University of Tartu

Faculty of Science and Technology

Institute of Ecology and Earth Sciences

Department of Geography

Master Thesis in Geoinformatics (30 EAP)

Constructing minimal zoning from overlapping coverage areas

Patrick Joan Thomson

Supervisor: PhD Alexander Kmoch

MSc Margus Tiru

Tartu 2021

ABSTRACT

Passive mobile positioning data has been used widely to study mobility, tourism, segregation etc. Few studies have focused on the accuracy of data interpolation to space. A homogeneity approach was applied to a dataset from one of the mobile network operators in Estonia to assess the accuracy of mobile positioning data-based results on local administrative unit levels. Also an algorithm was developed to explore the possibility of generating homogenous data-based zoning based on different input parameters. The homogeneity approach can be used to assess the accuracy of results but it needs to be developed further. Data based zone generation algorithm serves as a good starting point for future developments.

Keywords: Mobile positioning data, homogeneity, interpolation

CERCS code: P175 Informatics, systems theory; S230 Social geography

Annotatsioon

Passiivseid mobiilpositsioneerimise andmeid on laialdaselt kasutatud mobiilsuse, turismi, segregatsiooni jms uurimiseks. Vähesed uuringud ei ole keskendunud andmete interpoleerimise täpsusele. Eesti ühe mobiilsidevõrgu operaatori andmekogumile rakendati homogeensuse lähenemisviisi, et hinnata mobiilpositsioneerimise andmetel põhinevate tulemuste täpsust kohaliku haldusüksuse tasandil. Samuti töötati välja algoritm homogeense andmepõhise tsoneeringu genereerimise uurimiseks kasutades erinevatel sisendparameetritel. Tulemuste täpsuse hindamiseks saab kasutada homogeensuse lähenemisviisi, kuid seda tuleb edasi arendada. Andmepõhine tsoonide genereerimise algoritm on hea lähtepunkt edasiseks arenguks.

Märksõnad: mobiilpostitsioneerimise andmed, homogeensus, interpoleerimine

CERCS kood: P175 Informaatika, süsteemiteooria; S230 Sotsiaalne geograafia

TABLE OF CONTENTS

ABBREVIATIONS

BTS - Base Transceiver Station (same as mobile network cell, or simply cell)

ICT - Information and Communication Technology

IoT - Internet of Things

LAU - Local Area Unit

MNO - Mobile Network Operator

MPD - Mobile Positioning Data

PDM - Positium Data Mediator

PIEG – Positium Interpolated Event Grid

GLOSSARY

Country of Reference - The country for which MPD is processed for, e.g. Estonia is a country of reference when processing MPD from Estonian MNOs.

Stay, stay section - Element of continuity model with temporal duration and location (either a grid square or country)

Trip - Trip starts from subscribers' departure from and ends before subscribers' arrival to home (POR anchor in case of residents, COR in case of non-residents). Trip consists of one or many sub-trips which in turn consist of one or many stay and move sections.

INTRODUCTION

The usage of mobile positioning data (MPD) is widespread. With the development of information and communication technology (ICT) a significant impact on marketing processes has been made. In addition the availability and importance of using big data sources, such as MPD has increased in past years (Roberts et al., 2014). While most studies using MPD have focused on the population distribution (Järv et al., 2018), tourism (Ahas et al., 2008; Raun et al., 2016) and segregation (Silm & Ahas, 2014) not many have focused on the accuracy of spatial interpolation from MPD.

Different methods have been used for spatial interpolation of MPD. When using mobile positioning data people who have used the services of mobile operators can be traced back to the telecommunication antenna they were connected to at the time. Therefore, we could use the location of the antenna and give out results based on that. However, this is complicated by the fact that telecommunication antennas have large coverage areas and simply positioning people by the location of the telecommunication antenna would yield inaccurate results on a larger scale. Järv et al. used a combination of a statistical grid and the origin and destination of the peoples trips and filled the gaps using routing engines to predict movement. The interpolation method of the starting and ending points of trips was not specified (Jarv et al., 2018). Another widely used interpolation methodology is area weighted index. Aasa concluded that point-in-polygon and area weighted index have a poor performance, but housing weighted index and Adaptive Morton grid increase the descriptive power of estimating population based on MPD (Aasa et al., 2021).

Often users of MPD-based results expect that indicators are presented on existing administrative unit levels and would like to know how many people were at a specific location at a specific time. For example, telling clients or policy makers that 30% of all users in one mobile operator network have made Call Detail Records (CDR) in the coverage area of a certain telecommunication antenna would not be helpful. Generally, it is more comprehensible to give out statistics on more known formats, such as the amount of people in a certain administrative unit in a specific timeframe. If MPD is used to create widely used statistics and statistics are generally given out based on local administrative units or other custom areal distributions, then there is a need to verify the accuracy of the interpolated results. Ground proof data for a data source this large is hard to come by and gathering it would be unfeasible. This creates a need for a methodological approach on how to assess the accuracy of the presented statistical results that are based on MPD.

There are other ways of creating areas to give out statistics such as creating clusters of cell masts and generating Thiessen polygons around them (Ogulenko et al., 2021). Thiessen polygons (also known as Voronoi polygons) don't take into account the overlapping of different coverage areas. One way to improve the quality of output is to generate data specific zones in which the interpolation and data distribution is homogenous. To get an input dataset from which a data-based zoning is created methodology by Positium was used. Methodology by Positium in combination with Positium Data Mediator (PDM) software uses multiple different input datasets to improve the quality and accuracy of interpolating MPD data. PDM uses a probabilistic approach for calculating coverage areas for telecommunication antennas and using reference data such as a road, land use and building layers to determine more probable locations where the event in the mobile network operator range was made. PDM was used to create an input dataset in which the events have already been interpolated to space (M. Tiru, personal communication, May 2021). PDM is a mobile big data processing and analytics engine that uses several models in data processing to provide the highest quality results and analysis from mobile positioning data (*Positium Data Mediator*, n.d.).

There appears to be no theoretical or empirical research on the accuracy of interpolation from a telecommunication antenna to space which takes into account the overlapping coverage areas. Therefore this thesis aims to apply a methodology to assess the precision of interpolated MPD events through analysis and explore the possibility to create homogeneous zones algorithmically based on mobile positioning data interpolation and apply the methodology on an aggregated dataset that consists of 5 consecutive months and where the events have been interpolated to a custom grid. The Positium methodology for spatial interpolation used in this thesis is based on one approach of coverage area calculation model and spatial interpolation of network-antenna events to coverage area with various configuration parameters. The aspect of the original Positium interpolation method is not a subject of this thesis, but this thesis aims to assess the accuracy of its results.

The aim can be achieved through the following research questions:

Q1: Can the homogeneity approach be applied to assess the precision of MPD statistics?

Q2: Can zones be created algorithmically based on different input parameters?

1. THEORETICAL BACKGROUND

1.1. ZONING

Zoning is a method in which a municipality, person or other authority divides land into areas called zones. Zones may vary in size and other attributes. In the context of this thesis a zone is a collection of smaller grid units. The primary objective of zoning is to segregate different areas based on the homogeneity index.

1.2. HOMOGENEITY

1.2.1. HOMOGENEITY INDEX

The concept of homogeneity has been used in various fields to describe and quantify surfaces, landscape patterns, plant species etc (Afrianto et al., 2020; Lechthaler et al., 2020; Moreno-Mateos et al., 2008). Base Transceiver Stations (BTS) have coverage areas where mobile devices can connect to the antenna. MPD has no information on the spatial accuracy better than the location of BTS. PDM uses a probabilistic method to assign mobile devices to adaptive grid squares within the coverage area of cells. These probabilities are based on various reference data (e.g. land-use, buildings, roads, etc.). Because network cells are located more sparsely than required spatial result granularity (e.g. often there are no cells located within lower levels of LAU), spatial interpolation from cell coverage areas to adaptive grid is used. Though, the probabilistic method means that interpolation to the grid is statistically close to reality (when aggregating grid results to larger territories), it does mean that for individual subscribers, the location is not probabilistic, and not guaranteed. This creates the need for a measure of homogeneity for specific areas to show the probability of subscribers interpolated to grid squares in the specific area, were actually present in this area (if there are 1000 stays in area A, what is the probability that all of them were actually present in this area?). Stays are the smallest spatio-temporal units in PDM software and represent the location of a subscriber at a given time (see section 2.1.5). We are using a homogeneity index as this measure. H-index is just one approach to measure accuracy. There might be multiple methods for tackling this problem but in this thesis I focus on h-index.

In current methodology, we refer to the concept of homogeneity to the diversity resulting from the interpolation of cells to grid. The homogeneity index (h-index = 0…1) of an area is the proportion

of the number of stays in grid squares within the area to the number of all stays in grid squares interpolated from cells that have also been interpolated to the area. The formula for calculating h-index is following: **Hp = Sp / Sp'**

where

Hp - homogeneity index for territory p.

Sp - number of stays aggregated to grids within the territory p.

Sp' - number of all stays in grid squares interpolated from cells that have also been interpolated to the territory p.

H-index can be 1 in case of closed territories (isolated islands) or whole territories (whole country). Homogeneity index is 1 when it is 100% certain that all presences in the area are actual presences in this area. For example, h-index is 1 for the whole country as a territory. H-index can be 1 or very close to 1 for territories representing a distant island far from the mainland. H-index can also be close to 1 in closed urban areas with significantly high population count compared to suburban areas around it.

H-index can never be 0 as it should not be calculated for territories where there are no stays (e.g. cannot measure diversity where nobody is present - it is neither diverse nor uniform). Homogeneity index is close to 0 for very small territories where the majority of cell coverage areas are overlapping several neighboring territories.

In simple terms the homogeneity index shows the certainty that the number of subscribers (or stays etc.) were actually present in the area. If the index is 1, that means for sure all subscribers that have been indicated to be present in the area, are actually present in this area (and nowhere else). If the index is nearing 0, this means that there is a very high probability that the subscriber indicated to be present in the area may actually be present somewhere else (presumably in neighboring areas). For the whole country, the homogeneity index is always 1 because within a country, all interpolation is done within the area of the country. The smaller the territorial units (areas), the smaller the index gets.

In terms of zoning, the objective is to establish the measure of homogeneity to each territorial unit. It can also be an objective to generate custom territories where the homogeneity index is above some certain threshold (e.g. all units must have a homogeneity index above 0.6).

1.2.2. EXAMPLE CALCULATION OF H-INDEX

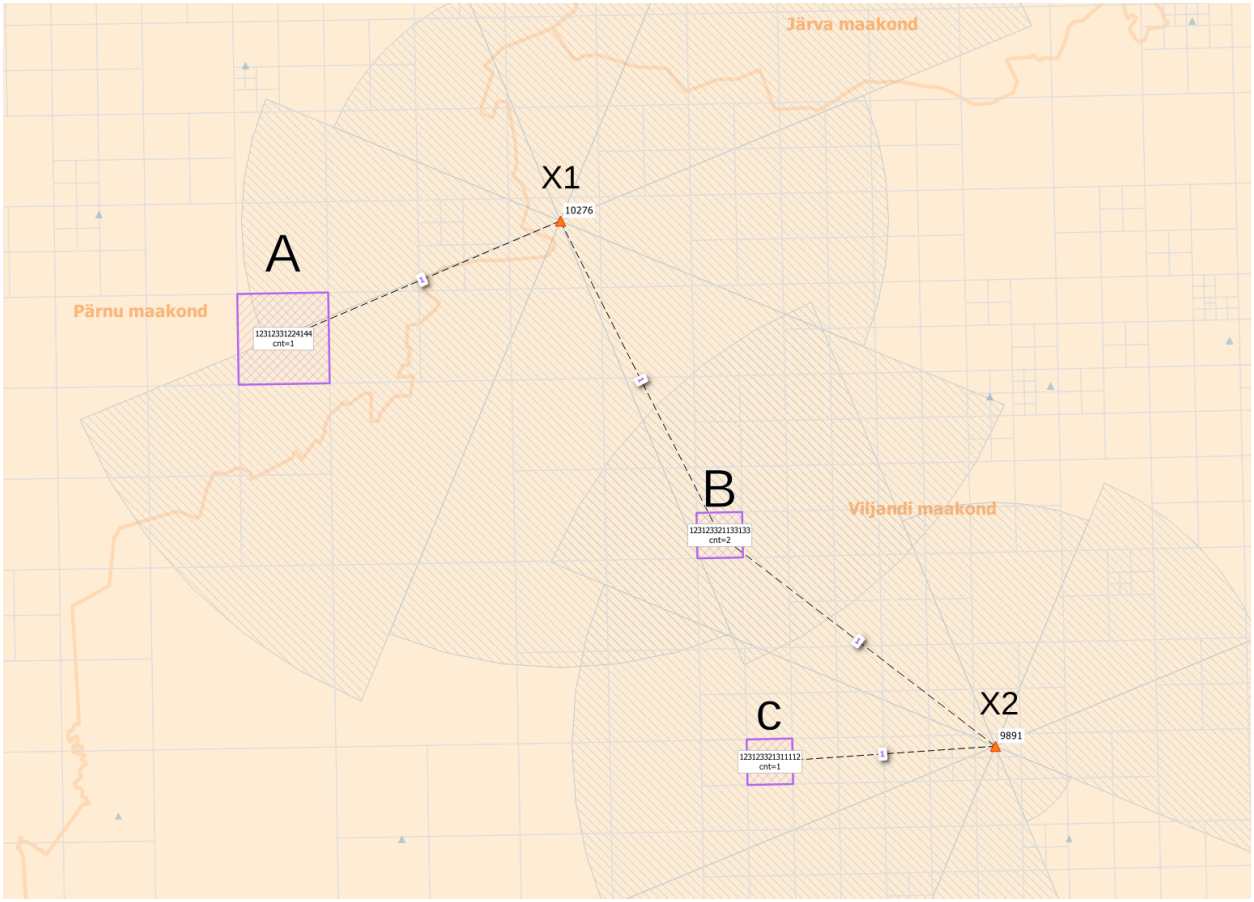

Figure 1. Example data for h-index calculation ***(M. Tiru, personal communication, May 2021)***

The example (see Figure 1) illustrates the situation where there are 4 stays in 2 cells (cell_id **X1** and **X2**) interpolated to 3 grid squares. Notice that one grid square (**B**) has been interpolated from both cells and houses 2 stays. Based on this example, the homogeneity index for two LAU in the illustration (Pärnu county and Viljandi county) are:

Pärnu county h-index = 1 / 2 = **0.5**.

There is one stay in a grid square in Pärnu maakond in grid square **A** which was interpolated there from cell_id **X1**, so the dividend in this equation is 1 (stay in Pärnu maakond). From cell_id **X1** there are a total of 2 stays (one in Pärnu county in grid square **A**, another in Viljandi county in grid square **B**), so the divisor is 2.

Viljandi maakond h_index = 3 / 4 = **0.75**.

There are three stays in grid squares in Viljandi maakond: two in grid square **B**, one in grid square **C**, which were interpolated there from cell_id **X1** and cell_id **X2**, so the dividend in this equation

is **3** (stays in Viljandi maakond). From cell_id **X1** there are total of 2 stays (one in Viljandi maakond in grid square **B**, another in Pärnu maakond in grid square **A**), and from cell **X2** there are 2 stays, both in Viljandi maakond in grid square **B** and in grid square **C**), so the divisor is **4**.

2. METHODS

The input data for this methodology are events from mobile positioning data which have been interpolated to an adaptive grid. Input data has already been calculated and will serve as a starting point on which the methodology is built upon. Stay sections are based on Positium's methodology that involves processing mobile positioning data. The methods section is structured as follows: 1) the source data and its origin will be described. 2) Exploratory spatial data analysis (ESDA) tasks with the source data will be explained to better understand underlying patterns that might influence the main part of the thesis - i.e. 3) the zoning algorithm, which will be described.

2.1. POSITIUM EVENT GRID DATA

2.1.1. MOBILE NETWORK CELLS

Location and coverage area of the mobile network cells are the basis for estimating the location of the MPD records spatially. Technically, all network events and operations (calls, messages, internet connection, etc.) are linked to at least one specific cell at the moment of the event. This event, represented by a record in MPD, therefore, takes place within the coverage area of the cell. All cells have a physical coverage area where mobile devices can connect and communicate to the mobile network. The base for MPD methodology is the ability to factually deduce that a specific mobile device was present within the coverage area of a mobile network cell at a specific moment.

2.1.2. CELL COVERAGE

Mobile network antenna' signal spread over a geographical area, which is called coverage area. Some antennae are omni-directional (i.e. they are a circular antennae and spread around themselves), some are directional. Network cells are mostly distributed unevenly as more clients are in urban areas and coverage in rural areas is generally sparser with fewer cell towers. The coverage area of cells varies from a couple of hundred meters in urban areas to up to 35 kilometers in rural areas (Sauter, 2011). Because network cells' coverage area is usually not provided by Mobile Network Operators (MNO), and instead only location of the antennae is received, the coverage area of the cells has to be constructed based on the available information and best available knowledge and assumptions.

Consecutive records between two or more cell towers within a specified (short) timeframe by the same subscriber is called oscillation. Oscillation might occur for example due to mobile network load balancing, buildings blocking signals, poor weather conditions and other signal altering circumstances. Those events might force a handover between BTS without the device moving in real life (Pukk, n.d.) Ideally cell towers that are directed towards each other and are located close to each other have higher oscillation count than cells that are further away from each other. Cells that are further apart should not have no oscillation at all. Overlapping coverage areas of cells highly affect oscillation between them, providing useful information for creating a coverage area in case when MNOs cannot provide the coverage area information for cells, and to detect cells whose locations are incorrectly provided by MNOs.

2.1.3. MOBILE POSITIONING DATA

MPD data consists of the records stored by the MNO in MNO databases representing the factual activity of the mobile subscriber in their, or their roaming partners' networks. Each record represents some kind of mobile network event, like initiation or termination of a mobile phone call, sending or receiving a message (SMS, MMS), sending or receiving internet packages, connecting a mobile device to the network, change of the location area within the network, handover of the antenna, etc. There is no finite list of the events that can and is stored by a specific MNO. The more events are stored, the denser (and thus higher quality) data this is. Each event (with some exceptions) should be treated as a fact that the specific mobile subscriber was present at this specific location (either within the coverage area of a specific mobile network antennae, or in a foreign country). The frequency of previously mentioned events from phone users is reflected on the CDR dataset. The location of the phone at the time of the event is automatically recorded as a by-product for billing purposes of the MNO. This method of data collection is known as passive mobile positioning (Ahas et al., 2008).

The frequency of CDR data depends directly on how often the phone user engages in call activities (calls in and out, SMSs in and out). In the collection of CDR data, the mobile operator is not positioning the phone's location *per se*, rather the location is automatically recorded as a by-product for billing purposes in the systems of mobile network operators. CDR data usually consists of a timestamp, device's unique id and BTS id.

The CDR dataset used in calculations is from one operators in Estonia consisting of 1.45 million unique subscribers, 816.11 million domestic and 5.71 million inbound records. The temporal range of the dataset is 5 consecutive months in 2019 and covers the entire country for that time period. A subset of data has not been chosen. This way the results would describe the average situation in Estonia. A sample of CDR data can be seen in Table 1, where "positoning_id" is an identifier given to the event, "mobile_subscriber_id" is the unique pseudonymous id given for the calculation, "positioning_time" is the time in datetime format, "cell_id" is code for BTS (Ahas et al., 2007; Wang & Chen, 2018) . Real locations for BTS are known.

Table 1. A sample of CDR data

positoning_id	mobile_subscriber_id	positioning_time	cell_id
22323675	1416992316447219200	2019-01-01 11:15:07+02	1
21712446	1416992316447219200	2019-01-01 11:30:58+02	1
20073248	4035464122881259072	2019-01-01 18:16:08+02	1
22750769	4035464122881259072	2019-01-01 11:45:13+02	1
24857987	2711886842895455104	2019-01-01 16:31:48+02	1

2.1.4. ADAPTIVE GRID

Methodology for spatial interpolation of inbound roaming and domestic MPD from mobile network cells uses adaptive grid (a.k.a. Morton order, Z-order curve) (Morton, 1966). Adaptive grid is based on the Euclidean space grids in the designated Cartesian coordinate system. Each grid square can be divided into four equal-size grid squares. The nomenclature of the grid squares is rule-based, and each smaller grid includes the parent's grid id (see Figure 2). The grid should cover all mainland areas of the country of reference with grid squares that are different in their sizes. The logic behind the size of the grid squares is that smaller-sized grid squares are located in places where there are more people, and the network cell distribution is dense (see Figure 3).

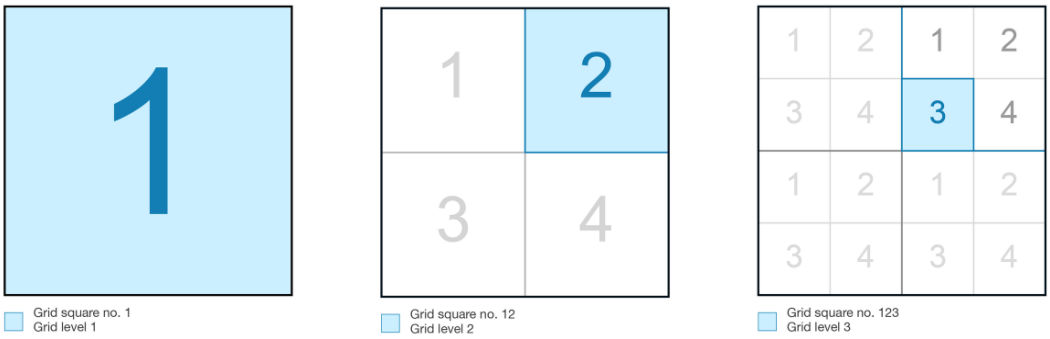

Figure 2. Grid naming nomenclature ***(M. Tiru, personal communication, May 2021)***

Positium uses road network reference data to "split" grid squares to smaller units based on the number of road intersections within the grid. This is done because road density corresponds to population density. The process of generating adaptive grid for a country of reference uses following steps:

- Generate minimum level grid (i.e. largest grid squares) covering whole land area of the country of reference based on LAU reference data (i.e. all country should be covered by minimum level grid squares);
- For each grid square, calculate the number of road sections intersecting the grid. If the number of road section is higher than the threshold, split grid square into four smaller grid squares;
- Continue with pt 2 until maximum grid level is reached.

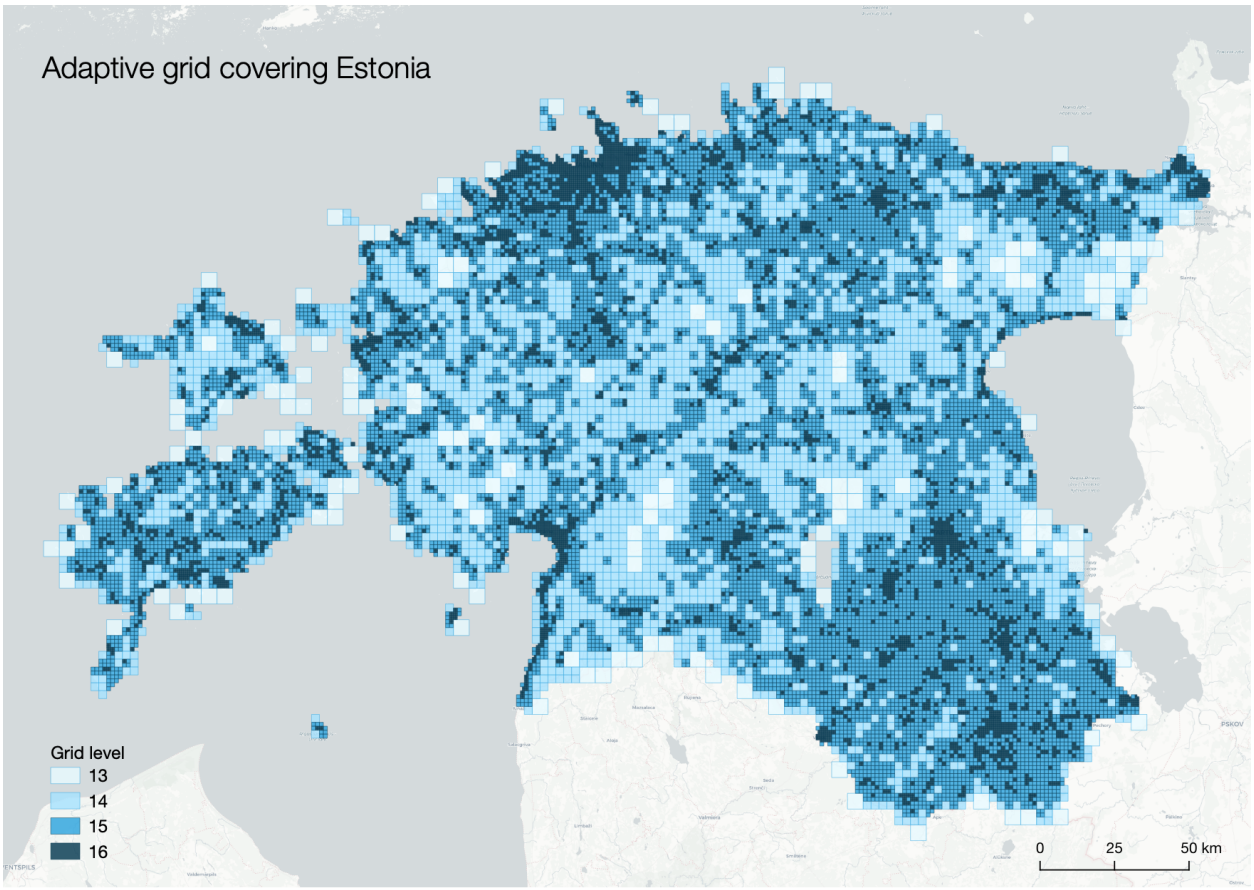

Figure 3. Adaptive grid Covering Estonia ***(M. Tiru, personal communication, May 2021)***

The spatial projection used for the adaptive grid is Pseudo-Mercator (epsg: 3857), the grid is based on the bounding box of the world (i.e., grid square #1 on level 1 covers the whole World). Because of this, the size of the grid square on the same grid level is different depending on the latitude. For example, grid square width on level 16 covering Estonia is roughly 641 m, in Oman, the same level grid is 1116 m and in Indonesia, 1225 m. Deciding what min and max grid levels to use in a specific country depends on several aspects. For example: what is the spatial precision necessary for results. If final results are presented on very high LAU level (e.g. regions of the country - LAU 1, 2), then max grid level should be rather low, so there are fewer grid squares (e.g. less storage and calculations required), the interpolation from cell to grid will take less time, the precision will be low, but accuracy will be high (the larger the grid square, the more probably the specific subscriber has actually been present in the grid square). It must be noted, the min and max grid levels should take into account that they are smaller than the smallest spatial unit in results. Otherwise, if adaptive grid squares are larger, or many of them are overlapping several resulting

spatial objects, the accuracy caused by aggregation from grid to polygon will decrease heavily. The grid level used in calculations for this thesis is 13. The side length of a level 13 grid in Estonia is roughly 5km.

2.1.5. STAY AND MOVE SECTIONS

The smallest spatio-temporal elements of the continuity model are stay and move sections. Stay and move sections constitute trips.

Stay sections are time periods during which a person is assumed to be present in a specific location. In the PDM continuity model, this location is represented as a grid in case of domestic and inbound data, and a country in case of outbound data. So, a stay section represents a person being in a location for a specific duration of a time.

Stay sections are constructed based on the subscriber being present in overlapping cell coverage areas using the following logic:

1. MPD record in cell A - if cell A's coverage area is not overlapping previous record's cell's coverage area, then this is the beginning of a new stay (stay 1) with the initial cell A;
2. Next MPD record in cell B. If cell B's coverage area is overlapping the last started stay's (stay 1) initial cell's (cell A) coverage area (overlap area>=20%), then record in cell B is a part of stay 1, otherwise it will become a new stay 2 with initial cell B;
3. All consecutive cells belonging to the same anchor point (Ahas et al., 2010) form one stay;

2.1.6. DATA INTERPOLATION

Because MPD records are spatially referenced to cells the initial data is geographically aggregated to cells. However, obviously, people are not present only in the exact locations of the cells but are rather spread around the country. Because of this, MPD cell-based data requires spatial interpolation over the land area of the country of reference. A subscriber linked to a cell can be present anywhere within the coverage area of the cell. Therefore, spatial interpolation of MPD to land is necessary. This is done using the coverage areas of the cells and adaptive grid.

2.1.6.1. INTERPOLATION TO ADAPTIVE GRID

Stays constructed from inbound roaming and domestic data are interpolated to an adaptive grid. This process is done using the grid and coverage area probabilities where each cell is linked to a

number of grid squares with specific probability that MPD records are assigned to the grid square (see Figure 4). This probabilistic interpolation ensures distribution of stays over the geographic area based on the best assumptions that are available, because there is no factual information in what exact location the subscriber was present within cells' coverage area. So, individually, a subscriber's location is probably not precise, but statistically, this distribution should provide an adequate model of people being present in grid squares all over the country.

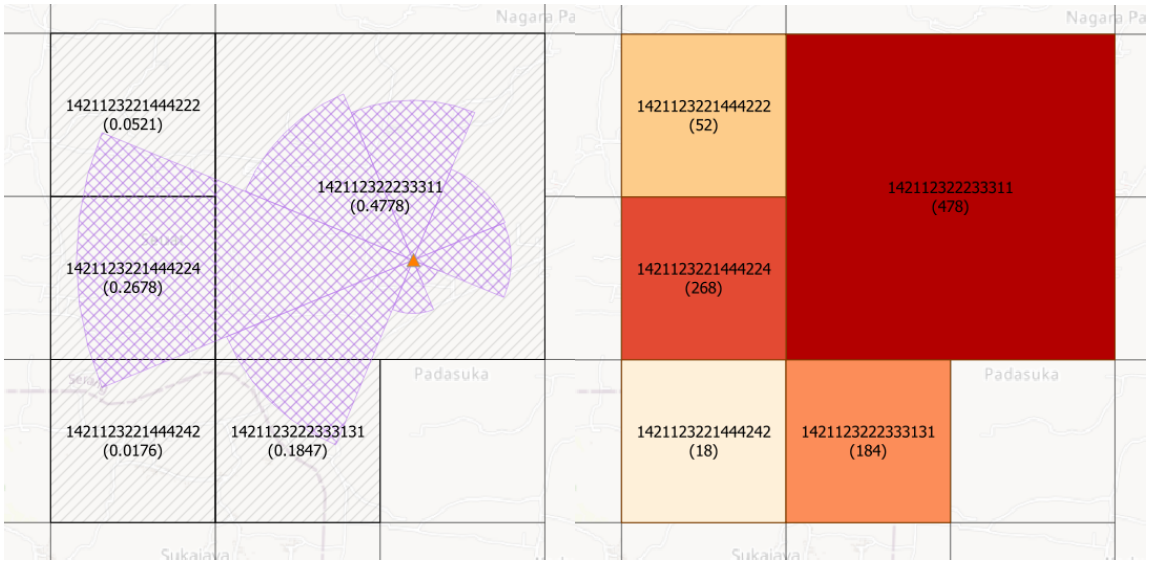

Figure 4. Interpolation from cell to grid ***(M. Tiru, personal communication, May 2021)***

This is the input data upon which the zones will be created. Since the geometry is given thus realignment of the grid squares with other grids such as the Estonian national grid is not in the scope of this thesis however this can be considered in the future when developing the methodology further.

2.1.6.2. EXAMPLE OF INTERPOLATION TO GRID

There are two sample cell towers A and B (see Figure 5). There is one stay in cell tower A and two in cell tower B. Two grid squares have been selected for interpolation (see Figure 6). Stays have been interpolated to grid squares (see Figure 7).

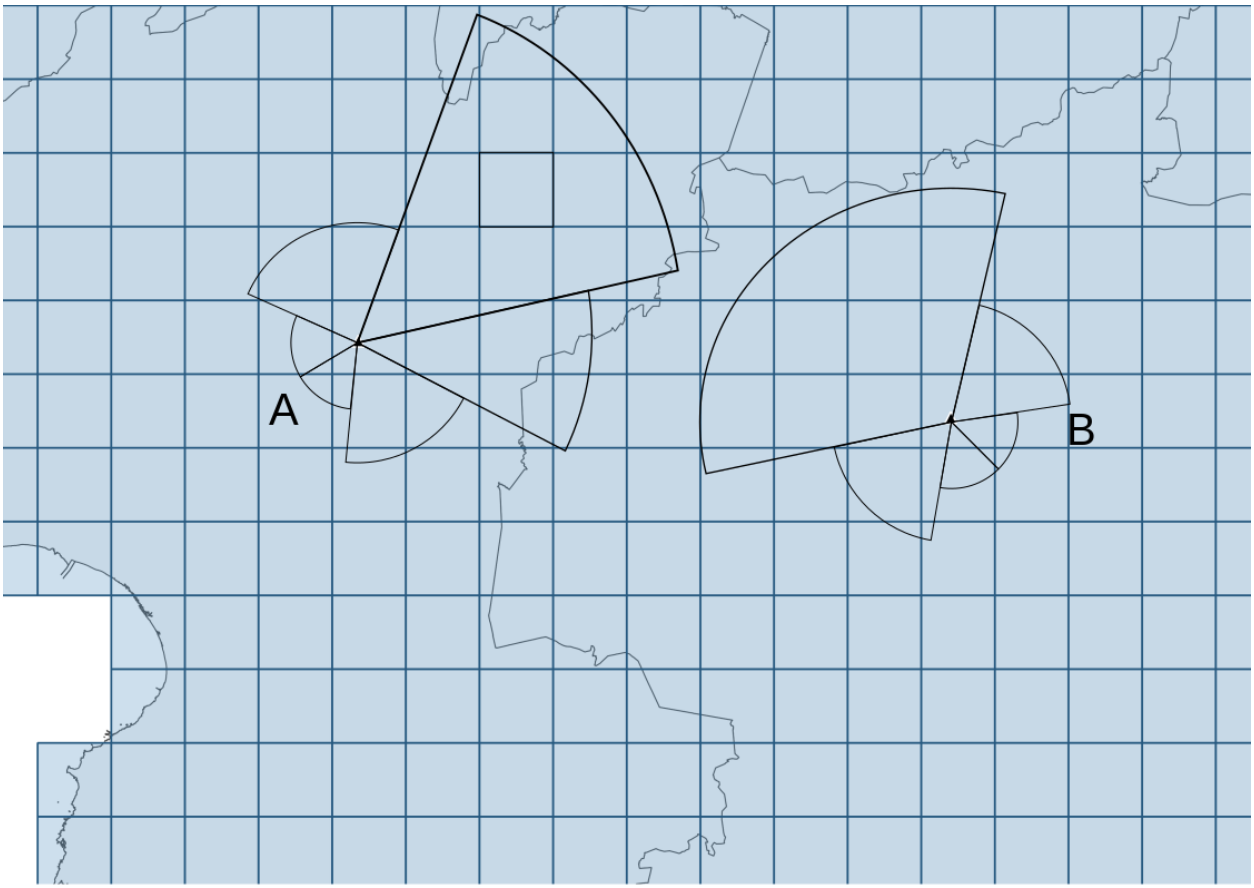

Figure 5. Sample areas of two cell tower masts

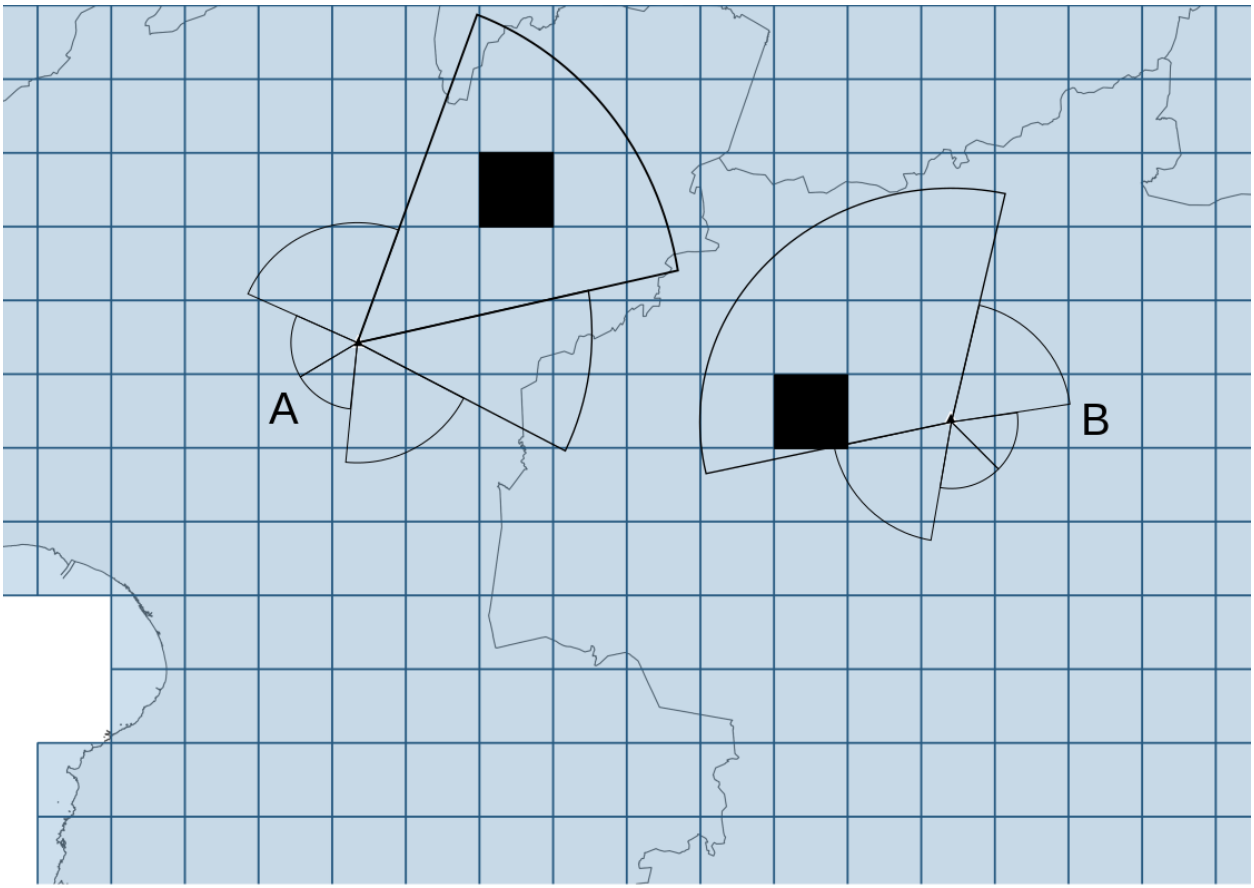

Figure 6. Spatial interpolation of 3 stays in example towers A and B

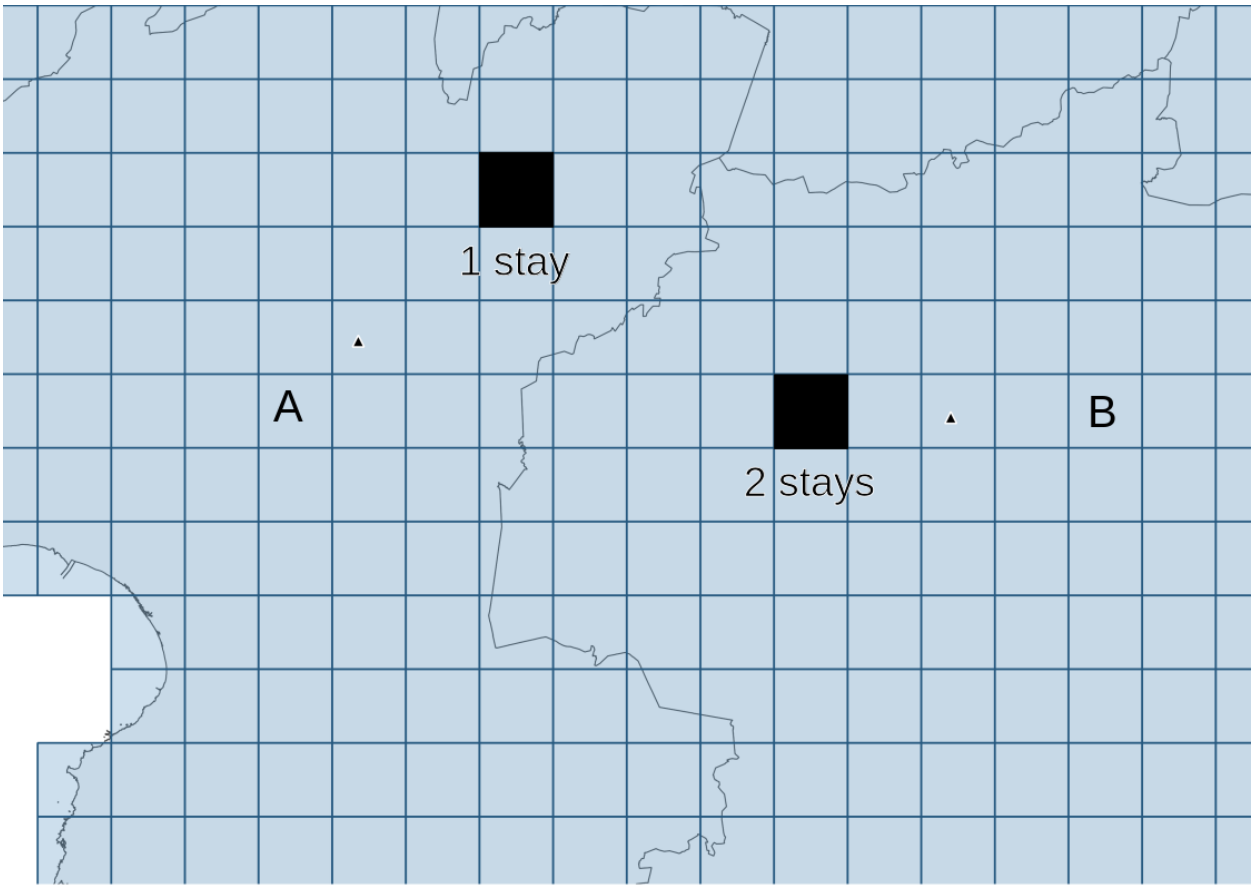

Figure 7. Locations of interpolated stays

2.2. INITIAL SPATIAL DATA ANALYSIS OF THE INPUT DATA

The input used for creating a zoning algorithm consists of BTS identifiers, grid squares to which stays have been interpolated from the BTS using the Positium methodology and the number of stays interpolated to the specific grid square. The grid level used here is 13. Level 13 grid was chosen in favor of higher level grids or multiple grid levels to simplify the zone generation by the algorithm and to reduce calculation times. Level 13 grid has 1929 units in total. A sample of the data can be seen in Table 2.

Table 2. Sample of input data

bts_identifier	grid_id	stay_cnt
4835	1231231431423	1
1703	1231231341232	6

When summing up all of the stays in the grids the outcome resembles the population density map of Estonia (see Figure 8Figure 9). The population density map was created using statistics Estonia 1 x 1 km grid (*VKR*, n.d.).

Maps of stays interpolated to grids with equal count in classes using 5 and 7 classes.

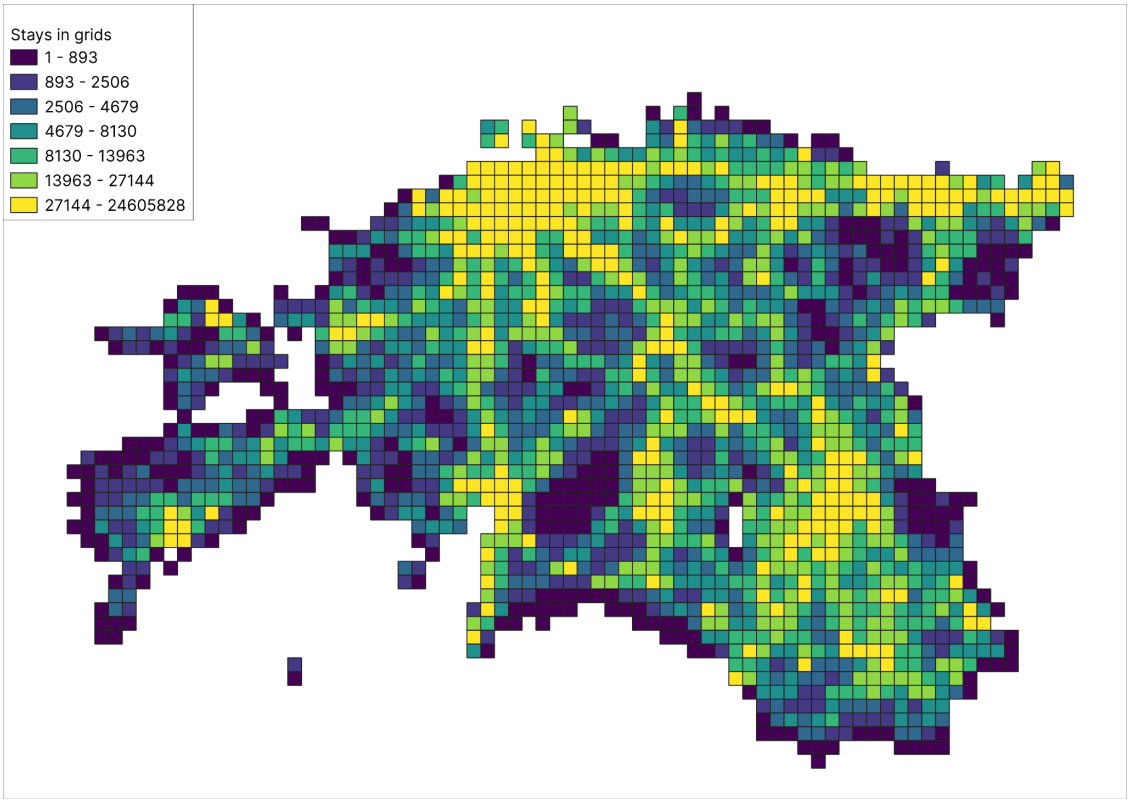

Figure 8. Aggregated stay count per grid squares

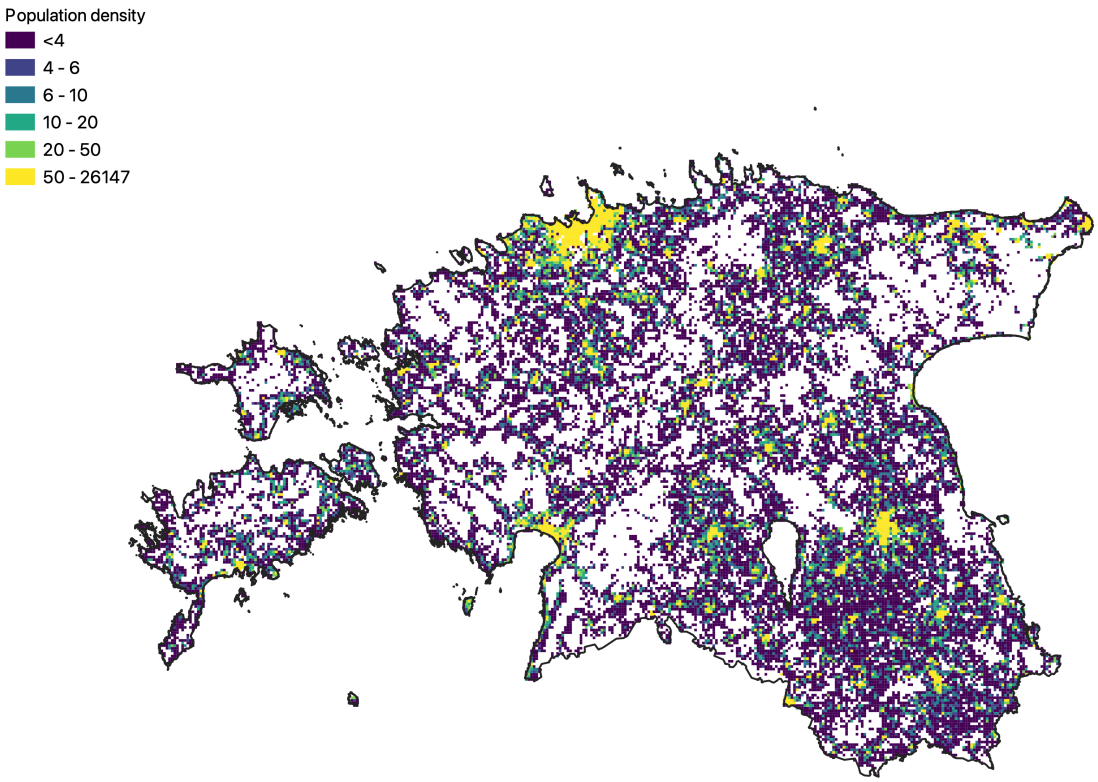

Figure 9. Population density using Statistical department of Estonia data

2.2.1. SPATIAL AUTOCORRELATION FOR POSITIUM INTERPOLATED EVENT GRID (PIEG)

The term "spatial autocorrelation" refers to the property of the spatial variable that exhibits systematic dispersion or segregation in space. A spatial variable is said to be spatially autocorrelated when neighboring locations or nearby areas exhibit similarity in terms of characteristics or magnitudes for a particular attribute or variable of interest (Anselin, 1995). Spatial autocorrelation measures the correlation of a variable with itself across space. Positive spatial autocorrelation means that the locations close together have similar values, while negative spatial autocorrelation means that locations close together have more dissimilar values than those locations further away.

Visual inspection allows for searching patterns in the map. If the spatial distribution on stays is random there should not be any clustering of similar values. However, we can see clear clustering of a high number of events in more populated areas near county centers. To eliminate human error

which can lead to detect false positives and statistical patterns where there are none spatial autocorrelation will be done. Visually detecting false positives is more common with differing polygon sizes which is not the case here.

Python library PySAL was used to generate the two similarity measures (Rey & Anselin, 2010).

2.2.2. GLOBAL SPATIAL AUTOCORRELATION

Spatial lag is a variable that averages the neighboring values of a location (*Documentation | GeoDa on Github*, n.d.). So the spatial weight between grid squares i and j indicates if the two are neighbors (i.e., geographically similar) (see Figure 10). There is also a need for attribute similarity to pair up with spatial similarity. The spatial lag is a derived variable that accomplishes this. For grid square i the spatial lag is defined as:

$$ylag_i = \sum_j w_{i,j} y_j$$

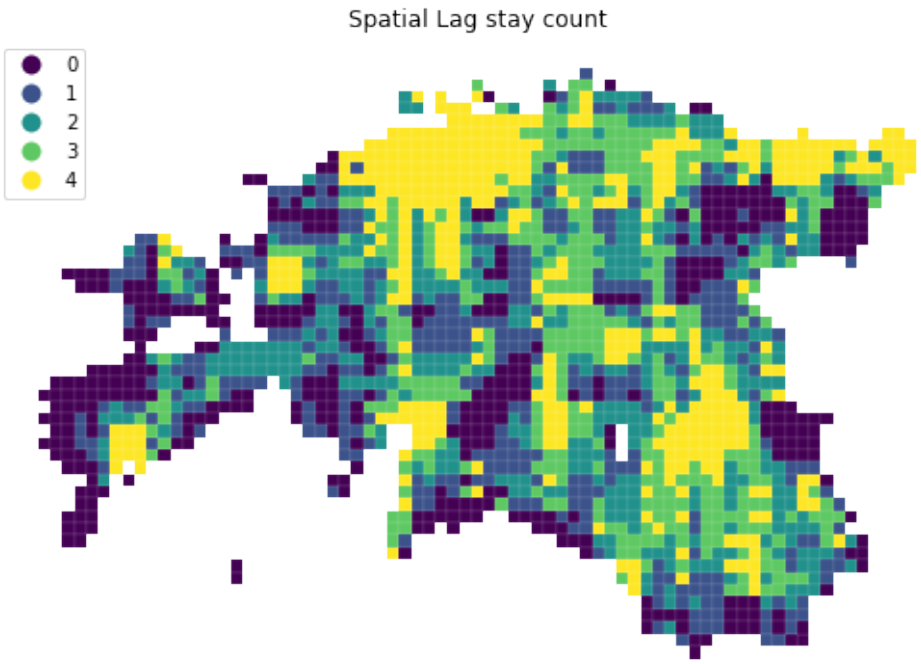

Figure 10. Spatial lag of aggregated stay counts per grid squares

The quintile map for the spatial lag tends to enhance the impression of value similarity in space.

However, we still have the challenge of visually associating the value of the number of stays in grid cells with the value of the spatial lag of values for the focal unit. To get a better understanding of the underlying patterns within the data further analysis must be done.

The data was split into two groups based on the number of stays in grid squares. The results were plotted on a two-color cartogram (see Figure 11). If this kind of map does not have many changes of color between neighboring grid squares, then we can say that there is a positive autocorrelation. If the color between neighboring grids changes often then we can say that there is a negative autocorrelation. If there are no regularities between the changes between the neighboring colors, we can say that the spread of values over the grid is spatially random and there is no autocorrelation.

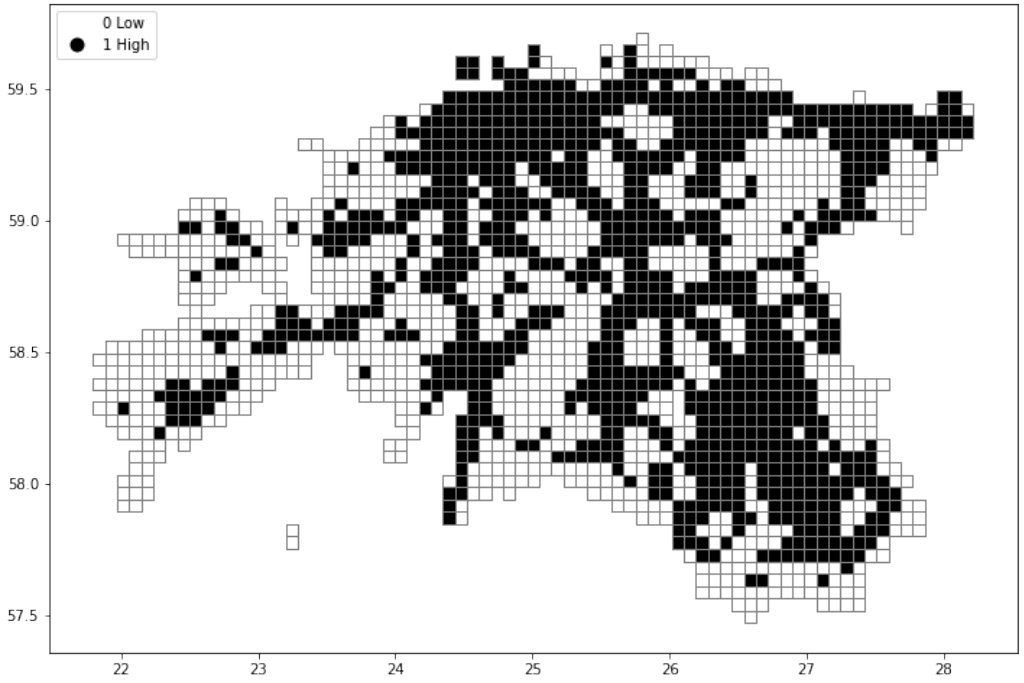

Figure 11. Binary distribution of high and low values in the input data

The median amount of stays in grid cells is 6072 and there are 959 grid squares above median out of 1919. Note this is 10 less grids than there are in level 13. Those 10 grids did not have any data interpolated to them and cannot be used in the calculations.

One way to formalize a test for spatial autocorrelation in a binary attribute is to consider the so-called *joins*. A join exists for each neighbor pair of observations. Each unit can take on one of two values "Black" or "White", and so for a given pair of neighboring locations there are three different types of joins that can arise:

- Black, Black (BB)
- White, White (WW)
- Black, White (or White, Black) (BW)

For our dataset we have 1101 BB joins, 1049 WW joins and 755 BW joins. To determine if the distribution of stays in grids is random or not, we need to check the null-hypothesis (complete spatial randomness) against our data.

Python library esda was used to carry out join analysis.

The join count analysis is based on a binary attribute, which can cover many interesting empirical applications where one is interested in presence and absence type phenomena. Spatial autocorrelation is multi-directional and multi-dimensional, making it useful for finding patterns in complicated data sets. It is similar to correlation coefficients; it has a value from -1 to 1. However, while other coefficients measure perfect correlation to no correlation, Moran's I is slightly different:

- -1 is perfect distribution of dissimilar values (perfect dispersion).
- 0 is no autocorrelation (complete spatial randomness).
- +1 indicates perfect clustering of similar values (perfect segregation).

Moran's I value for this dataset is 0.32129742478964. Which indicates that it has slight positive clustering.

Using Moran's tests I calculated Moran's I value and interpreted it against a reference distribution under the null hypothesis of complete spatial randomness using PySAL with 999 permutations (see Figure 12). Our observed value is in the upper tail. This shows that our observed Moran's I value is significantly higher from complete spatial randomness, and we can reject the null hypothesis in favor of spatial autocorrelation in stay counts in grid squares.

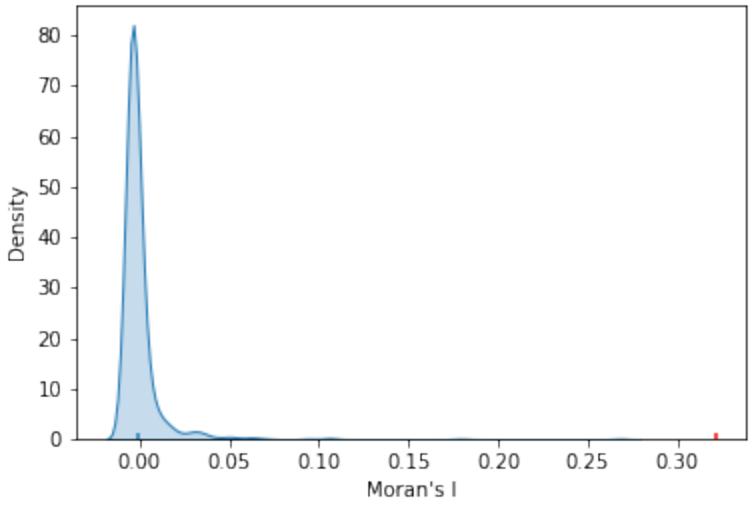

Figure 12. Observer value against complete spatial randomness, the red marker on the right side

2.2.3. LOCAL AUTOCORRELATION: HOT SPOTS, COLD SPOTS, AND SPATIAL OUTLIERS

Hot spots are the clusters of high values. In the case of PIEG there are 9 out of which 8 are county centers with the addition of another hot spot centered around Kohtla-Järve (see Figure 13). This information will be useful in the zoning algorithm section

Cold spots are clusters of low values. In the case of PIEG the number of cold spots is a lot larger than the number of hot spots. The cold spots are mostly located on border areas or areas with lower population density (See Figure 13). Visually located more in western Estonia. This information will be useful in the zoning algorithm section

Doughnuts are outliers where lower values are surrounded by higher values. In my case most reside next to county centers or hot spots with a larger cluster of doughnuts in north-eastern Estonia (See Figure 13). In the case of the algorithm zones starting in donuts have a higher chance of finding additional grid units to their zone. But at the same time doughnuts will be chosen less times since the algorithm prioritizes grids which increase h-index more.

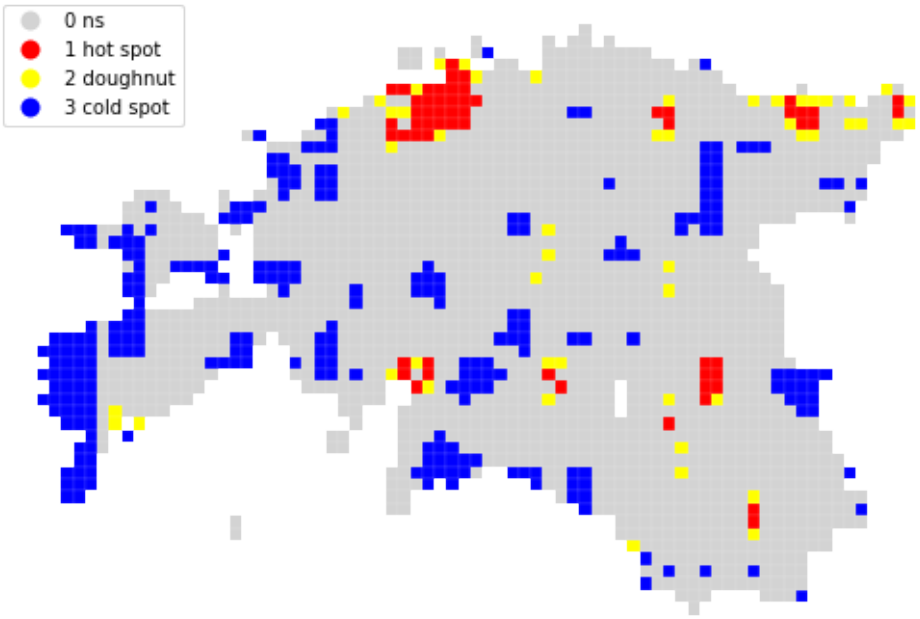

Figure 13. Hot spots, cold spots and spatial outliers for stay counts in grid squares

2.3. ZONING ALGORITHM

To create custom zones based on the input data several algorithms will be designed. Some important properties for these zones … one is that zoning will need the information of h-index. The following sections will at first introduce the algorithms used to create zones, then how h-index is used in building and evaluating zones, and then different other criteria, i.e. seeding and termination/convergence criteria.

2.4. BUILDING ZONES

2.4.1. PSEUDO ALGORITHM DESCRIPTION

The algorithm grows zones from seeds on a turn basis. A seed is the starting location of the zone generation. The seed can be a single location or a group of locations. Turn based zone growth means that in each cycle of the algorithm all of the initial zones (locations specified in seeds) have the possibility to grow by one grid. When a grid is added to a zone it is deducted from the list of available grid squares.

Algorithm description:

Prerequisites

- Choose a target number of initial zones and create a seed for each of the zones
- Create list of all grid squares that are available

Iteration for each zone:

- Create list of available grid squares directly adjacent around the current zone
- Calculate h-index values for possible combinations with full zone + each single one grid from the previously generated lists
- Take the grid which increases the h-index most and add it to the zone
- Remove added grid from overall pool
- Check termination criterion/break condition for zone
- Next iteration

The algorithm has three major parameters which change the outcome. The first is initial seeds. This determines the starting locations for the zones to grow. The second is distance from seed which limits how far the zones can expand from the initial location. The distance parameter is there to limit the generation of too large zones to avoid the edge-cases. Third is the threshold of h-index, under which the zone should not fall. A minor fourth input is the order of input seeds, because the algorithm starts growing the zones in the order of the input.

When going through the iterations in the algorithm, zones have two break conditions which finalize them and take them out of the calculation stage. The first break condition is when there aren't any more candidates (i.e. grid squares available) to increase the zone. This may happen when there is another zone nearby or it has grown to its limit size from the initial location. The second break condition is when the best candidate does not increase the h-index anymore. Then the algorithm checks if the new h-index of the zone with the best candidate is added would fall below the minimal threshold specified in the (input) parameter. Then there are two options, if the new h-index is above the minimal threshold then it gets added to the zone. When the best candidate would bring the h-index below the minimal threshold then no candidates are added and the zone is finalized.

Finalization/ Post-processing/Borderline cases

When all of the initial seed zones have been finalized there is a possibility that there are some leftover grid squares which have not been assigned to a zone yet. If all the initial zones have reached their conclusion the algorithm will choose the grid with the highest h-index from available grid squares and build zones based on the input parameters until its conclusion. This process will continue until there are no more available grid squares (see Figure 14).

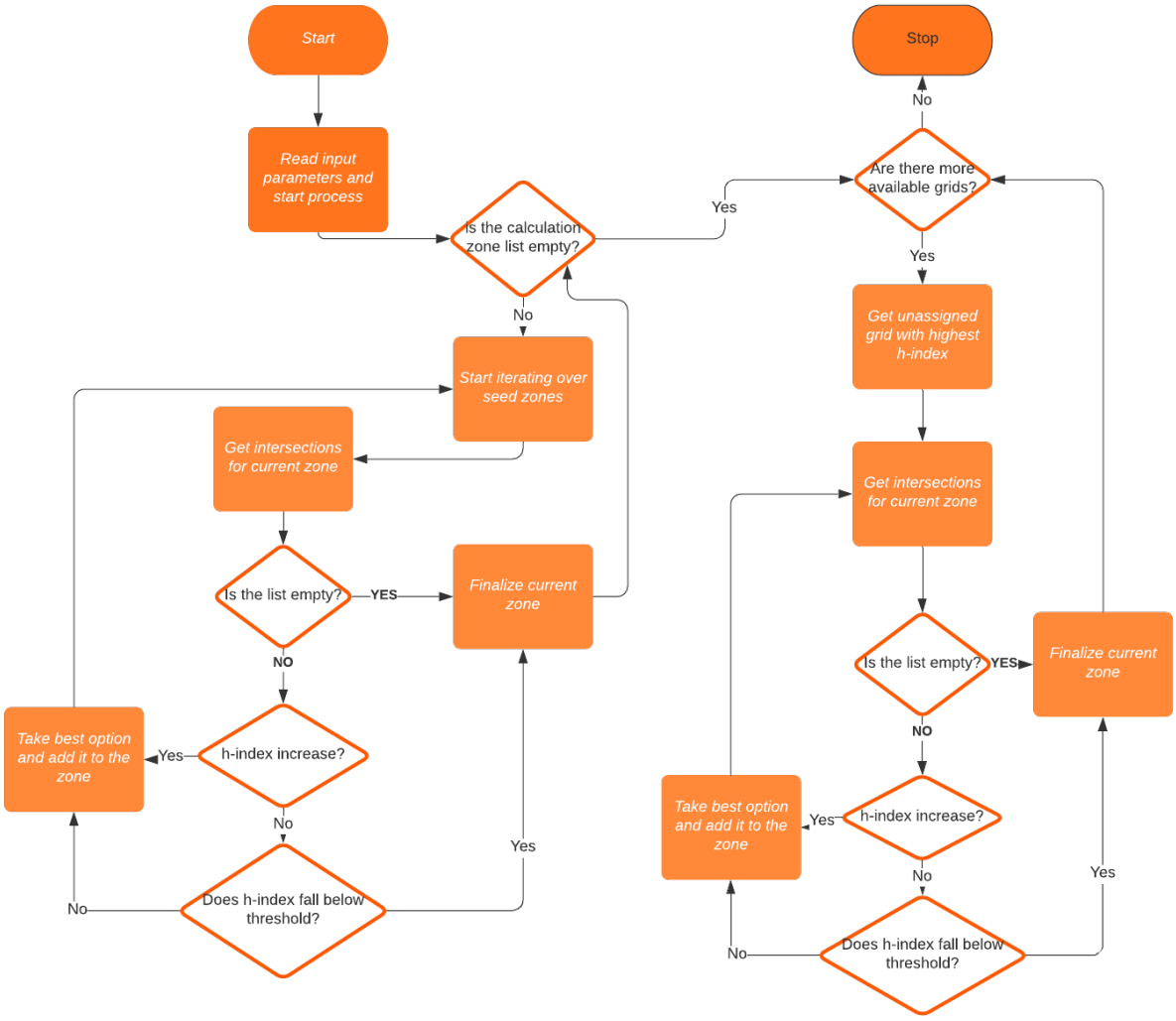

Figure 14. Algorithm process

2.4.2. EXAMPLE OF CALCULATION IN ALGORITHM

In the algorithm the base seed location h-index is calculated. For example, the base seed in Kuressaare would be here (see Figure 15).

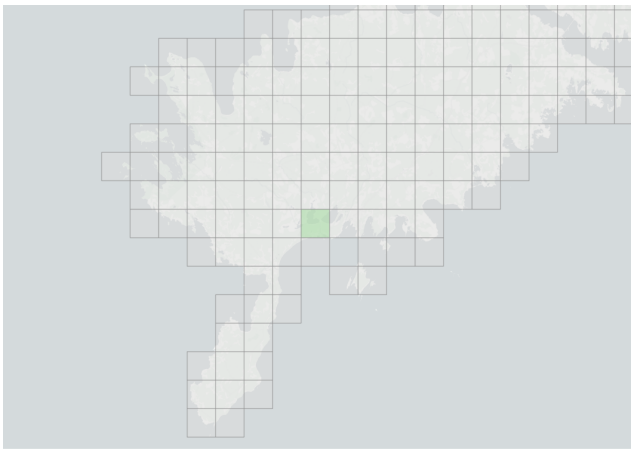

Figure 15. Starting location of zone generation

Then the algorithm gets a list of grid squares adjacent the initial seed location (see Figure 16).

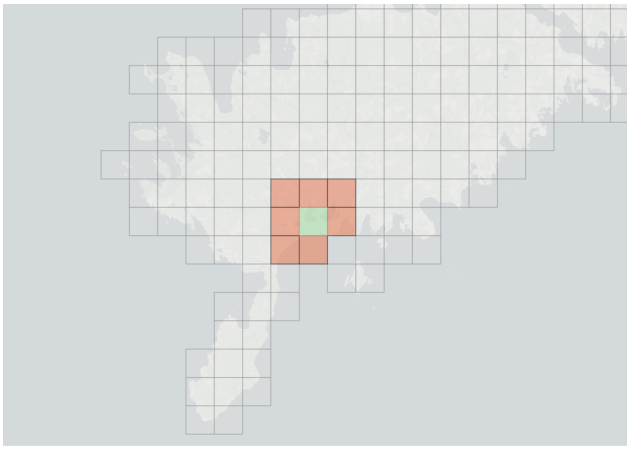

Figure 16. Candidates for the initial zone

For each one of the adjacent grid squares the algorithm calculates the theoretical resulting h-index if it would be joined with the zone. It then compares the highest calculated value with the current h-index of the zone. If the resulting h-index would be higher or least not under the specified threshold the grid is added to the zone and the new h-index is assigned (see Figure 17).

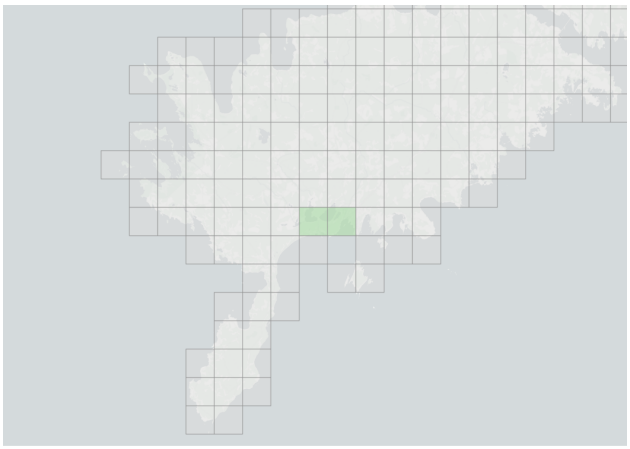

Figure 17. Updated zone

Now the zone has increased in size and the process will repeat.

2.5. HOW TO MEASURE RESULTS

One way to measure the results of the algorithm would be clients' feedback and if they like the new zones, this may cover topics such as similarities to current zones or aesthetically pleasing layouts. However, this is not a good measure as it is highly subjective and hard to back by data.

Ideally the end result would have a minimal number of zones with a high h-index but this would bring the edge case of having one zone as the whole country and having an h-index of 1. This would mean that decision making based on one zone would be useless. Therefore, ideally the minimal number of zones would be over 1. Another edge case may be a really high number of zones which would cut otherwise bigger zones into smaller chunks.

One measure that a zone might look for is compactness. Polsby-Popper and Schwartzberg measure the indentation of the zone, while Area/Convex Hull and Reock measure the dispersion of the zone. The measures can vary from 0 to 1 where zones having values that are 1 or close to it being the most compact. Since the zones are made out of square grids it is impossible for them to reach a score of 1 on most measures.

- Polsby-Popper - The Polsby-Popper measure is the ratio of the area of the zone to the **area** of a circle whose circumference is equal to the **perimeter** of the zone.
- Schwartzberg - The Schwartzberg score is the ratio of the perimeter of the zone to the **circumference** of a circle whose area is equal to the **area** of the zone.

- Area/Convex Hull - The Area/Convex Hull score is a ratio of the **area** of the zone to the **area** of the minimum convex polygon that can enclose the zone's geometry.
- Reock - The Reock score is a measure of the ratio of the **area** of the zone to the **area** of the minimum bounding circle that encloses the zone's geometry.

For reference the h-index will also be calculated for the LAU's. To calculate h-index for LAU levels it is necessary to assign grid squares to LAU-s. In Positium methodology the LAU that is chosen for a grid depends on the proportion of the grid that is covered by specific LAU-s. If a grid is entirely covered by one LAU then it is assigned to that specific LAU. This process is done while generating the adaptive grid layer (M. Tiru, personal communication, May 2021). To calculate h-index for a LAU all of the grid squares are grouped by LAU-s and follow the general h-index calculation methodology. This allows the comparison of newly generated zones and LAU's.

2.6. EXPERIMENTAL SETUP

Two parameters with different combinations were used for inputs to see how different parameters change the outcome of the algorithm. A single seed group was used as the starting point of the algorithm. The initial seed group was based on the county centers in Estonia to simulate growth of zones from the current administrative unit centers. The 15 initial seeds in the seed group were ordered in the list alphabetically. The use of a single seed group allows comparison of the different parameters and how they change the outcome of the algorithm.

There are three groups of parameter combinations. Each group is using the same distance to the seed location. The distances used were 30, 50 and 70km. The h-index threshold parameters chosen are starting from 0.5 and increased with a step of 0.05 until the threshold was 0.9. So altogether 27 instances of zoning output were created. The combinations can be seen in Table 3.

Table 3. Experimental setup

Distance from seed	30 km	50 km	70 km
Min h-index	0.5	0.5	0.5
	0.55	0.55	0.55

	0.6	0.6	0.6
	0.65	0.65	0.65
	0.7	0.7	0.7
	0.75	0.75	0.75
	0.8	0.8	0.8
	0.85	0.85	0.85
	0.9	0.9	0.9

The algorithm is deterministic and thus produces the same zoning if the same input parameters were chosen. Therefore, 27 different realizations were created and stored in JSON format for subsequent assessment and interpretation.

3. RESULTS

3.1. CALCULATED H-INDEX FOR ESTONIAN LOCAL ADMINISTRATIVE UNITS

For the whole dataset the main outlier was Ruhnu island which had a h-index of 1 because the cell towers from mainland and other islands do not reach there. The LAU 2 (municipality) and LAU 3 (settlement) level cover the entire island combined with no coverage area overlapping from different LAU units creates a scenario where the island would always have h-index of 1.

A histogram aims to approximate the underlying probability density function that generated the data by binning and counting observations. Kernel density estimation (KDE) presents a different solution to the same problem. Rather than using discrete bins, a KDE plot smooths the observations with a Gaussian kernel, producing a continuous density estimate:

H-index was plotted using Kernel Density Estimation algorithm to smooth histograms

LAU 3 level units had one outlier Ruhnu. The mean h-index was 0.007 and a median of 0.003. This was to be expected, since most cells cover multiple LAU 3 units, often even more than 5. Units with h-index higher that 0.6 were Ruhnu küla, Narva, Tartu, Pärnu (see Figure 18).Some LAU level 3 units did not get any h-index value assigned to them because grid level 13 sometimes covers more than one unit.

LAU 2 units had a mean h-index of 0.158 and median of 0.077. There were also 3 groups. Group 1 had an h-index of 1-0.69, group 2 had 0.46-0.25, group 3 had 0.15-0.01. Group 1 consisted of cities and island municipalities, group 2 had cities and municipalities which had larger populations, group 3 was made up of all the other municipalities with lower h-index scores (see Figure 18).LAU 1 level units had a mean h-index of 0.423 and median of 0.321. 6 counties had a h-index higher than 0.5 which is a baseline for a desired h-index value. 4 counties had h-index values below 0.21 with the lowest being Põlva county (see Figure 18).

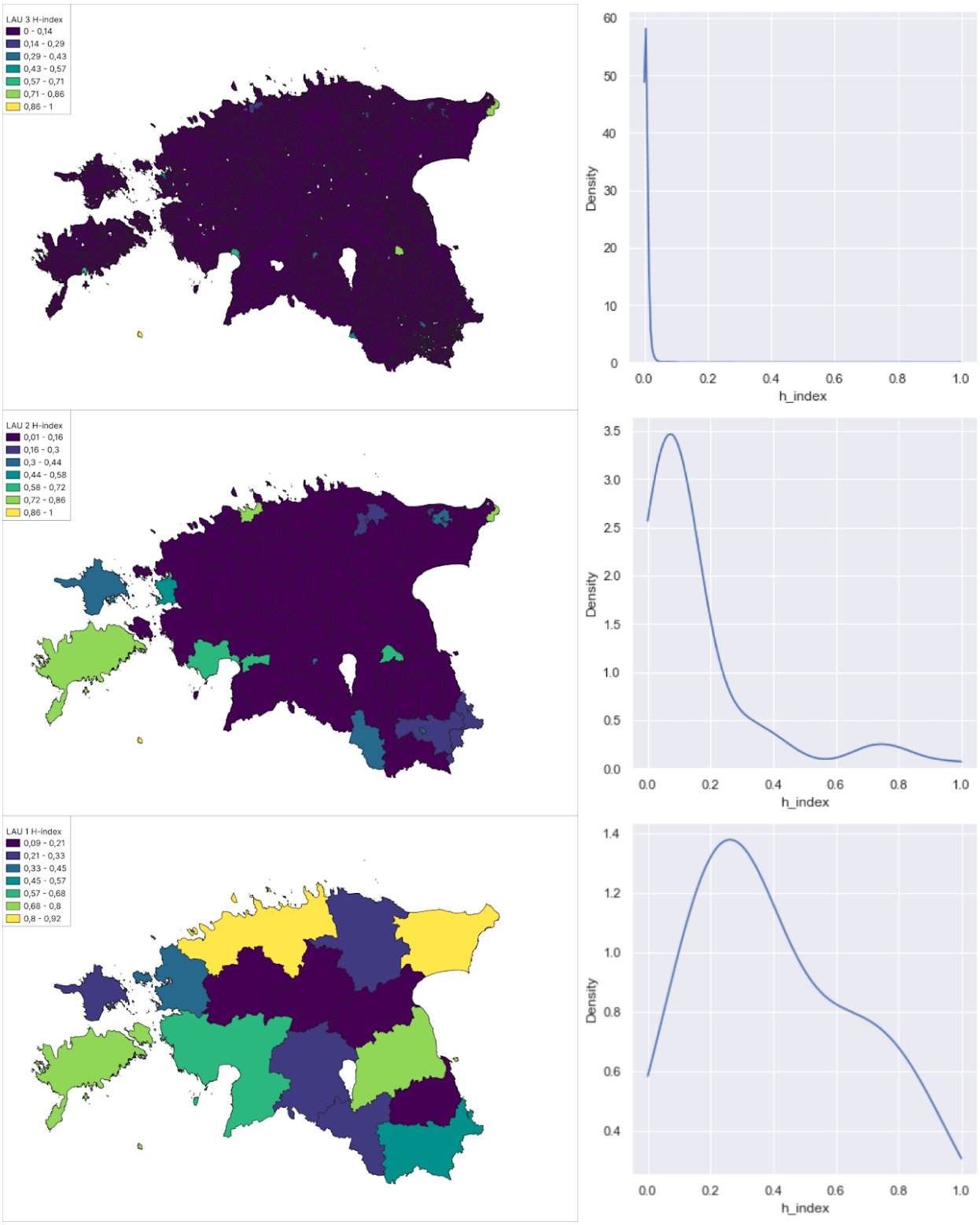

Figure 18. H-index and kernel density plot of h-index for LAU levels 1 to 3

3.2. RESULTS OF EXPERIMENTAL SETUP

3.2.1. H-INDEX OF GENERATED ZONES

The main outcome of the experimental setup is the similarities between the parameter combinations. Minimal h-index variable did not make a difference in the outcome of the algorithm. This may have happened due to the lower amounts of data outside of hot spots. Interestingly the results did not alternate between different minimal h-index parameters and produced exactly the same results across different minimal h-index parameter, however different distances from initial seed changed the outcome. The distance that a zone can grow from its initial position turned out to be the parameter that changed the outcome of the algorithm. If a zone could grow larger then the overall number of zones was lower and the average h-index was higher. This could be a bit misleading because a larger number of zones could bring the average and median values down. To confirm the relation between zone area and h-index a correlation analysis was done. The results indicated a weak to moderate positive correlation between the two outcome parameters (see Figure 19).

shows the differences between different seed types.

Table 4. H-index, zone count and correlation between h-index and zone area

Seed	H mean	H median	Zones	Index-area correlation
Distance 30 H-index 0.5-0.9	0.08	0.02	194	0.46
Distance 50 H-index 0.5-0.9	0.09	0.02	159	0.52
Distance 70 H-index 0.5-0.9	0.09	0.02	142	0.53

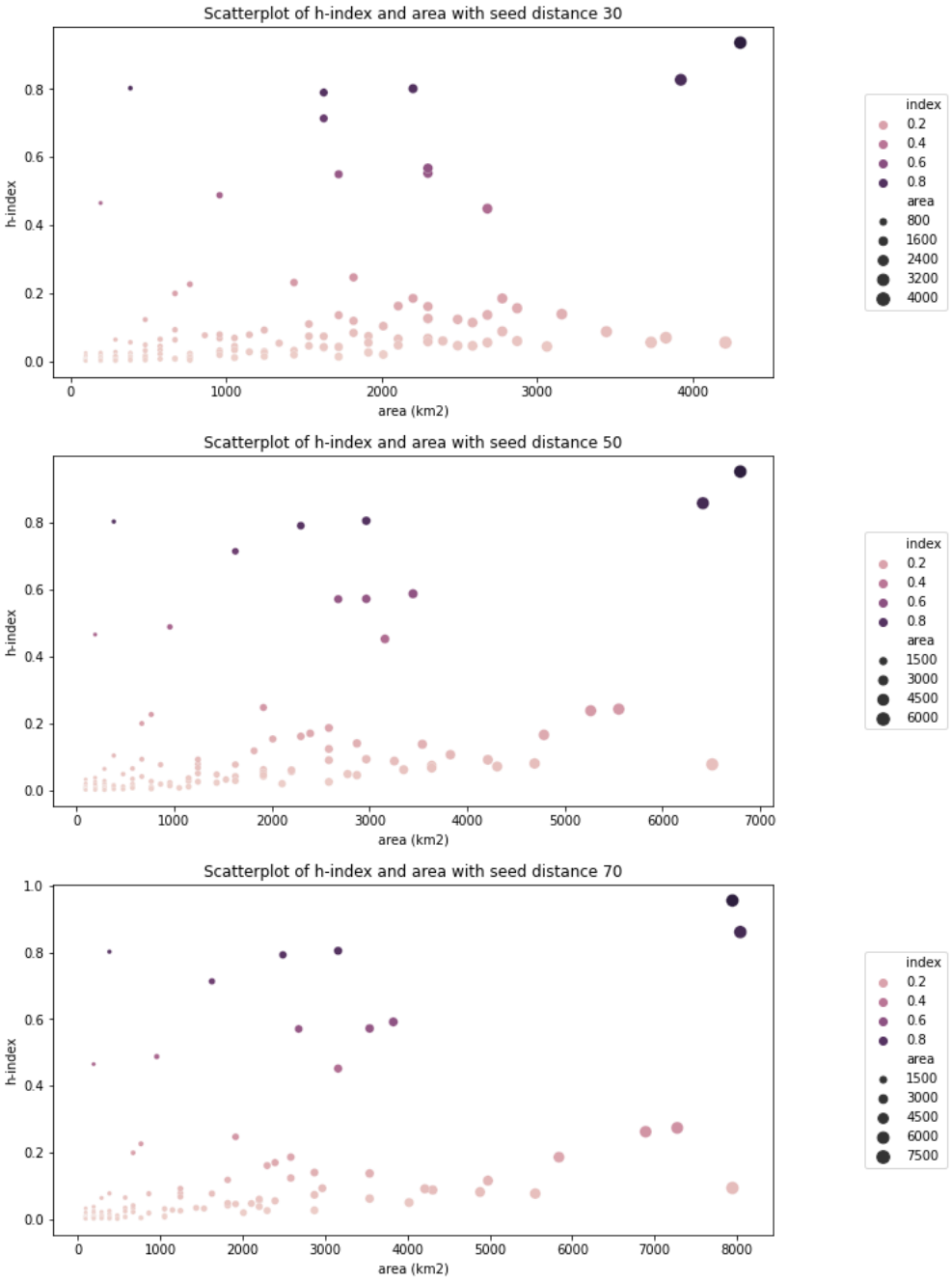

Figure 19. Scatterplots of the relations between h-index and zone sizes

The zones had mostly similar layout with the majority of them having a h-index under 0.2 The starting seeds that performed most similarly throughout different seeds were zones that started in Ida-Viru, Rapla, Valga and Viljandi county centers. These zones had the same composition with different seeds. Another zone with a relatively high h-index was centered around the hot spot in north-eastern Estonia where there was no initial seed (see Figure 20Figure 21Figure 22).

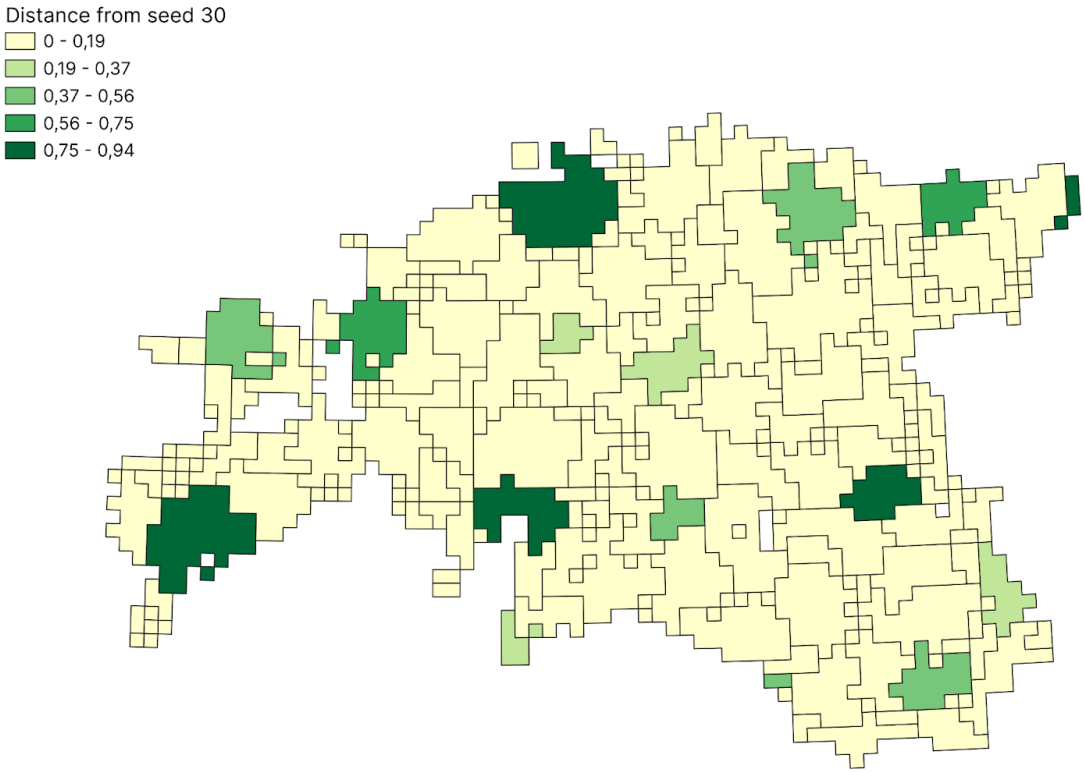

Figure 20. H-index of zones with distance from seed of 30 km

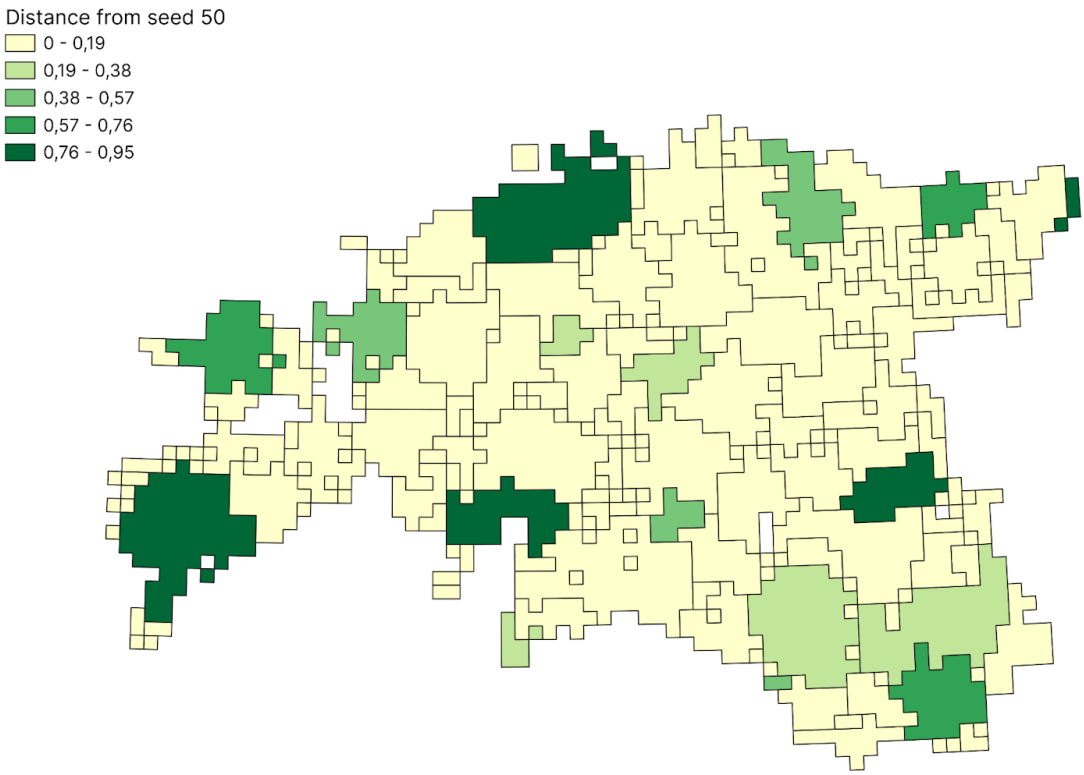

Figure 21. H-index of zones with distance from seed of 50 km

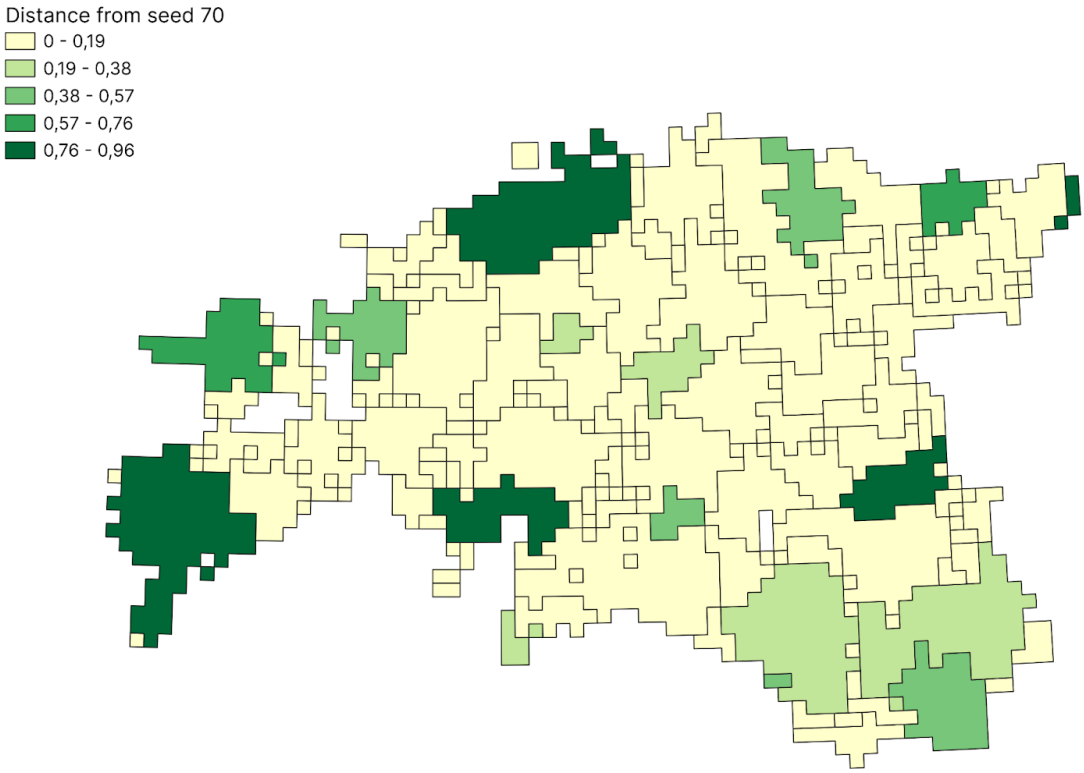

Figure 22. H-index of zones with distance from seed of 70 km

The best performing seeds from this experimental setup were Ida-Viru, Harju, Pärnu, Saare and Tartu county centers which produced zones with higher h-indexes on all occasions.

Another unique seed was the center of Valga county where the initial seed was a spatial outlier known as a doughnut where the surrounding values were larger. But the zone grew by only one grid on each occasion and had a relatively high h-index (0.42) compared to the average.

Despite the initial estimation of cold spots becoming borders for newly created zones there was one outlier, the center of Saare county, which almost took over the whole cold spot cluster. This probably happened because of a low amount of events in the cold spot cluster where the addition of new grids to the zone did not decrease the h-index substantially.

3.2.2. COMPACTNESS OF GENERATED ZONES

The compactness of the generated zones did not have much variety between different input parameters; however, the zones which had a limit of spreading 50km from their initial starting

point had highest compactness scores. The Polsby-Popper and Schwartzberg indicators favor inland zones since the shoreline complexity can add much more circumference to a zone which brings the overall score down. From the results we see that zones that can extend 50km from the initial zone were the most compact by a small margin. The algorithm produced multiple zones which consisted of a single grid (see Table 5).

Table 5. Compactness indexes for results

Polsby mean	Polsby median	Schwartzberg mean	Schwartzberg median	Convex hull mean	Convex hull median	Reock mean	Reock median
Distance 30 H-index 0.5-0.9	0.44	0.69	0.66	0.81	0.80	0.48	0.48
Distance 50 H-index 0.5-0.9	0.45	0.70	0.67	0.82	0.80	0.48	0.50
Distance 70 H-index 0.5-0.9	0.43	0.69	0.66	0.82	0.79	0.47	0.48

4. DISCUSSION AND CONCLUSION

With the increase of MPD usage in different projects and its availability for scientific research it is important to have a measure to determine the accuracy of interpolated MPD results. The policy makers and users of MPD-based results are likely interested in higher accuracy. This creates a need for a methodology to assess the accuracy of the results.

The main goal of my thesis was to implement a method that can give an estimation of accuracy for spatial interpolation of MPD data when interpolating to the current LAU and to use this approach to explore the possibility of algorithmically generating homogeneous zones. The approach was tested on data from one of the MNO-s in Estonia.

The research questions of this thesis were:

Q1: Can the homogeneity approach be applied to assess the precision of MPD statistics?

Q2: Can zones be created algorithmically based on different input parameters?

The homogeneity index is one approach to show the certainty that the interpolated subscribers, events, stays etc. were actually present at a specific location at a specific time. By calculating h-index values for the current local administrative units we can see that most of LAU levels 2 and 3 have generally really low scores. Which indicates that the actual location of the event probably differs from the interpolated location. While larger urban areas had higher h-index values the more rural areas had lower values. This can be because of a larger amount of 2G masts which are more common in rural areas and have larger coverage areas that can span multiple local administrative units. The opposite case in urban areas where the density of 3G and 4G masts is higher and most of the events will be interpolated closer nearby therefore increasing h-index. The combination of lower amounts of stays in rural areas and cell towers with larger coverage areas creates a need for a weighted approach which would tone down the impact of urban areas on creation of the zones resulting in more uniform zones across the dataset. Since the algorithm favors areas with higher amounts of data and does not have any weighted options as of now, areas with lower amounts of data will not increase the h-index as much. However, the patterns of zones seem to correlate with population density. This creates the question if event density is necessary for a high h-index. Homogeneity index is one approach to consider when assessing the accuracy of MPD-based results however the approach still needs tinkering to reach its full potential.

To answer the second research question a deterministic algorithm was developed and through the experimental setup different input parameters were given as inputs to see how the algorithm behave. Interestingly the outcome was not affected by the minimal h-index threshold parameter although the parameter affected the break condition of the zone generation part of the algorithm. The minimal h-index parameter was initially introduced to allow zones to lose h-index values in favor of adding more grid squares to the zone while not falling below a certain threshold. But due to the majority of zones having h-index well below the minimum the break condition did not assist in integrating new grid squares to zones in the case of low h-index areas and smaller zones. Instead, it had the reverse effect of finalizing zones when there were no suitable candidates, adding to the number of generated zones while not reaching the desired h-index within each zone.

The distance a zone could grow from its seed point caused also an increase in the average and median h-index of the output zones. This meant that zones could grow larger. There was a weak to moderate positive correlation between the area and the h-index of the zone. However, the increase of average and median h-index between 30, 50 and 70km was marginal. The number of zones decreased with the increase of distance parameter, which is logical because larger zones reduce the number of available grid squares for the generation of new zones.

The area indexes calculated for the zones might be a good way to sort out zones with less desirable shapes such as long-drawn-out segments. These indexes mostly favor circular shaped areas. Out of the different parameter combinations the highest average scores were with the distance from seed parameter as 50 km. Overall the highest average index value was for the Convex-Hull/Area method. The zones are made up of square elements which is the reason the Area/Convex hull score is substantially higher than other scores.

In the current version of the algorithm the desired h-index results for all the zones were not achieved. There were multiple zones with desired output but the majority of created zones had low h-index values and an abundance of single grid zones. At this point I cannot say what the ideal number of zones would be based on the input data. In the used dataset the better performing seeds throughout different input parameters were Ida-Viru, Harju, Pärnu, Saare and Tartu county centers which produced results with high h-indexes. It might be worth looking into selection of better seeds for future tests.

Overall, the thesis achieved its goals: the homogeneity approach was used to assess the accuracy of current local administrative units and served as a base for the second goal. The zoning algorithm created different outcomes based on input parameters. Even though the desired homogeneity index was not achieved on most of the created zones it serves as a good point to develop further methodology on. Generating new data-based homogenous zones gives new opportunities for policy makers and MPD-based result users alike.

5. FUTURE WORK

The algorithm can still be improved but the initial results show promise that generation of homogeneous zones based on data from interpolation of overlapping coverage areas is possible. When developing this methodology further it is important to keep it in mind to not overfit the algorithm on one dataset where it would perform worse with new data.

There are many parameters that could be added to improve the quality of the zones that the algorithm produces. One thing to add to the algorithm would be to have different weights based on the number of events interpolated to grid squares, because right now the model has a bias to grow large zones around hot spots if given the chance. Since the calculation of h-index takes into account the total number of events that have been interpolated into the zone, additional grid squares with lower amounts of events do not decrease the overall h-index marginally.

Secondly the algorithm creates a lot of zones that have low h-index. It might be necessary to create a wrapper function that would add smaller zones to larger ones to decrease the high number of zones created by the current version.

The selection of initial seeds could be derived from hot and spot clusters of input data. The model could also be fitted on different time frames of data. There might also be a better way of approaching the h-index on a contextual level.

KOKKUVÕTE

Minimaalse tsoneeringu loomine ülekattega levialadest

Patrick Joan Thomson

Mobiilpositsioneerimisandmete kasutus on laialdaselt levinud. Informatisooni ja kommunikatsioonitehnoloogia areng on turundusprotsessidele oluliselt mõju avaldandud. Lisaks sellele on suurandmete kättesaadavus ja olulisus on viimastel aastatel suurenenud (Roberts et al., 2014). Kuigi mitmed tööd on rõhku pannud rahvastiku jaotusele (Järv et al., 2018), turismile (Ahas et al., 2008; Raun et al., 2016) ja segregatsioonile (Silm & Ahas, 2014) pole paljud keskendunud andmete ruumilise interpooleerimise täpsusele.

Mobiilpositsioneerimisandmete interpoleerimiseks on kasutatud erinevaid meetodeid. Passiivse mobiilpositsioneerimise raskendavaks asjaoluks on see, et mobiilsideoperaatorite andmed on ainult telekommunikatsiooni masti täpsusega, kuigi inimesed võivad asuda tol hetkel ükskõik millises masti leviala punktis. Varasemalt on kasutatud alg ja lõppunkti metoodikat, kus inimese liikumise tühjad kohad kaeti marsruudimootorite abiga (Järv et al., 2018). Levinum on ka pindala kaalutud indeks. Aasa leidis, et laialdaselt kasutatud meetoditest punkt-polügonis ning pindalaga kaalutud indeksil on kehv tulemuslikkus, kuid eluasemega kaalutud indeks ja Kohanduv Mortoni võrgustik suurendavad mobiilpositsioneerimisandmetel põhineva populatsiooni hindamise kirjeldavat võimet (Aasa et al., 2021).

Sageli eeldavad mobiilpostitsioneerimisandmepõhiste tulemuste kasutajad, et näitajad esitatakse olemasolevatel haldusüksustel ja sooviksid teada, kui palju inimesi oli konkreetses asukohas konkreetsel ajal. Näiteks ei oleks kasulik öelda klientidele või otsusetegijatele, et 30% kõigist ühe mobiilsideoperaatori võrgu kasutajatest on teinud teatud telekommunikatsioonimasti levialas kõne. Kui mobiilpostitsioneerimisandmeid kasutatakse laialdaselt statistika loomiseks ja statistikat antakse tavaliselt kohalike haldusüksuste või muude kohandatud jaotuste põhjal välja, siis on vaja kontrollida tulemuste täpsust. Sellise andmeallika kohta on leida referentsandmeid, et tulemusi kontrollida. See loob vajaduse metoodilise lähenemise järele, kuidas hinnata mobiilpostitsioneerimisandmetel põhineval esitatud statistika täpsust.

Käesoleva magistritöö eesmärgiks on rakendada homogeensusindeksil põhinevat lähenemist et kontrollida tulemuste usaldusväärsust. Lisaks sellele uuritakse võimalust, kas algoritmilise

lähenemise kadu saab tekitada homogeenseid tsoone, mille sees on andmete ruumi interpoleerimise tulemused võimalikult usaldusväärsed. Töös kasutati Positiumi metoodikal põhinevat interpoleerimisemeetodit, mille aluseks on telekommunikatsioonimastide levialasse interpoleerimine erinevate konfiguratsiooniparameetritega.

Töö tulemusena selgus et kõrgema rahvastikutihedusega (linnalistel) aladel on keskmisest suurem usaldusväärsus homogeensus indeksi järgi. Maapiirkondadesse toimingute asukohta positsioneerides kasutades passiivseid meetodeid on suur tõenäosus, et inimene, kes selle toimingu tegi ei pruukinud asuda selles kindlas asukohas. Linnalise ala täpsuse ja maapiirkonna ebatäpset positsioneerimist põhjendab erinevate generatsioonide mastide ebaühtlane paiknemine. Linnalistel aladel on rohkem 4G ja 3G maste, mille leviala on väiksem, seetõttu on ka väiksem võimalus eksida. Maapiirkonnas seevastu on suurem osakaal ka 2G mastidel, mille leviala on palju suurem. Homogeensusindeks on lähenemisviis, mida võib kaaluda passiivse mobiilpositsioneerimise tulemuste hindamisel, metoodika täieliku potentsiaali saavutamiseks on vaja seda veel kohendada.

Töö käigus loodud algoritmile anti sisenditena erinevad sisendparameetrid, et näha, kuidas algoritm käitub ning millised tulemused on. Huvitaval kombel minimaalse h-indeksi lävendi parameeter ei mõjutanud tulemust, kuigi parameeter mõjutas algoritmi tsooni genereerimise osa katkestustingimust. Minimaalse h-indeksi parameeter võeti algselt kasutusele selleks, et tsoonid võiksid kaotada h-indeksi väärtust, et tsooni ennast suurendada uute võrgu osade arvelt, samal ajal jäädes kõrgemale minimaalsest lävendist. Kuid enamiku tsoonide puhul, mille h-indeks jäi alla miinimumi, ei aidanud murdetingimus madalate h-indeksiga alade ja väiksemate tsoonide korral uute võrguruutude integreerimist tsoonidesse. Selle efekt oli vastupidine, kui tsoonid arvati lõplikeks, sest sobivaid kandidaate polnud, suurendades genereeritud tsoonide arvu. Seetõttu ei saavutatud soovitud h-indeksit igas tsoonis.

Üldiselt saavutas magistritöö oma eesmärgid: homogeensusindeksi kasutati praeguste haldusüksuste tasemel täpsuse hindamiseks ja see oli aluseks teisele eesmärgile. Tsoneerimisalgoritm lõi sisendparameetrite põhjal erinevaid tulemusi. Ehkki enamiku loodud tsoonide puhul ei saavutatud soovitud homogeensuse indeksit, on see hea algus edasise metoodika väljatöötamiseks. Uute andmepõhiste homogeensete tsoonide loomine annab uued võimalused nii otsusetegijatele kui ka mobiilpositsioneerimispõhiste tulemuste kasutajatele.

AKNOWLEDGEMENTS

My biggest thanks goes to my supervisors Alexander Kmoch and Margus Tiru who were supportive throughout writing my thesis. Margus suggested the topic and helped me along the way. Alexander brought up different viewpoints of the work which I had not thought of and helped make it more scientific.

I also would like to thank the company Positium who provided me access for the data I was using as the input for this thesis. A special thanks goes to my team leader in Positium Marko Peterson who was understanding and supportive throughout writing the thesis and helped me whenever he could. A special thanks goes to my co-worker Eva-Johanna who helped me with formatting because I have no knowledge on how Microsoft Word works.

BIBLIOGRAPHY

Aasa, A., Kamenjuk, P., Saluveer, E., Šimbera, J., & Raun, J. (2021). Spatial interpolation of mobile positioning data for population statistics. *Journal of Location Based Services*, *0*(0), 1–22. https://doi.org/10.1080/17489725.2021.1917710

Afrianto, W. F., Hikmat, A., & Widyatmoko, D. (2020). Plant Species Diversity and Degree of Homogeneity after the 2010 Eruption of Mount Merapi, Indonesia. *Biosaintifika: Journal of Biology & Biology Education*, *12*(2), 274–281. https://doi.org/10.15294/biosaintifika.v12i2.23525

Ahas, R., Aasa, A., Roose, A., Mark, Ü., & Silm, S. (2008). Evaluating passive mobile positioning data for tourism surveys: An Estonian case study. *Tourism Management*, *29*(3), 469–486. https://doi.org/10.1016/j.tourman.2007.05.014

Ahas, R., Laineste, J., Aasa, A., & Mark, Ü. (2007). The Spatial Accuracy of Mobile Positioning: Some experiences with Geographical Studies in Estonia. In *Location Based Services & TeleCartography* (pp. 445–460). http://ezproxy.utlib.ut.ee/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=edo&AN=32999944&site=eds-live

Ahas, R., Silm, S., Järv, O., Saluveer, E., & Tiru, M. (2010). Using Mobile Positioning Data to Model Locations Meaningful to Users of Mobile Phones. *Journal of Urban Technology*, *17*(1), 3–27. https://doi.org/10.1080/10630731003597306

Anselin, L. (1995). Local Indicators of Spatial Association—LISA. *Geographical Analysis*, *27*(2), 93–115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x

Documentation | GeoDa on Github. (n.d.). Retrieved May 16, 2021, from https://geodacenter.github.io/documentation.html

Jarv, O., Tenkanen, H., Salonen, M., Ahas, R., & Toivonen, T. (2018). Dynamic cities: Location-based accessibility modelling as a function of time. *Applied Geography*, *95*, 101–110. https://doi.org/10.1016/j.apgeog.2018.04.009

Lechthaler, B., Pauly, C., & Mücklich, F. (2020). Objective homogeneity quantification of a periodic surface using the Gini coefficient. *Scientific Reports*, *10*(1), 14516. https://doi.org/10.1038/s41598-020-70758-9

Moreno-Mateos, D., Mander, Ü., Comín, F. A., Pedrocchi, C., & Uuemaa, E. (2008). Relationships between Landscape Pattern, Wetland Characteristics, and Water Quality in Agricultural Catchments. *Journal of Environmental Quality*, *37*(6), 2170–2180. https://doi.org/10.2134/jeq2007.0591

Morton, G. M. (1966). A computer oriented geodetic data base and a new technique in file sequencing. 1-20. Canada: IBM

Ogulenko, A., Benenson, I., Omer, I., & Alon, B. (2021). Probabilistic positioning in mobile phone network and its consequences for the privacy of mobility data. *Computers, Environment and Urban Systems*, *85*, 101550. https://doi.org/10.1016/j.compenvurbsys.2020.101550

Positium Data Mediator. (n.d.). Positium. Retrieved May 18, 2021, from https://positium.com/research/positium-data-mediator

Pukk, S. (n.d.). *Unmasking oscillation from mobile positioning data*. 56.

Raun, J., Ahas, R., & Tiru, M. (2016). Measuring tourism destinations using mobile tracking data. *Tourism Management*, *57*, 202–212. https://doi.org/10.1016/j.tourman.2016.06.006

Rey, S. J., & Anselin, L. (2010). PySAL: A Python Library of Spatial Analytical Methods. In M. M. Fischer & A. Getis (Eds.), *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications* (pp. 175–193). Springer. https://doi.org/10.1007/978-3-642-03647-7_11

Roberts, J. H., Kayande, U., & Stremersch, S. (2014). From academic research to marketing practice: Exploring the marketing science value chain. *International Journal of Research in Marketing*, *31*(2), 127–140. https://doi.org/10.1016/j.ijresmar.2013.07.006

Sauter, M. (2011). From GSM to LTE: An Introduction to Mobile Networks and Mobile Broadband. In *From GSM to LTE: An Introduction to Mobile Networks and Mobile Broadband*. John Wiley and Sons. https://doi.org/10.1002/9780470978238

Silm, S., & Ahas, R. (2014). The temporal variation of ethnic segregation in a city: Evidence from a mobile phone use dataset. *Social Science Research*, *47*, 30–43. https://doi.org/10.1016/j.ssresearch.2014.03.011

Tiru, M. (2021, May). *Positium Data Mediator methodology* [Interview].

VKR. (n.d.). Retrieved May 23, 2021, from https://estat.stat.ee/StatistikaKaart/VKR

Wang, F., & Chen, C. (2018). On data processing required to derive mobility patterns from passively-generated mobile phone data. *Transportation Research Part C: Emerging Technologies*, *87*, 58–74. https://doi.org/10.1016/j.trc.2017.12.003

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Patrick Joan Thomson, annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose, Creating minimal zoning from overlapping coverage areas, mille juhendajad on Alexander Kmoch ja Margus Tiru, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 3.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Patrick Joan Thomson

23.05.2021