave similarly. In other words, their features are highly correlated, but the observed values might be far apart in terms of Euclidean distance. Therefore the thesis proposes a correlation-based dissimilarity measure that ensures the right properties. It focuses on the activity pattern of the observations rather than the magnitudes [JWHT13].

Given a series of n measurements of the pair (X_i, Y_i) indexed by $i = 1 \dots n$, the sample correlation coefficient, is defined by a formula [Ros10]

$$\rho_{i,j} = \text{cor}(X_i, Y_i) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}. \quad (3)$$

Generally, it is said that the pair is highly correlated when the absolute value of $\rho_{i,j}$ is close to 1 [Ros10]. Moreover, the pair is positively correlated when $\rho_{i,j} > 0$ and negatively when $\rho_{i,j} < 0$. Thus in this thesis, a measure $1 - \rho_{i,j}$ is taken as a measure of dissimilarity to find objects that are close to each other. The dissimilarity measure has been used for similar purposes in other papers as well, such as in [Ize13].

Next, for defitinion purposes it can be checked, whether the dissimilarity measure is a metric. Let $X = \mathbb{R}^n = \{x : x = (x_1, \dots, x_n), \xi_i \in \mathbb{R}\}$. Defining a distance between elements $x = (x_1, \dots, x_n) \in X$ and $y = (y_1, \dots, y_n) \in X$ as

$$d(x, y) = 1 - \rho_{x,y} = 1 - \frac{\sum_{i=1}^{n}(\xi_i - \overline{x})(\nu_i - \overline{y})}{s_x s_y}. \quad (4)$$

Proof of three properties on the chosen measure.

First axiom: $d(x, x) = 0$:

$$d(x, x) = 1 - \rho_{x,x} = 1 - 1 = 0.$$

Second axiom: $d(x, y) \geq 0$:

$$d(x, y) = 1 - \rho_{x,y} \geq 0 \Leftrightarrow 1 \geq \rho_{x,y}.$$

inequality holds because the correlation between any variables is always between $[-1, 1]$. Therefore, shifting the interval, measure $d(x, y) = 1 - \rho_{x,y}$ falls into an interval $[0, 2]$, which is always greater or equal than 0.

Third axiom: $d(x, y) = d(y, x)$:

$$\begin{aligned} d(x, y) = 1 - \rho_{x,y} &= 1 - \frac{\sum_{i=1}^{n}(\xi_i - \overline{x})(\nu_i - \overline{y})}{s_x s_y} \\ &= 1 - \frac{\sum_{i=1}^{n}(\nu_i - \overline{y})(\xi_i - \overline{x})}{s_y s_x} = 1 - \rho_{y,x} = d(y, x). \end{aligned}$$

Consequently, d is a dissimilarity measure satisfying the three axioms. The measure is a metric when it satisfies the triangle inequality. Showing that, in this case, the triangle inequality does not hold.

Triangle inequality: $d(x, y) \leq d(x, z) + d(z, y)$:

$$d(x, z) + d(z, y) = 1 - \rho_{x,z} + 1 - \rho_{z,y} = 2 - \rho_{x,z} - \rho_{z,y} \not\geq d(x, y).$$

The last inequality does not always hold. Consider two independent random variables X_1, X_2 with equal variances $\text{var}(X_1) = \text{var}(X_2)$. Define new variables $Y_1 = X_1$, $Y_2 = X_2, Y_3 = X_1 + X_2$. Now notice, that $\text{cor}(Y_1, Y_2) = \frac{\text{cov}(Y_1, Y_2)}{\sqrt{\text{var}(Y_1)\text{var}(Y_2)}} = 0$ and

$$\begin{aligned} \text{cor}(Y_1, Y_3) &= \frac{\text{cov}(Y_1, Y_3)}{\sqrt{\text{var}(Y_1)\text{var}(Y_3)}} = \frac{\text{cov}(X_1, X_1 + X2)}{\sqrt{\text{var}(X_1)\text{var}(X_1 + X2)}} = \\ &= \frac{\text{cov}(X_1, X_1) + \text{cov}(X_1, X_2)}{\sqrt{\text{var}(X_1)[\text{var}(X_1) + \text{var}(X2)]}} = \frac{\text{cov}(X_1, X_1)}{\sqrt{\text{var}(X_1)\text{var}(X_1) + \text{var}(X_1)\text{var}(X2)}} = \\ &= \frac{\text{var}(X_1)}{\sqrt{2[\text{var}(X_1)]^2}} = \frac{1}{\sqrt{2}} \approx 0.71 \end{aligned}$$

.

Therefore the correlation-based distance between Y_1 and Y_2 is $d(Y_1, Y_2) = 1 - 0 = 1$. Between variables Y_1 and Y_3 is $d(Y_1, Y_3) = 1 - 0.71 = 0.29$, analogously $d(Y_2, Y_3) = 1 - 0.71 = 0.29$. Therefore $1 = d(Y_1, Y_2) \not\leq d(Y_1, Y_3) + d(Y_3, Y_2) = 2 \cdot 0.29 = 0.58$. Consequently d is not a metric.

The suitability of the correlation-based distance is illustrated with an following example. The same example describes why the traditional Euclidean distance does not always measure the desired aspect in this case. Let's consider the activity patterns of three locations.

In the figure 3 there are three different locations that have measured activity levels in 4 variables $V1, \ldots V4$. From the figure, location 1 has zero activity in variable $V1$, a peak in $V2$, a decrease to $V3$ and another peak in $V4$. Location 2 shows a similar pattern with low activity in variables $V1$ and $V3$ but higher values in $V2$ and $V4$. The third location with blue peaked in variable $V1$ and decreased activity in values $V2$ to $V4$.

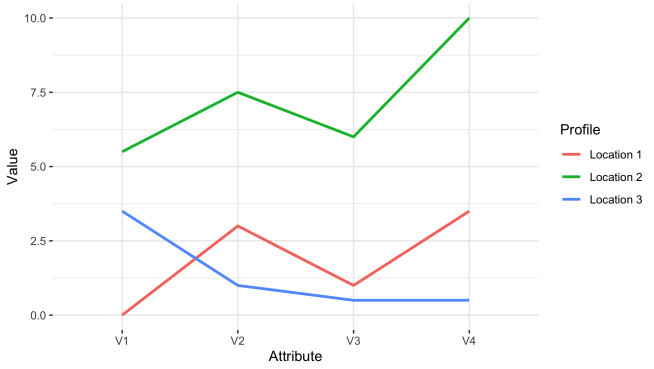

Figure 3. Example profiles

	Location 1	Location 2	Location 3
Location 1	0	0.09	1.71
Location 2	0.09	0	1.59
Location 3	1.59	1.71	0

Table 3. Correlation distances.

	Location 1	Location 2	Location 3
Location 1	0	10.85	5.05
Location 2	10.85	0	12.91
Location 3	5.05	12.91	0

Table 4. Euclidean distances.

b

It is evident that location 1 and location 2 have similar activity patterns; their profiles are parallel in the plot. However, the Euclidean distance between them is 10.85 (table 4), which is twice as big as the Euclidean distance between location 1 and location 3. The distance between location 1 and location 3 is 5.05, the smallest distance between any of the 3 location points. Therefore, according to Euclidean distance in this case, location 1 and location 2 are further apart than location 1 and location 3.

Since the idea of the method proposed in the thesis is to measure similar activity patterns, the correlation-based measure is more suitable. Recall that correlation distance near 0 shows very high similarity, and near 2 shows a high dissimilarity. From table 3 the correlation-based distance between locations 1 and 2 is 0.09, which indicates that the locations behave very similarly, even though location 2 has been more active. However, the correlation-based distance between locations 1 and 3 is 1.59, which shows very high dissimilarity.

The example can be generalised to multiple location points with numerous attributes describing the activity pattern. Some locations might be more frequently visited, but the activity level will not affect classification results with this method. Therefore, when distinguishing similarly behaving location points, it is more meaningful to measure the similarity of the patterns with a correlation-based measure.

2.2.3 PC1 as a centroid

K-medoids algorithm chooses a new centre in each iteration by measuring the distance between each data point and choosing the point most in the centre to be the next medoid, as stated previously. An alternative way of choosing the centroid is proposed as follows.

K-means uses the calculated centre of the data as a centroid. The centroid is calculated, so the mean distance between data points and the centroid is minimised. In the case of Euclidean distance, the arithmetic mean minimises the mean distance. Arithmetic mean might not be optimal for calculating the centroid when using a different distance measure. When using correlation-based distance, the first principal component can be used as the centroid instead of the arithmetic mean. The first principal component maximises the mean covariation of the component to each variable. When the data is standardised, the first principal component maximises the mean correlation to each variable. In this thesis, the data matrix must be transposed because the aim is to maximise the correlation between the principal components and location profiles, not variables. Therefore, each location is considered as a variable and each variable as a data object for the purposes of PCA in this thesis.

In this thesis, the PCA is performed on each subset. Hence, after dividing the data points into initial classes, the principal components are calculated for each cluster of points. The first component for each class is chosen as the new centroid (anchor profile) of the cluster. Therefore each centroid describes the k-th class of points as a linear combination of the points included. With the PC1 as a centroid, new distances are calculated from each new centroid to each data point. The points are re-divided into classes by the closeness to each centroid as in the k-medoids method. The action is repeated for desired number of iterations, for example until consecutive centroids do not differ from each other.

3 Practical part

The purpose of the practical part is to describe implementing the discussed method for meaningful location classification. Firstly, the data is shown, how the sample is taken, and aggregations are made. Next, the creation and modification of anchor profiles are described in detail. Along with that, how to use the profiles to find similarly behaving locations. Lastly, there is a validation of the method and conclusions.

3.1 Assumptions and subsets

MPD is considered big data and working with all the data for method development is computationally not reasonable. Therefore a random sample of ten thousand subscribers was made to investigate the patterns in the data. Additionally, some assumptions must be made due to human mobility.

The anchor model detects meaningful locations within the time of a month and again for each month. In this thesis, the meaningful locations are also looked at in one month period to avoid possible relocation. Therefore the sample of the selected subscriber's one-month data is taken for investigation. The chosen month is January of 2019.

The accuracy of the data records is defined by the extent of the antenna coverage area. Therefore the antennae id-s are grouped in regards to the output of the anchor model, which of them have overlapping reception areas. This is done during data processing prior anchor model as described in chapter 1.2. Therefore this work does not concern single antennae but works with the set of neighbouring antennae. Each antenna group, now called a location, can be described by the call activities undertaken using antennas in this location.

Moreover, a constraint is made. Some subscribers have too little activity to investigate since a subset of one month was taken. Therefore, the final subset contains subscribers that have been active in at least five different locations in January of 2019.

In the final data set, there are 39,539 unique subscriber and cell id pairs and a total of 3,177 unique subscribers in January 2019, as seen in the table 8. It would suggest that, on average, each subscriber has 6-7 cells they constantly use in a month. From the plot 12, most people have 5-12 unique cells used in the time period. There are 18 people out of 3,177 who have calling activities through more than 40 locations.

From the plot 11 it appears that in January of 2019, mostly each person makes a calling activity through one cell 3-4 times. However, most records per unique user and cell combinations are between 3 and 20. There are also people who have made hundreds of calling activities in the same certain areas.

In conclusion, firstly, a sample of 10,00 random people was compiled, with restrictions to all data records being from one month, January 2019. Secondly, the locations of each row are grouped into groups of neighbouring areas. Lastly, because of the lack of balance in the data, the individuals counted more than five different cells were extracted

to perform analysis. Consequently, in the ideal scenario, all different anchor profiles can be assigned to each individual.

The final processed data has 2,453,765 rows, where each row is one specific calling activity. There is an ID of the entry, the ID of the subscriber, the ID of the cell and lastly, an exact timestamp of the activity as seen in table 7. As mentioned above, the time interval was from January 1st to January 31st in 2019.

3.2 Preliminary work

The next step was to create multiple new variables to describe the behaviour of each person's every location in the time period. This thesis does not consider the spatial behaviour of the cells, apart from the grouping of the overlapping areas of cells. Therefore, new variables are solely created based on the timestamp attribute each record has. The new variables were created according to different units of time, such as time intervals and weekdays. Additionally, all entries were grouped by unique user and cell combinations which were aggregated into new attributes.

The first group consists of the most general new attributes- 'days of all' and 'weeks of all'. 'Days of all' considers the percentage of all days in January 2019 the location was used to make calls. Based on the table 8, person A used cell two about 16.13% of the days in January, which makes five unique days. On the other hand, the individual has a calling activity on 80% of all weeks of January. Consequently, the cell is used most probably once a week, and with great probability, this is a constantly visited location, for example, a store, free-time activity, family etc.

The second group consists of attributes more precise. The variables describe the distribution of calling activities by each weekday of the month. All the variables from 'mon_of_all', which describes the percentage of calls in the week falling onto a Monday, to 'sun_of_all', which describes the percentage of calls in the week falling onto a Sunday, sum up to 1. Based on table 9 for the previous example, the person used the cell mostly on Fridays, a few times at the start of the week and not once at the weekend.

It was considered to create variables for each weekday in a way that they describe the usage percentage of the specific weekday. For example, a value for Monday of 0.8 would mean that in January 2019, the person used the cell on 80% of all Mondays. However, with the consideration of an expert's opinion, it was decided that this poses certain problems and, in some cases, loss of information. Consequently, the analysis remained with the distribution across all weekdays.

The third group is the most precise group of new variables. They describe the distribution of calling activities at 4-hour intervals. The variables from '00_04', which describes the percentage of calls in the day falling into the specific time period, to '20_24' similarly sum up to 1. Based on table 10 and the same example, the person clearly uses the cell mostly in the afternoon and daytime.

The fourth group consists of three variable pairs. The first pair is the distribution of calling activities between working days and the weekend. The two variables sum up to 1 and describe the proportions of the calling activities between weekdays and weekends. This proportion overlaps with the weekday distribution variables but gives a general direction. For example, if the exact distribution of weekdays is not known, but it is known that the variable should be close to zero at the weekends. Based on the table 11, the same person mentioned before has conducted all the calling activities on working days.

The second pair describes the distribution of calling activities for each person's every cell between working and non-working hours. In this thesis, the time interval 08:30-17:30 is considered to be working hours. The interval distinguishes the people who work during business hours based on the time usage survey data from Statistics Estonia [Est10].

The last pair consists of daytime and nighttime distribution of calling activities. Nighttime is defined to be 23:00-07:00 in accordance with the expert's opinion. A high nighttime value describes a location with a large proportion of calls made in the nighttime. Together with other variables, it could describe a location where the person spends their nights or can call a primary home in terms of meaningful locations. However, this thesis uses CDR data which might affect the results. Since people are supposed to sleep in their primary homes, they do not use their mobiles at the nighttime. Therefore, for validation purposes, DDR data could deliver more accurate results in home detection.

3.3 Anchor profiles

The main idea is to create anchor profiles that behave in a certain way. The profiles are created by experts and are used as the initial centroid points. Next, the profiles are compared to the data, and the closest data points to each cell are filtered to create subsets. Therefore each subset contains points that behave in a desired way or similarly. Each subset is further processed with PCA and fine-tuned according to data.

There are initially five different anchor profiles: Work-time, Nighttime work, Home, Secondary home, and Evening free time (table 12, figure 4). Work-time anchor profile corresponds to a traditional Monday to Friday business hours location. Nighttime work is an odd hours location, with activity mostly at night. Home is a location where the person usually spends the night and often visits. A secondary home is a location for weekend activities. This might be a summer house, family, or a constantly visited hotel. Evening free time is a general location with the most activity after business hours. This might be a constantly visited store, sporting place, friends, family, or restaurant.

By the table 12, primary home cells are used on the majority of days of the month, 73% of the days, and every week. Weekday distribution is uniform because there should not be a clear distinction between activities on different weekdays in activity at home. All other variables have values analogously under the assumption that a person sleeps at

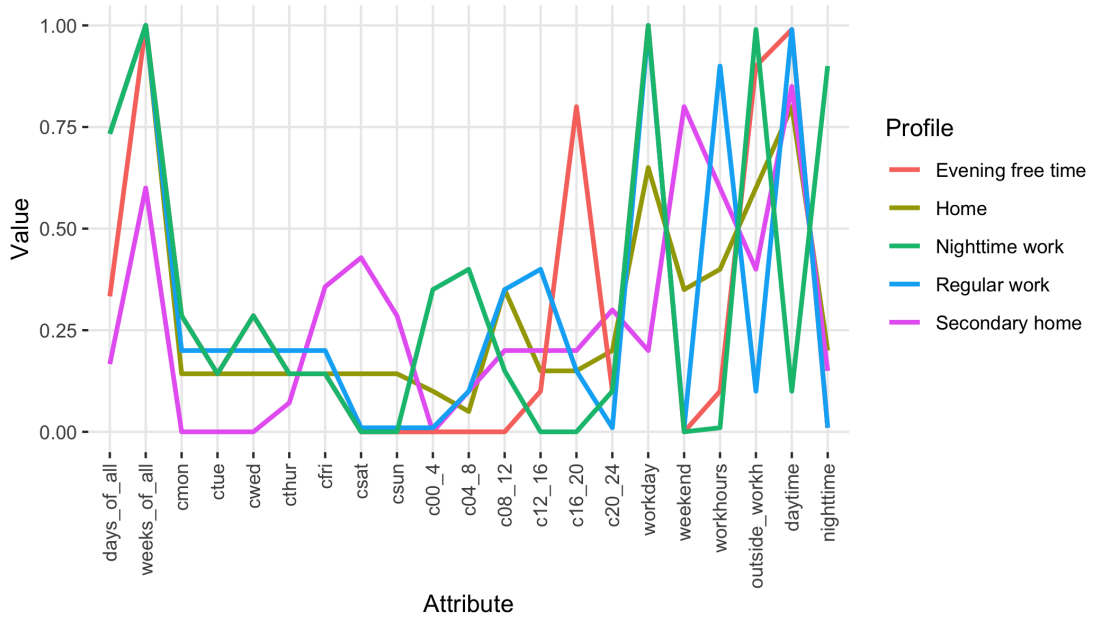

Figure 4. Anchor profiles constructed by an expert.

night and that there should be no significant distinction between some hours or days for recognising home.

Evening free time, regular work and nighttime work all have almost no activity at the weekends. Regular work should have a relatively uniform activity distribution throughout working days; nighttime work and evening free time might vary. Firstly, the free time activities probably fall into a couple of specific days, e.g.Tuesdays and Thursdays, and nighttime workers work on a schedule. Secondary home, on the contrary, has the most activity on the weekends and the day before it.

Four-hour distribution for regular work is focused on the business hours 08:00-16:00 with permission of a small deviation around the business hours. Evening free time activity has most of the activity after business hours during 16:00-20:00 because it is considered mainly an after-work activity location. The secondary home has an overall fairly uniform distribution of calling activities with increased activity in the evening. The nighttime working location has activity mostly at nighttime 00:00-08:00 and most calls in non-business hours and predefined nighttime.

3.3.1 Initial anchor classes

The anchor profiles based on experts' opinions are specific descriptions of a location's behaviour. Some individuals are less active at locations, and some are more active. Their

activity is different, but the overall behaviour might be similar. This subsection describes the classification of locations that behave similarly to an expert's defined profile.

For every anchor profile, the most similarly behaving cells can be found using the distance measure $d(x, y) = 1 - \rho_{x,y}$. Firstly, a distance matrix D is calculated, where the columns correspond to each anchor profile's distance from each location point. Next, for each anchor profile, a subset of cells is chosen to be determined as the type of the profile.

The method to choose a threshold for a subset is proposed as half of the minimum distance to another centroid, using the predefined distance d again. In short, a distance matrix $D_{profiles}$ is calculated where elements are distances between the anchor profiles. For each anchor profile, the shortest distance to another is chosen and halved. Therefore, when performing initial subsetting, there are no overlapping clusters. When two centroids are close to each other, the radii, from which to subset cells, are small and analogously large when the centroid is further from others.

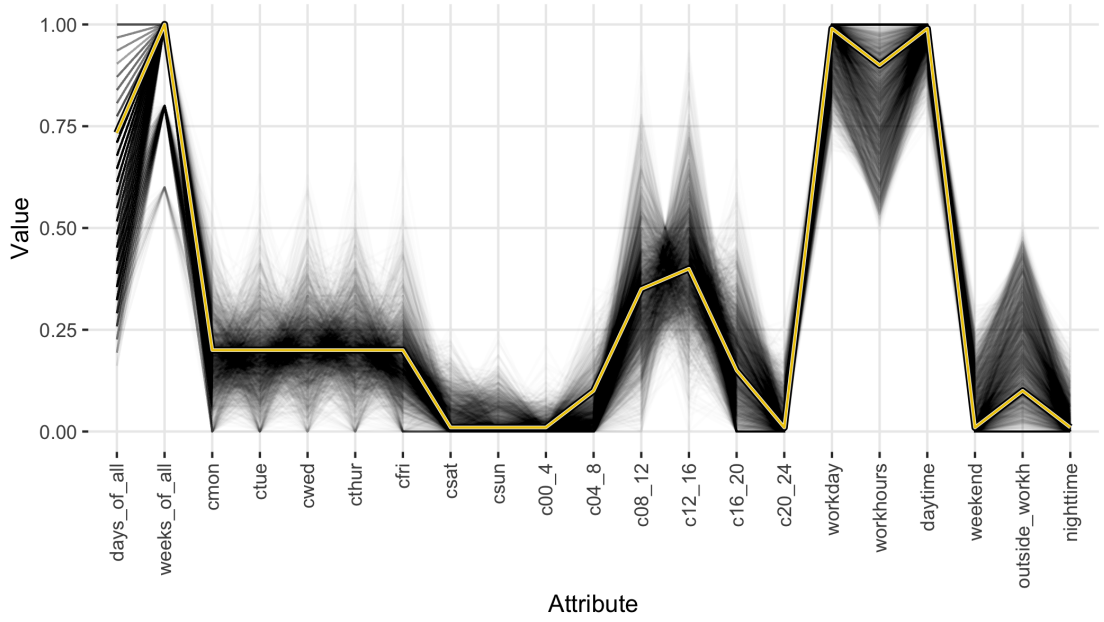

Figure 5. Initial work-time cells.

Describing the initial grouping step results in the example of the work-time anchor points. The yellow line on the figure 5 is an experts' conducted profile corresponding to values of the fourth row in table 12. According to the distance measure, the grey lines are all the locations that exhibited a similar activity pattern to the work-time anchor profile.

The threshold by which the locations were separated was 0.09. Consequently, if a location's activity pattern has a distance from the work-time anchor profile less than 0.09,

it is classified as a work-time location. Other groups' thresholds can be found in the Appendix table 14.

It is evident, that the most similar shape occurs in the last six variables ('workday','weekend','workhours','outside_workhours', 'daytime', 'nighttime'). The result is as expected because these are very distinctive variables to a work-time location. The cell should only be used on working days, work hours, and during the daytime. Other variables of the six should be close to zero. As anticipated with a correlation-based measure, the first two variables ('days_of_all', 'days_of_all') have caught active and less active individuals' locations. For example, cells that have been used almost on all days in the month and some only on eight days. Moreover, there is a distinctive peak during the daytime between 8:00-16:00 and a drop on Saturday and Sunday.

3.4 First principal component

The principal component analysis was used to improve the fit of anchor profiles to the data. The idea is to improve the profiles that experts had constructed. Describing the method in the next subsection in the example of work-time anchor points. Other anchor profiles are adjusted analogously.

Usually, with PCA, it is under interest which variables contribute to which principal components and how much. Each component is a linear combination of the variables. Generally, a PCA with a maximum of 21 components could be created to describe this data table. In this thesis, the concentration is on the first component. Using only the first component, all data points can be reconstructed. The advantage of the first component is its size- it is one-dimensional.

In this thesis, to find the first principal component, this data has to be transposed. The reason is that we are interested in creating a new anchor profile that projects best the data set under observation. Therefore the PCA on the transposed data set creates a linear combination of all the cells that were firstly filtered to be similar to the expert's work-time anchor profile. Moreover, this creates a weighted universal profile of all the locations in the subset. Therefore the work-time anchor profile is shifted by all cells in the subset. Some cells contribute more and some less to the new adjusted work-time profile.

The first component, from table 13, explains nearly 91% of the total variance. This is a high percentage and ensures that it is meaningful to use the first component to describe this subset. Therefore, the first component can be used as an adjusted work-time anchor profile. However, the values are not in the original scales anymore, but when calculating the correlation-based distance, it does not affect the resulting distance.

Consequently, a new distance matrix D_2 is calculated where the columns correspond to each modified anchor profile's distance from each location point from the data. Analogously new thresholds are calculated to select a new set of work-time locations as well

as each other location types. This iterative profile adjusting is done until the consecutive profiles do not differ significantly.

Distances between the new adjusted profiles for each profile are also calculated. When the distance between two consecutive modified profiles is less than ε, the iteration is stopped. In this thesis, the cue for an ending is $\varepsilon < 0.005$. The cue was chosen by analysing the distances between each of the profiles. Additionally, visual aid was taken into consideration.

3.4.1 Reversing the component

The original data set consisted of values in the interval [0,1], but the components range from [-112, 51]. From the perspective of classifying objects, the range is not important because, as stated, the shape of the profile is under investigation rather than the range of values. Nevertheless, for comparing the updated shapes, it is important that they can be visualised on the same scale. This subsection describes reversing the components to interpretable original scales and expresses the changes in the example of the work-time anchor profile.

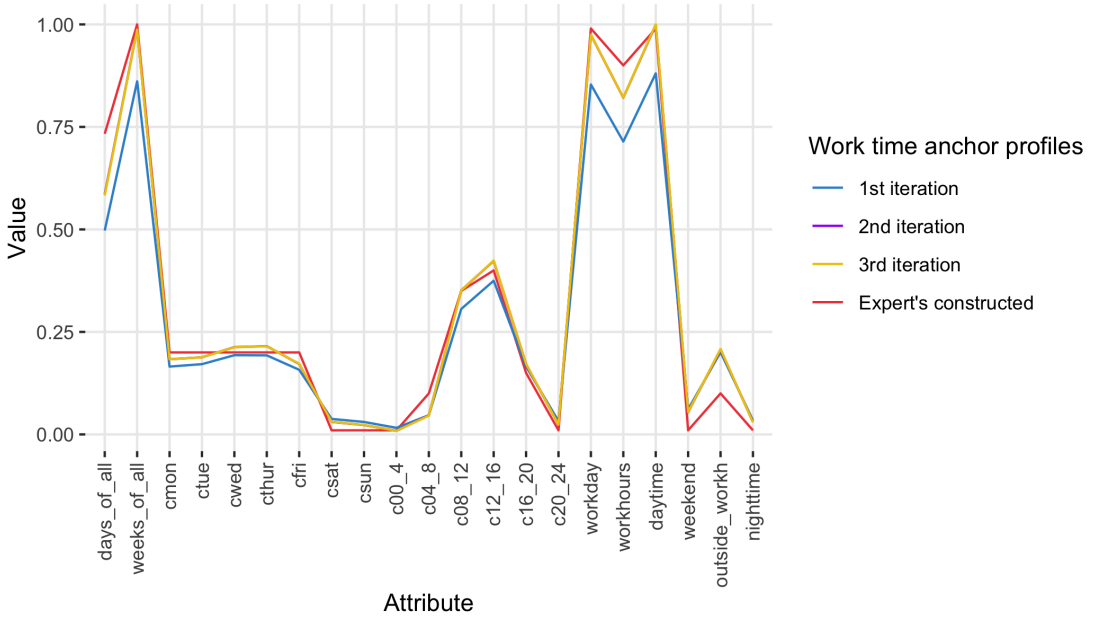

Figure 6. Changes in the work-time anchor profile.

The process of reverting the principal components was described in section 2.1.1. It must be noted that when reverting the components, the dimensions stay the same as the

original data set. Therefore for visual aid, one row must be chosen. For the comparison of the profile changes, the location that mostly contributed to the $PC1$ is drawn.

From the figure 6 it is seen that interestingly after the first iteration, the profile is less strict with working days and hours. However, the second and third iterations move back closer to the expert's constructed profile. On the weekdays, the new anchor profile has a softer decrease from Thursday to Saturday and a small percentage of activeness on the weekends. Secondly, the work hours are more flexible. The percentage of calls during work hours can be less and higher outside working hours.

Moreover, the second and third iteration anchor profiles overlap because the purple line for the second iteration is overlined with yellow, a third iteration anchor profile. This is an indicator that the iterating should stop.

3.5 Validating the method

In the current work, it is difficult to establish ground truth for validation purposes, as is often the case with big data sources where individuals in the data set remain anonymous. This means that it is not possible to check whether the anchor classification resulted in the semantics that the user themselves would assign to the location. One way to provide validation to the chosen methodology would be to use the help of an expert. There is an additional data set of a conducted survey, where the respondents' exact home locations are known. Therefore a second validation is performed on a data set with ground truth.

3.5.1 Expert assessment

After each iteration, a subset of new work-time anchor points is created. Moreover, the locations that did not make it to the following subset and locations that were added to the next subset were distinguished. Eventually, a data matrix is created, where each row is a cell that was removed when making a more accurate subset of work-time anchor points. Analogously a data matrix with locations that were added within the next iteration is created.

Therefore the labels of the locations changed during the iterations. It is under interest whether the changes were in the right direction. The method is considered adequate when locations that behave less like a work-time point are left out of the group, simultaneously adding locations that behave similarly to the work-time profile to the group. The subsection aims to ascertain whether the changes improved the quality of the work-time anchors' group.

A random sample of $n = 20$ cells was picked from both data sets to perform a sign-test. Therefore, 20 pairs were created, where one was the cell that was previously something else but added to work-time anchor points, and the second was a location that was initially a work-time point but removed from the subset. The expert was instructed to decide which profile of the pair is most likely a work-time anchor point. If the expert

chooses the location that was added to work-time points, the switch is considered a success.

The thesis proposes a way to validate results using a sign-test. The test is conducted in order to distinguish the consensus between the output and opinion of an expert.
The hypothesis pair to test is the following:
H_0: The probability of change in the right direction is 0.5.
H_1: The method changes the labels systematically towards the right direction (or towards the wrong direction).

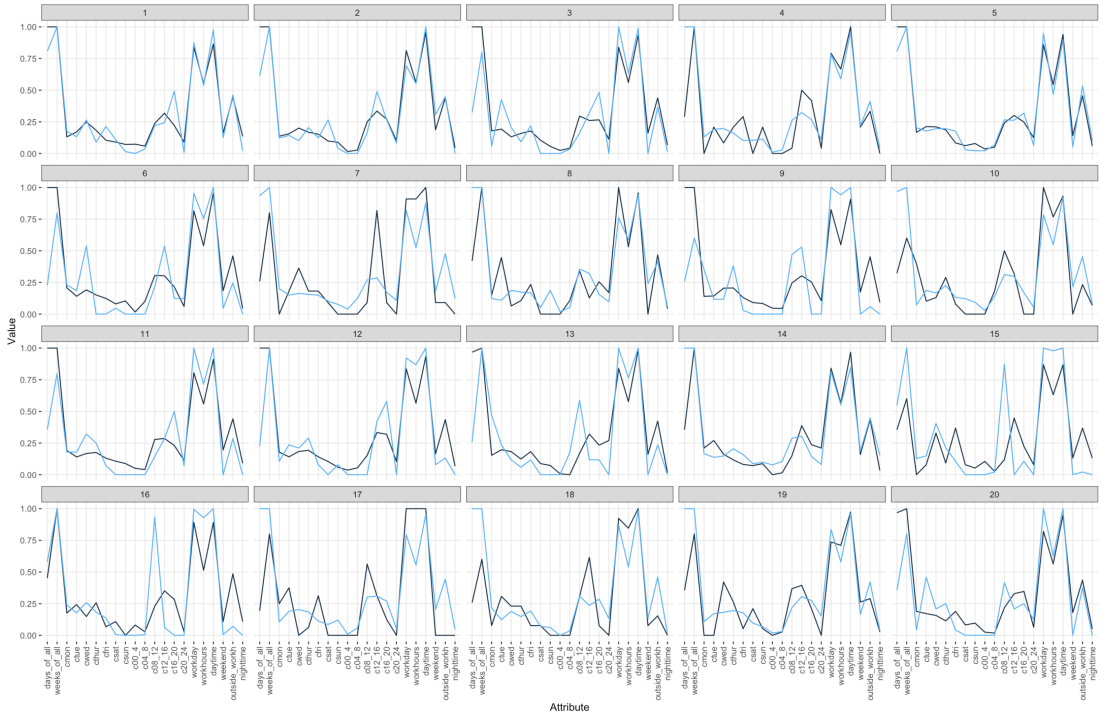

Figure 7. Sign-test arrangement.

In the figure 7, the two profiles have been given a random colour to avoid each group having a constant colour. The expert inspected figure 7 and concluded for every pair of $i = 1 \ldots 20$, which of the two is a work-time anchor point. Next, the results were compared to the methods' output values. When the expert and method were in agreement, it was marked as "+", otherwise "-". The resulting vector of agreement consisted of 18 "+" -s and 2 "-"-s. Therefore, the expert agreed with 18 cells out of 20 that were added to the group during any iteration.

There is a $\frac{1}{2}$ chance that the expert chooses the same work-time anchor as the method

for all i. Consequently, binomial distribution $B(20, \frac{1}{2})$ was used to test the hypothesis. An R function 'binom.test()' was used with number of successes $x = 18$ and number of trials $n = 20$ to get a ground for the decision.

Based on the resulting p-value 0.0004 of the test, the null hypothesis is rejected. It can be concluded that the changes in locations between the groups were adequate. Therefore, the method was able to distinguish between work-time points and other types of points.

3.5.2 External data set

In addition to validating the methods consensus with the expert, a validation was conducted on another data set. The test data set consisted of 151 people with known real home locations and, in total, 764 location points. The data set was conducted with the consent of the 151 people who disclosed their home location in regards to a study [SL18]. The time frame used for validation is March 2017.

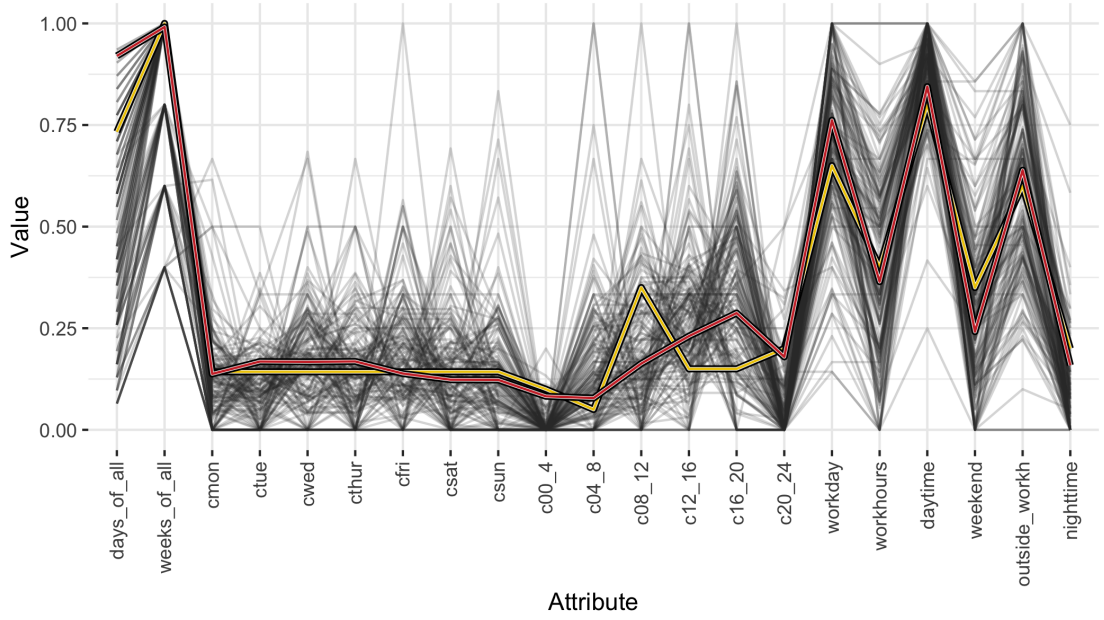

Figure 8. Real home locations and anchor profiles.

The plot 8 shows the behaviour of known real homes of the data set with grey; the yellow line shows the expert's constructed profile, and the red line shows the updated profile. It can be seen that the locations behave variously. The home locations are mostly moderately active, with points at both extremes. During the weekday, the distribution

of activity is seemingly random. There might be slightly less activity on Mondays and higher activity towards the end of the week. Additionally, there are groups of locations that have zero activity on a random weekday.

Almost all real home locations have nearly no activity during night hours, 20:00-04:00, with increased activity from morning to evening. Mostly the locations are more actively visited during working days, daytime and outside working hours. It is evident that the real home locations differ from the two profiles. For example, the expert assumes low activity on the weekends, while in reality, the activity is nearly zero in many cases. However, the expert described higher activity in the morning-noon, but the corrected profile has shifted the peak into the evening. In the figure 8, it is evident that the home profiles do have higher activity in the evening rather than in the morning.

There are a total of 151 home locations out of all 764. This means that with an approximate probability of 0.2, the correct home location is chosen when picking a random location. When classifying locations with their closest anchor profile, a total of 78 locations were classified as home location points. Of the 78 locations, 38 were classified correctly, giving a probability of 0.49 for correct classification.

Using the final fine-tuned anchor profile and a predefined threshold, out of 764 location points, 34 were selected to be homes. Out of the 34 locations, 23 locations were actually according to real homes. This leaves 11 to be wrongly classified. The accuracy of 67% is adequate considering the inconsistent behaviour of the real home locations.

It is more meaningful to classify fewer locations as homes if it is known that the probability of correctly classifying is high. Therefore in this thesis, the locations are left rather unclassified than grouped to a wrong class. Some locations might have similar activity patterns to the home anchor profile but are left unclassified because the classification rule was too strict.

Additionally, the idea of the method is to search for similarly behaving location points with the given initial shape. The location that the individuals have reported to be their homes might not behave how the expert describes them to behave. Meaning that multiple different profiles can potentially describe a home location. In home detection, adjusting the expert's constructed home profile did not improve the results. Furthermore, the home anchor profile that this thesis focused on is based on the primary location for spending the night.

There is an additional problem that might have occurred during the classification. The threshold to classify locations to home anchor group might have been too low. The threshold was calculated for the large data set and might be too restrictive for testing. Therefore a ROC curve can be drawn to assess how the sensitivity increases.

Figure 9 shows two ROC curves in comparison. The red curve describes how the created anchor profile's sensitivity increases with specificity. Moreover, PCA was also conducted on the reported real home locations. The first component of the PCA was used as renewed home anchor profile. Next, distances from each data point in the set

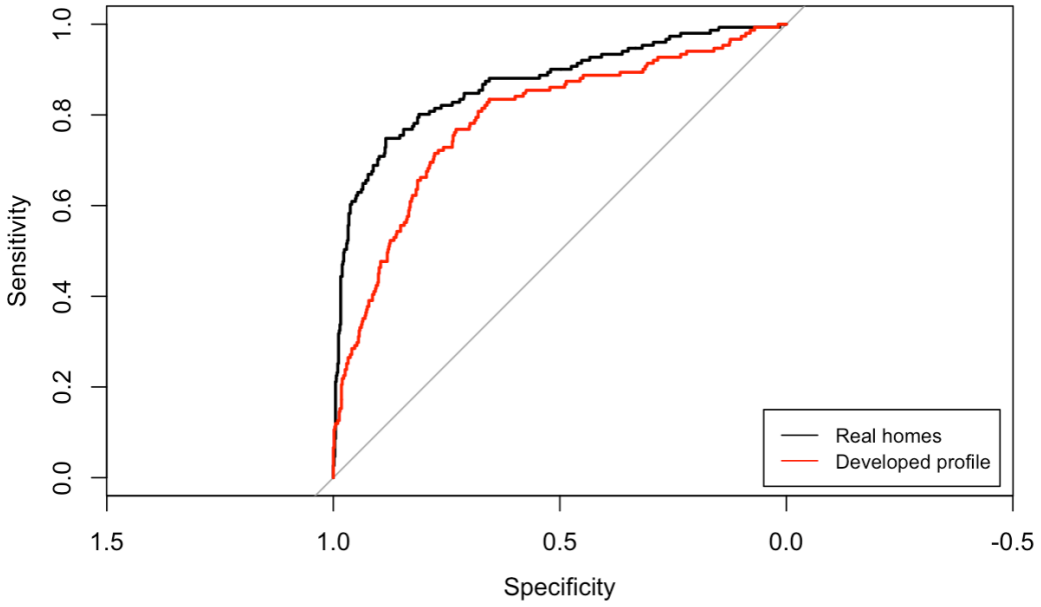

Figure 9. ROC curves of different methods

were calculated from the new profile. The black curve shows the diagnostic ability of the new profile based on the distances.

It can be seen that the profile based on real homes has better diagnostic ability than the profile created on unclassified data with the expert's opinion. With the increase in specificity, the real home profile has higher sensitivity at each point. The area under the curve for the real home profile is 0.87 with a 95% confidence interval of 0.83-0.91. The area under the curve for the constructed profile is 0.79 with a 95% confidence interval of 0.75-0.83. Therefore the constructed profile is rather good but not as good as the profile from real data with training labels.

3.6 Results

The thesis created a method for classifying the locations of subscribers with profiling. Five profiles were constructed by experts and adjusted to find similar locations from the data set. The profiles consisted of 21 different time-based variables that described the behaviour of a location type. These profiles were used as an input for the classification.

The method is promising on the data it was developed on. For each person, the location points were aggregated and assigned a behavioural pattern based on the activity at that location. Based on the pattern and predefined anchor profiles, groups of similar

locations were distinguished. Figure 10 shows the final work-time anchor class. The class has remained similar in shape to the expert's constructed profile. However, the updated class is smaller and has less variance.

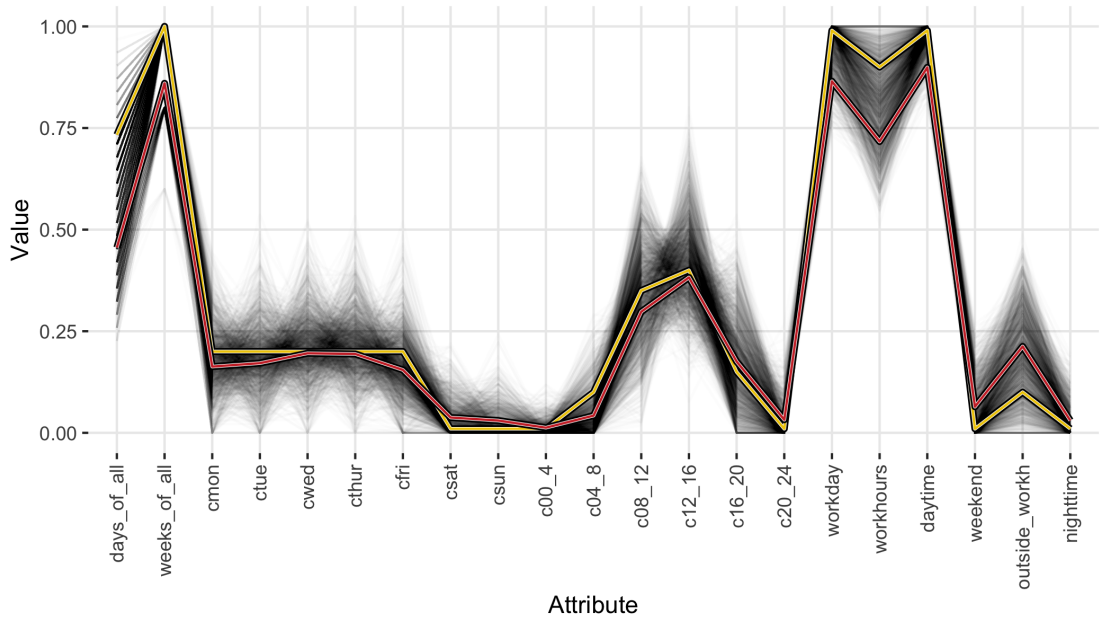

Figure 10. Final work-time anchor class.

Other 4 location type groups are given in the Appendix, figures 13-16. In each of the figures, grey lines are the detected locations, yellow is the expert's constructed profile, and the red line marks the adjusted profile. The primary home location profile was adjusted to be more evening concentrated. The method did change a lot with secondary home anchor points (figure 14). The emphasis on the weekdays falls strongly into Saturday. Interestingly the night time-interval 20:00-24:00 and working day variables have opposite directions. The adjusted profile has lower evening activity and higher workday activity than the expert's profile.

The nighttime work type location group (figure 15) also changed the shape of the pattern. The weekday and time interval distributions have smoothed out without no longer differentiating the essential properties of a nighttime working place. Additionally, the daytime proportion has significantly increased, concluding that there might have been no nighttime anchor points in the data set. Alternatively, the nighttime locations behave differently in some variables than the expert described and hence were not detected.

Evening free time location points (figure 16) have been detected seemingly better. The weekday distribution has also smoothed out, but the distinction of 16:00-20:00 has

remained in the adjusted profile. Moreover, the specific variables, such as high workday and low work hours variables, have stayed similar to experts.

A test was conducted with the help of an expert to test the method's ability to classify similarly behaving locations. The expert needed to distinguish between a work-time and a non-work-time location profile. The test was successful, as the expert detected 18 work-time locations that the method also classified as work-time locations. Therefore, with the help of the expert, the method was considered to find similarly behaving locations.

Interestingly, contrary to promising results from the sign-test, the external data set did not give the expected results. There can be multiple reasons behind this. Firstly, it is possible that the method was adjusted to the original data and is not suitable for use on other data sets. The method might have picked up very specific cases that were not suitable for generalisation. It is also possible that the variables to profile the locations were insufficient.

Another possible reason for this contradiction is the reliability of the test data. It is not known how the question about the real home was asked from the respondents. There might be nuances that can affect the answer. For example, when asking where a person's home is, their understanding of "home" might be different. It might not be the primary location they sleep at night, but a place that is home in their hearts. For example, students in Tartu, who live in a dormitory, tend to mark their home as their parent's home. This leads to errors in conclusions because their actual activity lies in a different location. According to detecting meaningful locations, their parent's home is considered their secondary home location.

One possibility is to include a more extended period of data in the future to overcome the problem. Even though the meaningful locations are calculated with monthly accuracy, location classifications can be improved in retrospect. For example, if one month's location is left unclassified, but in the next month, it is in a specific class, it indicates that it could have been in the same class previously. However, this needs a more complex analysis out of this thesis scope.

Conclusion

The method of classifying similarly behaving locations from MPD was discussed in the thesis. Each subscriber's visited locations were described by different variables to profile the activity pattern in the location. An expert constructed the desired profiles to classify locations by the variables. The expert's anchor profiles were adjusted with the method described in this thesis without knowing the ground truth. Additionally, a profile was created when the ground truth was known.

Multiple aspects should be considered while concluding the method. Clearly, there exist limitations to the method. Even if there is an expert's constructed profile for a very specific location type, it necessarily does not mean that the type can be found in the data. The method still allows to classify of the locations according to the predefined profiles, which can give extra semantics to the location – e.g. location is predominantly daytime or weekend location. This way, profiling has allowed to bring in new possibilities to characterise locations that were previously just marked as regularly visited locations.

In the future, the profiling variables could be closely examined. There might be additional variables to add behavioural value to the location points. In addition to the expert's opinion on the attributes, the necessity of each variable should be established. Additionally to time-based variables, spatial-temporal variables could be considered in the profiling.

Moreover, there are multiple methods for choosing a suitable threshold for subsetting the profile types. In this thesis, the smallest distance to another centre divided by two was chosen. In the future, the threshold could depend on the proportion of the type of locations in the population. Therefore, when the expert is constructing initial anchor profiles, additionally, the expected proportion of subscribers having the type of location should be considered.

All in all, the method fulfilled its purpose of finding groups of location points that have similar activity patterns as the given input. The classification of similar locations provides additional semantics to the "Other regularly visited places" class in the anchor model.

References

[AAR+08] Rein Ahas, Anto Aasa, Antti Roose, Ülar Mark, and Siiri Silm. Evaluating passive mobile positioning data for tourism surveys: An estonian case study. *Tourism Management*, 29:469–486, 06 2008.

[ASJ+10] Rein Ahas, Siiri Silm, Olle Järv, Erki Saluveer, and Margus Tiru. Using mobile positioning data to model locations meaningful to users of mobile phones. *Journal of Urban Technology*, 17:3–27, 04 2010.

[BBG+18] Hugo Barbosa, Marc Barthelemy, Gourab Ghoshal, Charlotte R. James, Maxime Lenormand, Thomas Louail, Ronaldo Menezes, José J. Ramasco, Filippo Simini, and Marcello Tomasini. Human mobility: Models and applications. *Physics Reports*, 734:1–74, 2018. Human mobility: Models and applications.

[Com14] European Comission. Feasibility study on the use of mobile positioning data for tourism statistics. consolidated report eurostat contract no 30501.2012.001- 2012.452. 2014.

[Est10] Statistics Estonia. Ak091: 10-aastased ja vanemad nädalapäeva, ajavahemiku ja põhitegevuse järgi. 2010.

[HKT01] Jiawei Han, M. Kamber, and Anthony Tung. Spatial clustering methods in data mining: a survey. *Data Mining and Knowledge Discovery - DATAMINE*, 01 2001.

[Ize13] Alan Izenman. *Cluster Analysis*, pages 407–462. 01 2013.

[JC16] Ian T. Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

[JWHT13] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer Texts in Statistics. Springer, 2013.

[LRS+18] Titi Lestari, Rifa Rufiadi, Erki Saluveer, Siim Esko, and Sarpono Dimulyo. Indonesia's experience of using signaling mobile positioning data for official tourism statistics indonesia's experience of using signaling mobile positioning data for official tourism statistics. 11 2018.

[MST+17] Sidharth Mishra, Uttam Sarkar, Subhash Taraphder, Sanjoy Datta, Devi Swain, Reshma Saikhom, Sasmita Panda, and Menalsh Laishram. Principal component analysis. *International Journal of Livestock Research*, page 1, 01 2017.

[Pan22] Eesti Pank. Välisreiside statistika koostamise metoodika. 2022.

[Ros10] Sheldon M. Ross. Chapter 3 - using statistics to summarize data sets. In Sheldon M. Ross, editor, *Introductory Statistics (Third Edition)*, pages 71–143. Academic Press, Boston, third edition edition, 2010.

[SL18] Kaja Sõstra and Kristi Lehto. Using mobile positioning for improving the quality of register data. 2018.

[SP17] Kalpit Soni and Atul Patel. Comparative analysis of k-means and k-medoids algorithm on iris data. 13:899–906, 01 2017.

[Tir14] Margus Tiru. Overview of the sources and challenges of mobile positioning data for statistics. *In Proceedings of the internation conference on Big Data for official statistics*, pages 8–18, 2014.

Appendix

I. Overview of data

Table 5. Raw data example of 10,000 people sample, 2019 January.

	pos_id	**ms_id**	**pos_time**	**cell_id**
1	90463376	A	2019-01-01 09:54:44	196
2	89664350	A	2019-01-01 10:51:43	196
...	...	...	...	...
$4,140,665$	283180429	ABC	2019-01-31 11:52:35	6634

Table 6. Raw data example 10,000 people sample, 2019 January and more active individuals.

	pos_id	**ms_id**	**pos_time**	**cell_id**
1	23240032	B	2019-01-01 14:13:59	2297
2	20592097	B	2019-01-01 13:59:37	3440
...	...	...	...	...
$2,453,765$	928452621	BCD	2019-01-31 12:29:49	16629

Table 7. Validation test data set.

	pos_id	**ms_id**	**pos_time**	**cell_id**	**home_cell**
1	8239	C	2017-03-01 10:08:39	2348	2201
2	8240	C	2017-03-01 16:52:28	2201	2201
...	...	...	...	...	...
$15,421$	74118	CDE	2017-03-01 12:59:20	3763	9514

II. Data aggregation

Table 8. Grouping data by the percentage of days and weeks.

	ms_id	**cell_id**	**days_of_all**	**weeks_of_all**
1	A	2	0.16129032	0.8
2	A	402	0.96774194	1
...	...	...	...	...
$39,539$	ABC	16174	0.1290323	0.6

Table 9. Grouping data by the percentage of weekdays.

	ms_id	***cell_id***	***mon_of_all***	***tue_of_all***	***wed_of_all***	***thu_of_all***	***fri_of_all***	***sat_of_all***	***sun_of_all***
1	A	2	0.143	0.285	0.143	0.0	0.429	0.0	0.0
2	A	402	0.115	0.169	0.188	0.139	0.188	0.055	0.15
...	...	...	...	...	...	...	...	...	...
$39,539$	ABC	16174	0.125	0.0	0.5	0.375	0.0	0.0	0.0

Table 10. Grouping data by the percentage of 4-hour time intervals.

	ms_id	***cell_id***	***00_04***	***04_08***	***08_12***	***12_16***	***16_20***	***20_24***
1	A	2	0.0	0.0	0.143	0.857	0.0	0.0
2	A	402	0.121	0.042	0.29	0.248	0.194	0.103
...	...	...	...	...	...	...	...	...
$39,539$	ABC	16174	0.0	0.0	0.25	0.75	0.0	0.0

Table 11. Grouping data by other variable pairs.

	ms_id	***cell_id***	***work_day***	***weekend***	***work_hours***	***non_work_hours***	***daytime***	***nighttime***
1	A	2	1.0	0.0	1.0	0.0	1.0	0.0
2	A	402	0.8	0.2	0.485	0.515	0.818	0.182
...	...	...	...	...	...	...	...	...
$39,539$	ABC	16174	1.0	0.0	1.0	0.0	1.0	0.0

Table 12. Anchor profiles.

	days_of_all	weeks_of_all	mon_of_all	tue_of_all	wed_of_all	thu_of_all	fri_of_all	sat_of_all	sun_of_all
Home	0.73	1.0	0.143	0.143	0.143	0.143	0.143	0.143	0.143
Secondary home	0.167	0.6	0	0	0	0.071	0.36	0.429	0.286
Evening free time	0.33	1.0	0.286	0.143	0.286	0.143	0.143	0.0	0.0
Regular work	0.73	1.0	0.2	0.2	0.2	0.2	0.2	0.01	0.01
Nighttime work	0.73	1.0	0.286	0.143	0.286	0.143	0.143	0.0	0.0

	00_04	04_08	08_12	12_16	16_20	20_24	work_day	weekend	work_hours	non_work_hours	daytime	nighttime
Home	0.1	0.05	0.35	0.15	0.15	0.2	0.65	0.35	0.4	0.6	0.8	0.2
Secondary home	0.0	0.1	0.2	0.2	0.2	0.3	0.2	0.8	0.6	0.4	0.85	0.15
Evening free time	0.0	0.0	0.0	0.1	0.8	0.1	1.0	0.0	0.1	0.9	0.99	0.01
Regular work	0.01	0.1	0.35	0.4	0.15	0.01	0.99	0.01	0.9	0.1	0.99	0.01
Nighttime work	0.35	0.4	0.15	0.0	0.0	0.1	1.0	0.0	0.01	0.99	0.1	0.9

III. Model outputs

Table 13. PCA summary of work-time cells.

	PC1	**PC2**	**PC3**	**PC4**	**...**	**PC20**	**PC21**
Std. deviation	20.6297	3.9796	2.5632	2.14715	...	$2.096e-15$	$1.358e-15$
Prop. of variance	0.9076	0.03378	0.01401	0.00983	...	0.0	0.0
Cum. proportion	0.9076	0.94140	0.95541	0.96524	...	1.0	1.0

Table 14. Class radii.

	Initial	**1st iteration**	**2nd iteration**
Home	0.090	0.077	0.048
Secondary home	0.194	0.162	0.145
Evening free time	0.116	0.077	0.065
Regular work	0.090	0.085	0.087
Nighttime work	0.191	0.109	0.048

IV. Plots

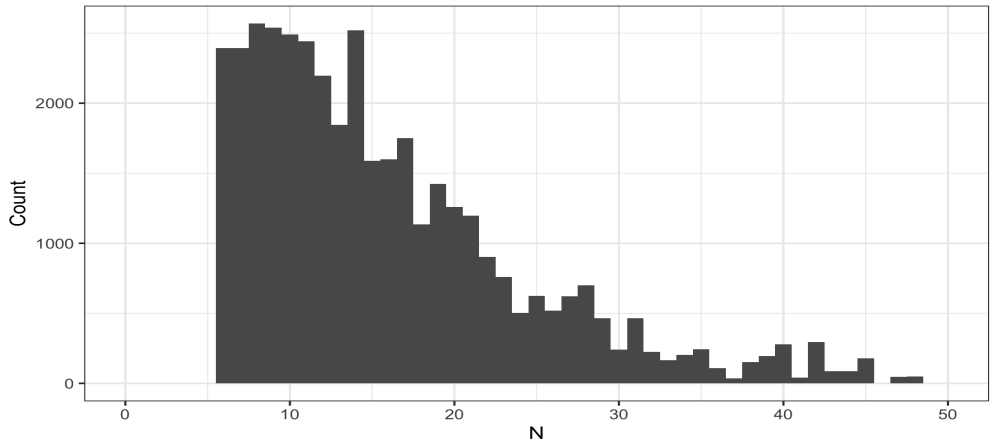

Figure 11. Histogram of records per unique user-cell combinations in 2019 January.

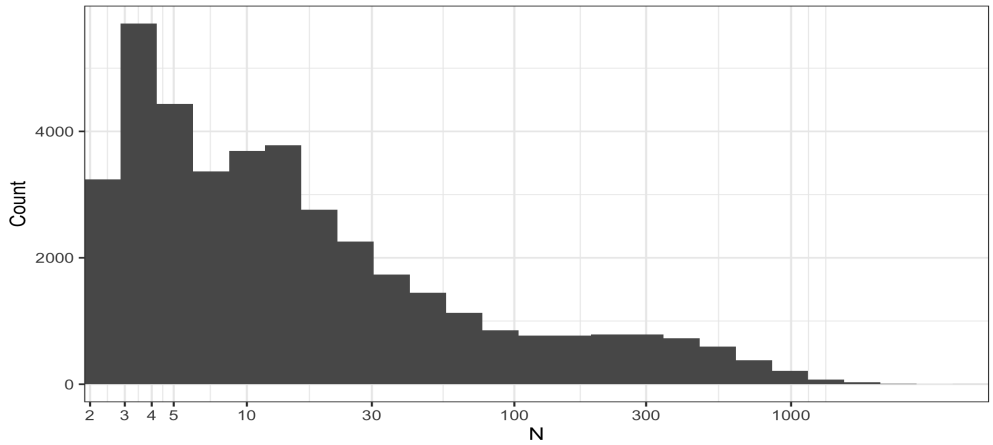

Figure 12. Histogram of unique cells per user in 2019 January.

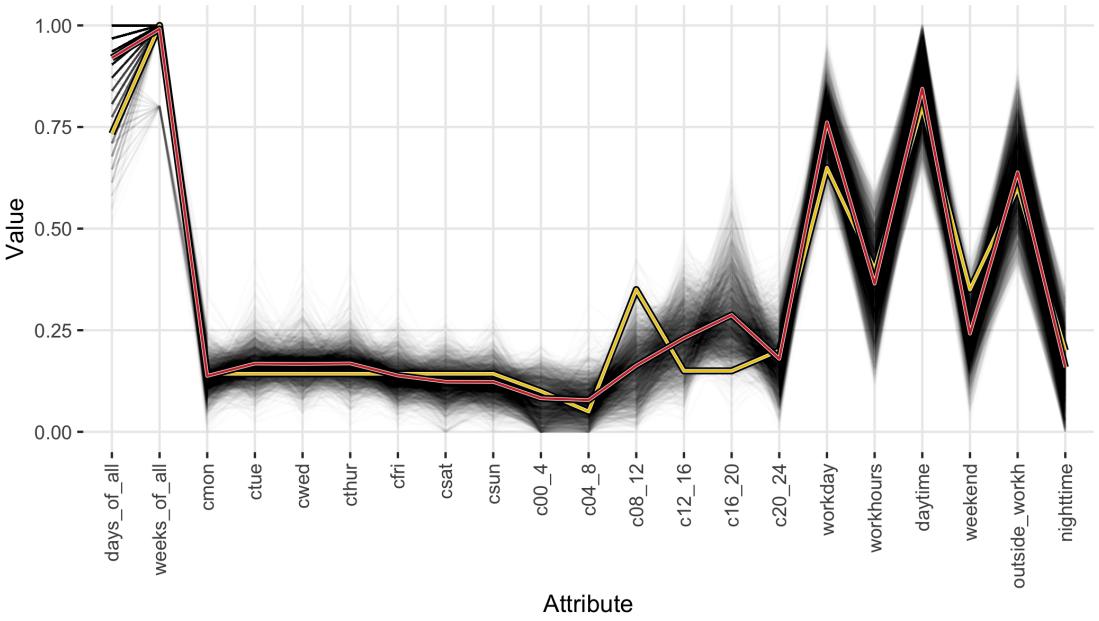

Figure 13. Final home anchor points.

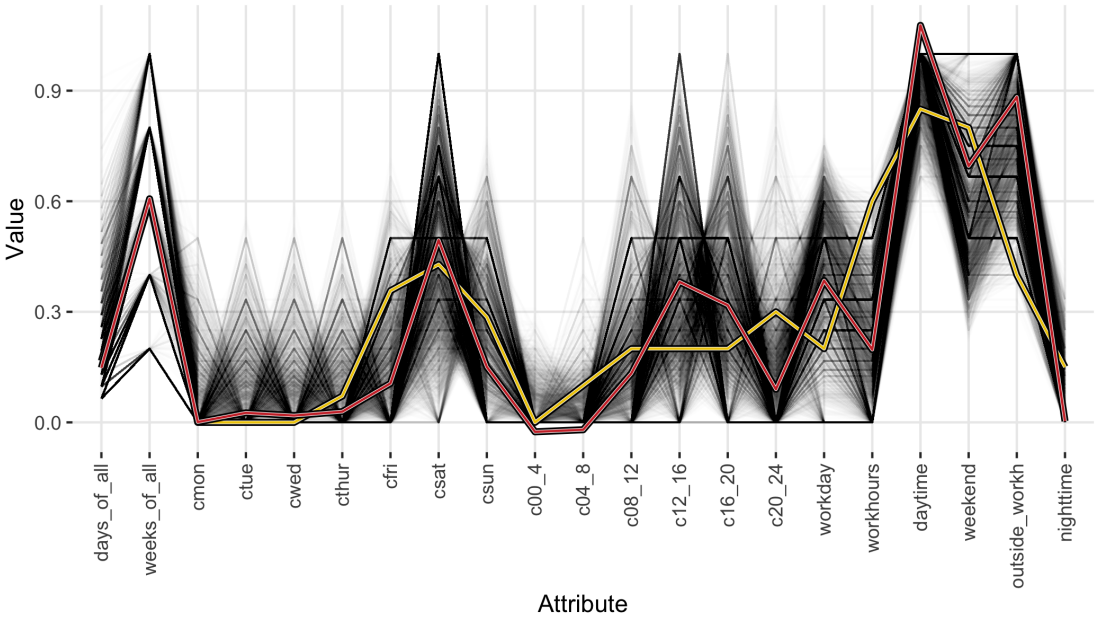

Figure 14. Final secondary home anchor points.

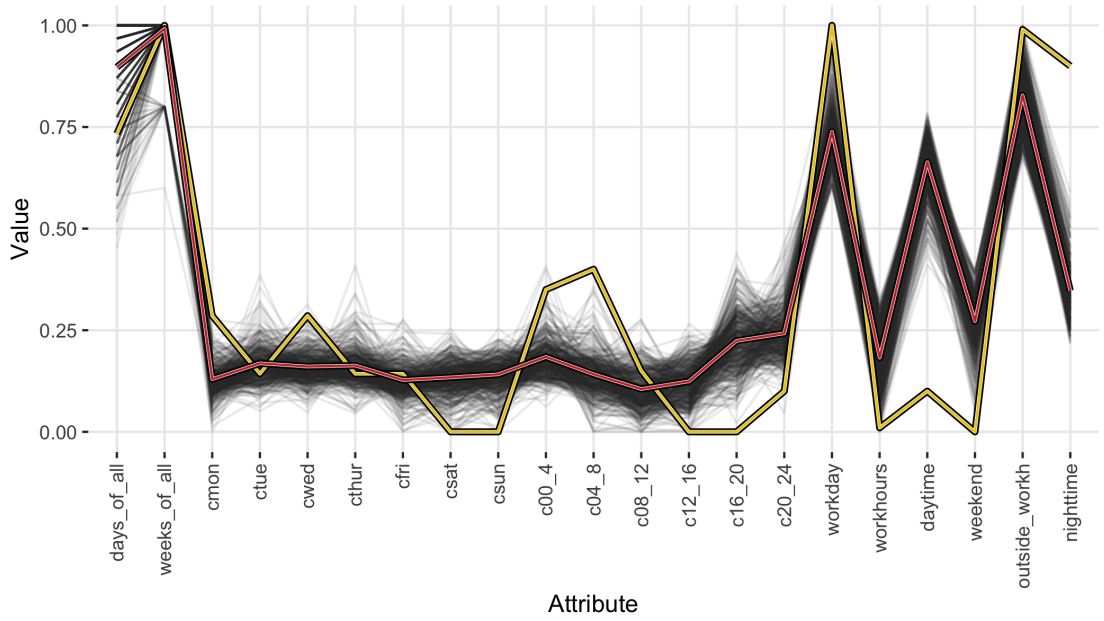

Figure 15. Final night time work anchor points.

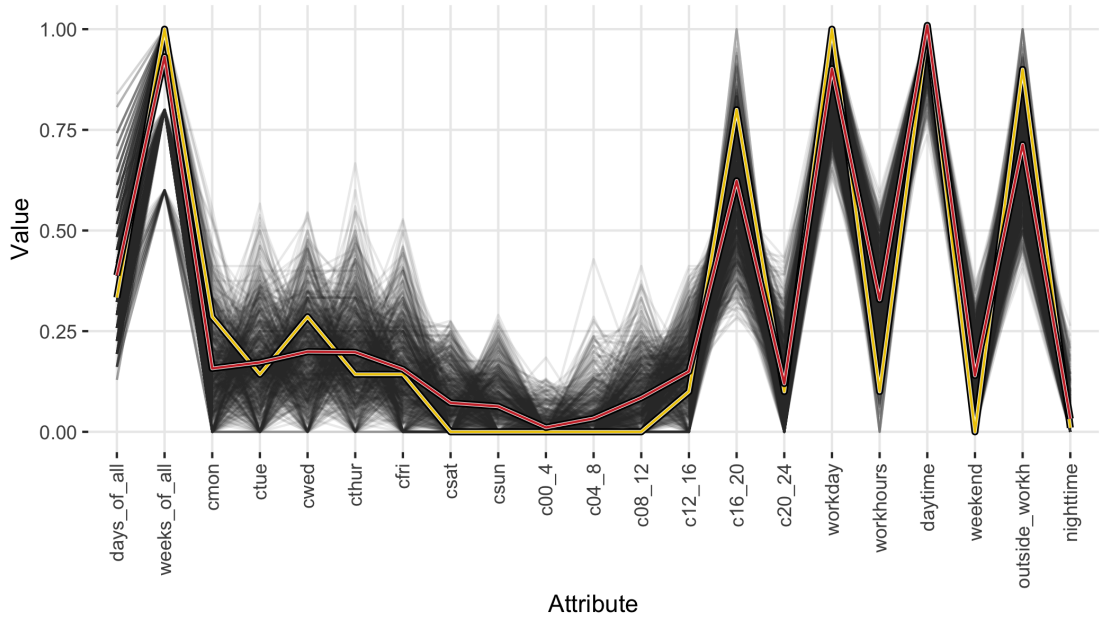

Figure 16. Final evening free time anchor points.

V. Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Brigitta Rebane**,
(author's name)

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

   reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

   Detection of meaningful locations from passive mobile positioning data using location profiling,
   (title of thesis)

   supervised by Märt Möls and Kaisa Vent.
   (supervisor's name)

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

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

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

Brigitta Rebane
17/05/2022