Universiteit Leiden
The Netherlands

Computer Science

Spatial Subgroup Discovery

Nils Burghouwt

Supervisors:
dr. Matthijs van Leeuwen & M.Sc. Ioanna Papagianni

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

26/07/2022

Abstract

With the rise of big data it becomes more and more interesting to recognize patterns of interest in data. For finding patterns in geographically coherent datasets, the spatial information of an instance is crucial. This thesis introduces the problem of finding the optimal list of subgroups that together explain the most relevant deviations in the data with respect to a given target variable and that adhere to the imposed spatial constraint. An algorithm is proposed that combines the concepts of subgroup discovery and spatial data mining. The spatial subgroup discovery algorithm finds a list of subgroups of which the instances in a subgroup are not more than a defined distance away from each other. It makes use of three hyperparameters: the beam width, the maximum search depth, and the radius. The algorithm has been tested on one dataset varying the three hyperparameters and evaluated on the quality measure: the sum of Weighted Kullback-Leibler divergences. The algorithm was able to find spatial subgroups, in which most of the subgroup lists had a quality between 0.0025 and 0.0035. The results show that the larger the beam width and the radius, the higher the quality and the higher the runtime. Varying the maximum search depth did not cause an effect on the quality of the subgroup list, only the runtime did increase whilst increasing this hyperparameter. In comparison with clustering, the results of this thesis show that due to the use of subgroup discovery in combination with spatial constraints, the algorithm is not only able to find patterns, but is also able to explain the patterns because of the subgroup description.

Contents

1 Introduction

Due to the upswing of available data, recognizing patterns in data becomes more and more interesting. Data mining makes that possible. It aims at generating new information by examining a large pre-existing database [15]. Common data mining techniques are classification, association, regression, and clustering.

A distinction between data mining techniques can be made by looking at the occurrence of a target. In unsupervised data mining there is no target variable. It tries to turn the data into relevant information. Clustering is an example of an unsupervised data mining technique. It tries to partition the dataset into clusters. The objects in the same cluster should share similarities, but should not be homogeneous with objects of other clusters. Wang et al. [24] used clustering to group transactions. A transaction consists of a set of items. A basket of items purchased during a shopping trip is an example of a transaction. A cluster contains similar transactions, dissimilar transactions are in different clusters.

In contrast to unsupervised analysis, supervised data mining uses a target variable. The goal is to predict or forecast the target value of new data. Subgroup discovery is a supervised data mining task. It tries to extract interesting rules with respect to a target variable [6]. An example could be in the commercial domain. In this case the target variable is whether the customer buys the product or not. Subgroup discovery tries to find rules that show a significant deviation in the distribution of the target variable [7]. A subgroup could be the people that have an age of more than 50 and live in the Netherlands. The target variable whether they bought an electric bike is here mostly 'yes'.

The features of the dataset define the found patterns. Information about the location of the instances could also be considered as a feature. For example, companies can adjust their marketing strategy based on the found patterns of the customers behaviour and actions on their website. Apart from that, also the country and region of the customers could be taken into consideration when trying to find interesting patterns. Special marketing campaigns could therefore be performed to the specific regions that show interesting patterns. Besides the commercial sector, we are also able to gain knowledge by finding patterns in the ecological world. For example, by taking the location of the recorded animals into account, targeted actions could be performed to the regions where animals show a different behaviour. This thesis tries to find regions in which the target variable, e.g., animal species, shows an interesting pattern. These patterns could be seen as groups with similar characteristics, thus subgroups.

Van Leeuwen et al. proposed a clustering method based on the Minimum Description Length (MDL) principle [22] to group the European mammals, see Figure 1. The MDL principle ensures that the total compressed size of the components is minimized. The figure shows the best found decomposition where 6 groups (clusters) are distinguished. The used dataset consists of 2670 instances, which represent each grid cell (dot in the figure) in Europe. The presence/absence of 194 mammals is specified per cell. Within a cluster you can find the cells that show the same patterns; cells that have similar presence/absence records should belong to the same cluster.

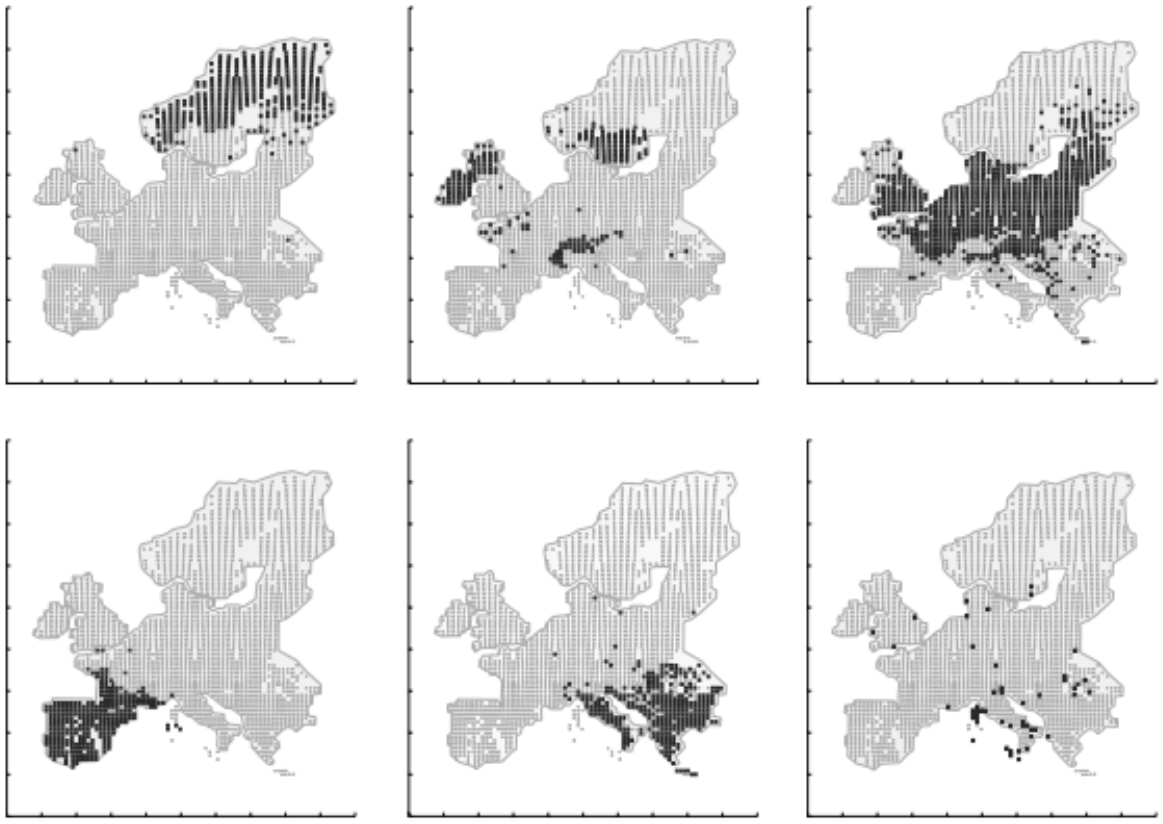

Figure 1: Clustering the *mammals* dataset. Source: 'Identifying the components' (M.van Leeuwen, J. Vreeken, A. Siebes, 2009) [22]

However, this decomposition was made without exploiting any prior knowledge or geographical constraints of the cells. It can therefore be observed that not all cells within a decomposition are near each other. For example, if we zoom in to one of them (Figure 2), we can identify three different regions within one group. Besides that, there are also some cells visible that are not even in one of these regions.

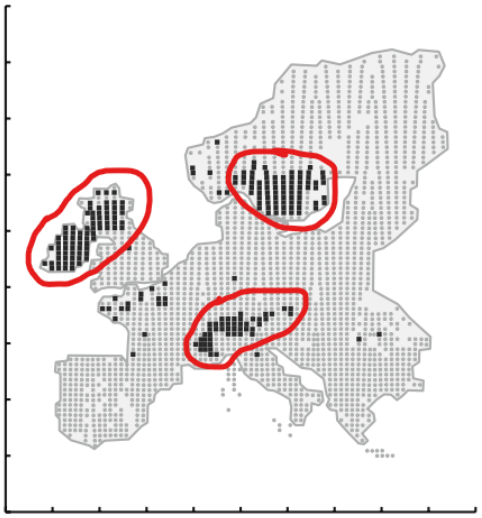

Figure 2: One cluster of the decomposition of the *mammals* dataset. Source: 'Identifying the components' (M.van Leeuwen, J. Vreeken, A. Siebes, 2009) [22]

Heikinheimo et al. clustered the same *mammals* dataset [5]. The k-means clustering method was used here instead. The 6 found clusters are visible in Figure 3. Again can be observed that not all the cells within a cluster are near each other. For example, the cluster that covers most of Italy and Greece (yellow), has also some cells in Germany, France and the UK.

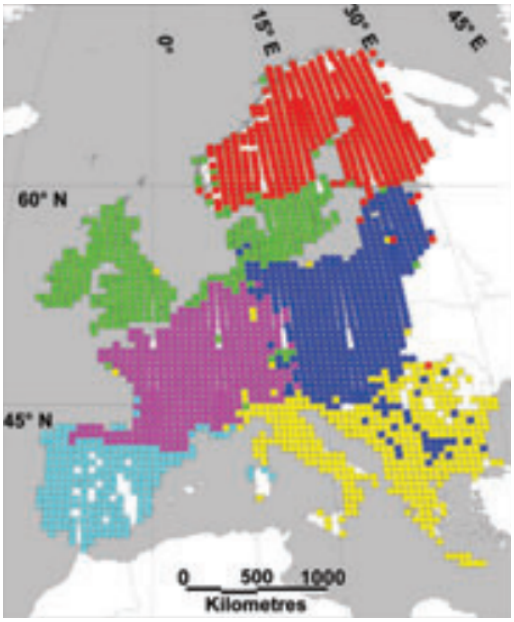

Figure 3: The 6 found clusters of the *mammals* dataset using k-means. Each color represents a cluster. Source: 'Biogeography of European land mammals shows environmentally distinct and spatially coherent clusters' (H. Heikinheimo, M. Fortelius, J. Eronen and H. Mannila, 2007) [5]

In order to test the relation between the environment in a specific zone and the recorded mammal species there, the decomposition in Figure 3 was compared by looking whether the generated clusters differ significantly in the values of environmental variables [5]. The more similar the environmental variables in one cluster, the more the coherence. Instead of comparing the results with the environmental variables, this thesis will use the environmental variables as features and the mammal species as target. That allows to decide which environmental features are related to the different mammal species. This thesis differentiates therefore from the aforementioned paper and article because we do not apply a clustering technique but rather a spatial subgroup discovery method. Subgroup discovery allows the use of a target variable that is needed to find groups that stand out with respect to a target variable. Therefore, subgroup discovery will be used instead of clustering. Besides that, the goal is not to make a complete decomposition as is done in Figure 1. In order to add the constraint, this thesis combines subgroup discovery with another data mining concept: spatial data mining. Spatial data mining explores possibly unknown and useful patterns from spatial datasets. Spatial data mining has already been applied to the well-known data mining techniques classification [16], clustering [9] and association rule mining [2].

We try to find interesting groups in a dataset with the constraint that the instances within a group should not be more than an acceptable radius away from each other. An 'acceptable' radius could for example be the length of an average country in Europe. An example of this thesis algorithm performed on the same *mammals* dataset in combination with climate data can be seen in Figure 4. The red surface represents a found subgroup. It can be observed that all the instances within one group are near each other.

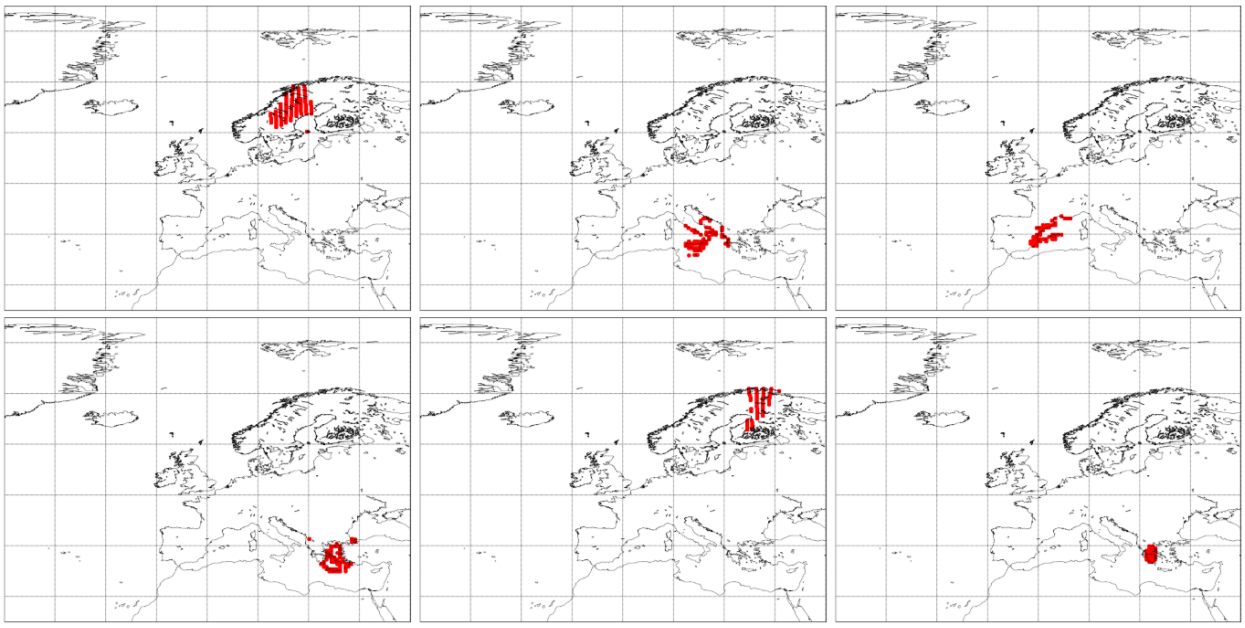

Figure 4: Example of spatial subgroup discovery algorithm performed on the *mammals* dataset in combination with climate data. In total 6 different subgroups are identified. The mammal species is the target and the features are the locations of the mammals and the climate variables. The used dataset will be further explained in Section 5.1

To the best of our knowledge we are the first to investigate a subgroup discovery task given spatial constraints. With finding subgroups that follow (geo-)spatial thresholds, we aim to learn possibly better subgroups with more interpretable explanations.

The aim of this project is to adapt standard subgroup discovery methods such that spatial constraints can be imposed in order to find subgroups that share similar locations. After modifying the subgroup discovery problem statement, an algorithm will be developed and empirically tested on a dataset containing spatial information.

The problem can be informally stated as follows:

Find the optimal set of subgroups that 1) together explain the most relevant deviations in the data with respect to a given target variable and 2) that adhere to the imposed spatial constraint.

This problem statement is formalised in Section 2.

1.1 Thesis overview

This bachelor thesis is written in cooperation with LIACS and is supervised by dr. Matthijs van Leeuwen and M.Sc. Ioanna Papagianni.

In order to clarify this thesis project first standard subgroup discovery and spatial constraints are thoroughly explained in Chapter 2. Then, Chapter 3 reviews related work. Furthermore, the spatial subgroup discovery algorithm will be explained in Chapter 4 by mentioning the modified separate and conquer algorithm and beam search using spatial information. Chapter 5 describes the experiments and their outcome and finally, Chapter 6 concludes the thesis with directions for future work.

2 Preliminaries

This chapter describes the information that is needed to understand the methodology. Firstly, the relevant definitions of subgroup discovery needed for the formal problem statement will be mentioned (Section 2.1). Subsequently, the spatial terms and intuition will be explained (Section 2.2). A summary of the notation of the previous sections can then be found in Table 5. With that information the formal problem statement can be defined in Section 2.3.

2.1 Subgroup discovery

Subgroup discovery aims to discover meaningful descriptions of subsets of a dataset among different variables with respect to a target of interest [17]. A description is a conjunction of conditions on $\mathbf{X}$, each specifying a specific value or interval on a variable [18]. For example, given two features $X_{max(temp(Nov))}$ and $X_{\mu(temp(Jun))}$, a candidate condition of a subgroup, denoted by s, is $X_{max(temp(Nov))} > 17.16$. A description a is then for example 'maximum temperature in November more than 17.16 and mean temperature in June less than 21.22'. It has $|a| = 2$ conditions. A formalisation of this description a is:

$$X_{max(temp(Nov))} > 17.16 \wedge X_{\mu(temp(Jun))} < 21.22$$

Subgroup discovery can also be described as the data mining method that aims to find all subgroups within the inductive constraints that reveal a significant deviation in the distribution of the target attribute. The common constraints are the minimum coverage, which states that each subgroup should have at least a certain number of objects, and the minimum quality, which states that each subgroup should have at least a certain quality. Due to the fact that this thesis is about spatial subgroup discovery, we introduce also a spatial constraint. This will be further explained in Section 2.2. An ordered set of subgroups that describe different parts of the data, is called a subgroup list (M) [18]. An example of a subgroup list is given in Table 1.

			Pr($mammal\_species = ...\mid s$) in %				
s	description	n_s	acomys minous	alces alces	alopex lagopus	rattus norvegicus	mus domesticus
1	$max(temp(Nov)) > 17.16$	30	40	10	0	10	40
2	$bioclim4 < 682.74$	15	20	40	0	0	40
3	$min(temp(Jun)) > 8.16$	50	0	100	0	0	0
4	$prec(Nov) > 67.83$	5	40	30	0	20	10
dataset distribution		0	30	40	10	15	5

Table 1: This table contains a subgroup list derived from a toy example dataset of the *mammals* dataset. The dataset has 10 numeric features and one nominal target consisting of 5 different mammal species. Each of the 4 subgroups s are covered by n_s instances and are defined by 'description'. The rest of the numbers are denoted by $\Pr(mammal\_species = ...|s)$, which indicates the estimated probability (in %) of each class label (mammal species) occurring within the subgroup.

For each subgroup the probability distribution can be defined. If for example, we take the first subgroup in Table 1 with the description $max\_temp\_nov\_utm > 17.16$ given, the statistics can be defined as

$$\hat{\Theta}^{a_1} = \{\hat{p}_1 = 40; \hat{p}_2 = 0.10; \hat{p}_3 = 0; \hat{p}_4 = 0.10; \hat{p}_5 = 0.40\}.$$

$\hat{\Theta}^{a_i}$ is the vector of all parameter values of the rule i. The a_i represents the description of subgroup i. Here i has a range of 1 to 4. The classes 1 to 5 represent the mammals species. The $\hat{p}_c$ is the marginal probability for that class [18]. The target variable follows the following categorical distribution:

$$mammals\_species \sim Cat(\hat{p}_1, \hat{p}_5 = 40; \hat{p}_2, \hat{p}_4 = 10; \hat{p}_3 = 0)$$

In order to decide the quality of a subgroup the distribution of the subgroups target attribute is compared with the distribution of the datasets target attribute. The distribution of the target attribute of the dataset is denoted by:

$$\hat{\Theta}^d = \{\hat{p}_1 = 30; \hat{p}_2 = 0.40; \hat{p}_3 = 10; \hat{p}_4 = 0.15; \hat{p}_5 = 0.5\}.$$

The bigger the deviation and the larger the subgroup, the better the subgroup is. In order to calculate how interesting a subgroup is, a quality measure is used. In general, a quality measure for subgroup discovery consists of two components. The coverage (n_a) of the subgroup should be represented in the formula, and a function of the difference between the subgroups and datasets target distribution ($f(\hat{\Theta}^a, \hat{\Theta}^d)$) [18].

As the dataset in this thesis has a multinominal target, the Weighted Kullback-Leibler divergence will be used. This quality measure supports a multinominal target. It is defined as the Kullback-Leibler divergence between a subgroup's and dataset's target distribution linearly weighted by its coverage [18]. The formula is given by:

$$WKL(\hat{\Theta}^a; \hat{\Theta}^d) = n_a KL(\hat{\Theta}^a; \hat{\Theta}^d) \text{ [23]}$$

The formula for the Kullback-Leibler for categorical distributions is given by:

$$KL_{Cat}(\hat{\Theta}^a; \hat{\Theta}^d) = \sum_{c \in \mathcal{Y}} \hat{p}_{c|a} \log(\tfrac{\hat{p}_{c|a}}{\hat{p}_c}) \text{ [23]}$$

The $\hat{p}_{c|a}$ is the maximum likelihood estimate of the conditional probability of the target c (element of $\mathcal{Y}$) given the subgroup a. The intuition behind this formula is that it compares the probability distribution of the subgroup with the probability distribution of the dataset. The target variable is given and therefore the different labels of the target variable can be counted. The count of each category of the subgroup can be compared with the count of each category of the whole dataset by looking at their distributions ($\hat{\Theta}^a$ and $\hat{\Theta}^d$). If the distributions differ significantly from each other, it means that their distributions of the target variable also differ. Due to the fact that subgroup discovery aims at finding subgroups that stand out with respect to a given target variable, comparing the probability distributions of the target variable of the subgroup and of the whole dataset is therefore a good option. That is what the Weighted Kullback-Leibler divergence does. The bigger the difference of the distributions, the bigger the WKL will be. This concludes to a better subgroup in comparison to a subgroup with a lower WKL.

We can for example compare s_1 and s_2 from Table 1 on their quality using the WKL. Without even calculating the WKL of both subgroups, we can know by looking at numbers that the WKL of s_1 will be bigger. Not only is the coverage of s_1 bigger ($n_1 > n_4$), but also the $\hat{\Theta}^{a_1}$ differs

more from $\hat{\Theta}^d$ than $\hat{\Theta}^{a_4}$ does. Therefore, the $KL_{Cat}(\hat{\Theta}^{a_1}; \hat{\Theta}^d)$ will be bigger and because the n_1 is also bigger, the WKL of s_1 will subsequently also be bigger.

In order to measure the quality of different subgroup lists, the *Sum of Weighted Kullback-Leibler divergences* (SWKL) is introduced. This can be defined as the sum of weighted KL divergences for the individual subgroups [18]:

$$\text{SWKL}(M) = \frac{\sum_{i=1}^{w} n_i KL_{Cat}(\hat{\Theta}_j^a; \hat{\Theta}_j^d)}{|D|}$$

In order to normalise the sum of the individual weigted Kullback-Leibler divergence of the subgroups, it is divided by $|D|$, the number of instances in dataset D. The bigger the SWKL, the better the quality of the subgroup list. Figure 4 represents the plots of the subgroup list performed by the in this thesis proposed algorithm. The WKL of each of the green surfaces was calculated. Subsequently, the SWKL of the subgroup can be calculated.

2.2 Spatial constraints

In this thesis we work with data that contains spatial objects. This is an attribute that captures location in 2D or 3D space. This could be in the form of points or as geometrical spatial objects. In order to only allow subgroups that are geographically connected with each other, a spatial constraint should be imposed. It is important that only areas that are closely and geographically linked are being considered.

The data that is used for the spatial subgroup discovery algorithm has three spatial features: the latitude (φ), the longitude (λ) and the identification (ID) of the cell, utm_ID, which will be explained in Section 5.1. For now it is sufficient to know that the dataset makes use of a grid that consists of cells. Each cell is defined by an unique ID. An example of such a dataset is given in Table 2 and the corresponding grid of the data is given in Figure 5.

index	$mammal\_species$	λ	φ	utm_ID	$features$	
0	rattus norvegicus	1.8	5.1	2F	...	...
1	alopex lagopus	2.1	3.4	3D	...	...
2	rattus norvegicus	2.3	7.1	3H	...	...
3	mus domesticus	3.1	6.2	4G	...	...
4	alopex lagopus	3.4	3.7	4D	...	...
5	mus domesticus	4.1	5.7	5F	...	...
6	alopex lagopus	4.2	4.8	5E	...	...
7	alopex lagopus	5.1	6.4	6G	...	...
8	mus domesticus	5.4	3.6	6D	...	...
9	rattus norvegicus	6.2	6.0	7G	...	...
10	mus domesticus	6.6	0.9	7A	...	...

Table 2: Toy example dataset of the *mammals* dataset. Each row represents a recorded mammal. The $features$ columns are all the climate features, such as $max(tem(Nov))$, that are merged with the *mammals* dataset.

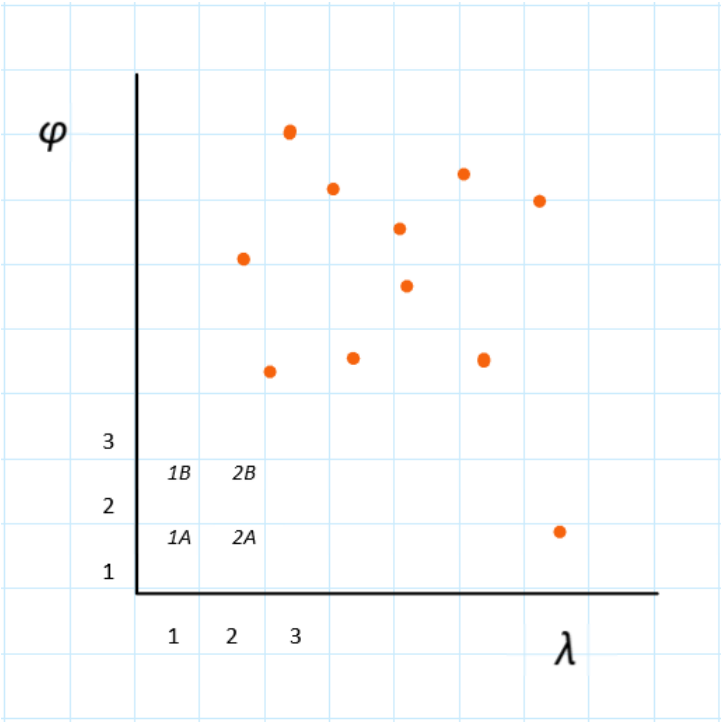

Figure 5: Toy example grid of the *mammals* dataset. The values within the cells (eg. '2B') represent the cell ID. Each orange dot corresponds to a recorded mammal from the toy example dataset 2.

In order to introduce the spatial constraint a circle with the predefined radius (r) is made around a sampled instance. The φ and the λ form the centroid of the constraint. A centroid will be formally written down as a coordinate $(x^c_\varphi, x^c_\lambda)$, in which x^c_φ gives the latitude value of the instance c and the x^c_λ the longitude value of the instance c. If we have sampled for example the instance with the index of 3 from Table 2, the centroid $(3.1, 5.2)$ will be formed. If the predefined radius is 1.5, the circle visible in Figure 6 around the centroid will be made.

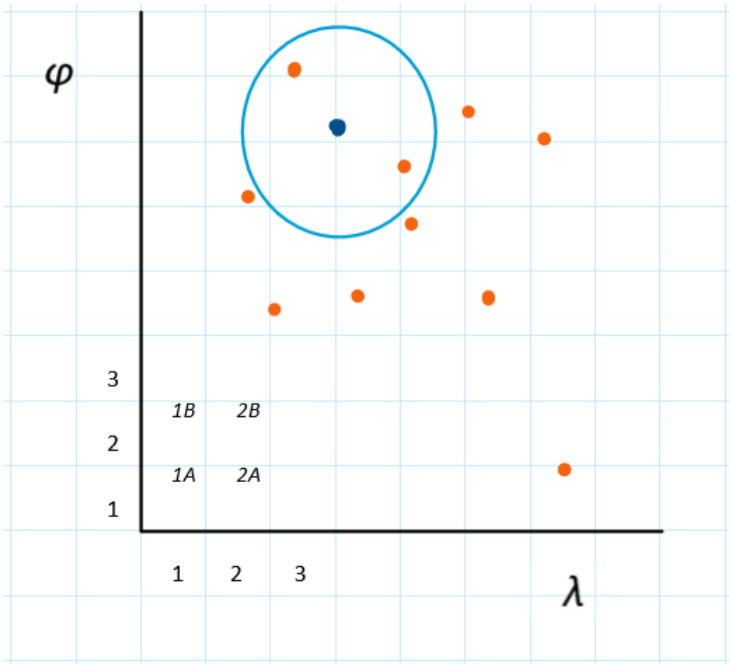

Figure 6: Toy example grid of the *mammals* dataset. The blue dot represents the sampled instance and forms the centroid of the constraint.

The centroid defines the constraint. Now we are going to define which instances belong to the formed constraint. All the instances within the circle could for example be in the formed subgroup, thus all the instances j that adhere to:

$$(x_\varphi^j - x_\varphi^c)^2 + (x_\lambda^j - x_\lambda^c)^2 \leq r^2$$

The x_φ^j and x_λ^j refer here respectively to the latitude and longitude value of the instance j. Excluding the instance that forms the centroid, the instances with indices 2 and 5 will then be part of the subgroup:

index	*mammal_species*	λ	φ	*utm_ID*	*features*	
2	rattus norvegicus	2.3	7.1	3H	...	...
3	mus domesticus	3.1	6.2	4G	...	...
5	mus domesticus	4.1	5.7	5F	...	...

Table 3: Subgroup formed by the constraint $(x_\varphi^j - x_\varphi^c)^2 + (x_\lambda^j - x_\lambda^c)^2 \leq r^2$.

This will however not be the constraint. In order to find interesting subgroups, each instance has not only the three spatial objects as features, but also the environmental information about the location of the instance are features. Making use of the cells in a grid makes this possible. Lakshmanan et al. [11] used the cell ID as spatial object and researched that 'the most straightforward approach to automated analysis would be to look within a small neighborhood (say 10 km) of an event of interest (say occurrence of a lightning flash) and see if certain values of a spatial gridded field

(say the maximum radar reflectivity observed at 6 km or higher) are associated with the event'. Features about a specific cell, such as the maximum temperature in March, can easily be linked with the cell in which the instance has been recorded. Therefore, we will not only use the latitude and longitude value, but also the cell ID (utm_ID) as part of the constraint. The instances that are allowed in the subgroup are all the instances that are in the covered or partly covered cells by the circle. If we apply this to Figure 6, we get Figure 7.

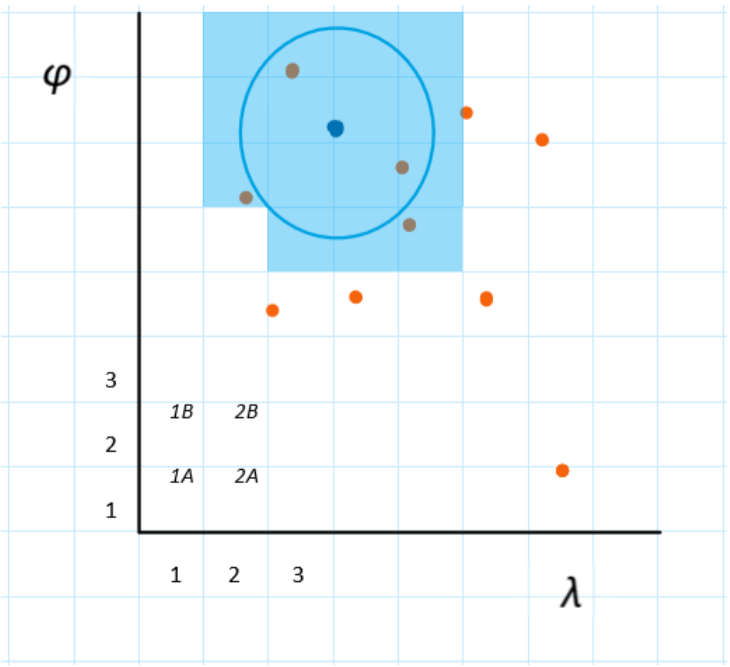

Figure 7: Toy example grid of the *mammals* dataset. The blue dot represents the sampled instance. Around the longitude and latitude coordinate (λ, φ) the circle with the predefined radius r is made. The blue surface covers the allowed cells of the formed subgroup. All the instances within the blue surface are in the subgroup.

In order to decide whether an instance is allowed in the subgroup, we make a set of the allowed cells. The allowed cells form a new grid, defined by the centroid(s) and denoted by $grid(C)$. In Section 4.5 will be explained that a subgroup can have multiple centroids. Therefore, C is a set of centroids. All the cells that are (partly) covered by the circles made arounds the centroids, defined by the (λ, φ) coordinate of the instance, are added to the set. A cell is defined by the coordinates of its corners, $corners(\phi)$. Here, ϕ refers to the ID of a cell (utm_ID) and $corners(\phi)$ is a set of the corners of that cell. For example, if we take the cell with $utm\_ID =$'2B' in Figure 7, $corners$('2B') is defined by {(1.0,1.0),(1.0,2.0),(2.0,1.0) and (2.0,2.0)}. The length of a cell is the same for every cell in the grid and denoted by $dist$. We can now define the list of allowed cells $grid(C)$:

$$\forall_{k \in corners(\phi)} \exists_{c \in C} (x_\varphi^k - x_\varphi^c)^2 + (x_\lambda^k - x_\lambda^c)^2 \leq (r + dist)^2 \Rightarrow \phi \in grid(C) \quad (1)$$

The x_φ^k and x_λ^k refer respectively to the value of the latitude and longitude value of the kth corner. The $dist$ is added to the radius to make sure that also the cells that are partly covered by the circle are added to $grid(C)$. The cell with the ID ϕ is an element of $grid(C)$ if all the corners of the cell

ϕ are in at least the surface made by one of the centroids in C.

Now that we have the set of all the allowed cells, the next step is adding the instances that are covered by these cells to the subgroup. We therefore compare the ϕ's of the instances with the ϕ's of the set. If the ϕ (utm_ID) of an instance is in the set of allowed cells, the instance x^j is allowed in the subgroup:

$$x^j_\phi \in grid(C) \Rightarrow x^j \in s_i \quad (2)$$

For example, if we take the instance with the index 3 as centroid, $C = \{3.1, 5, 2\}$. If we then apply (1) to all the cells (ϕ) in the grid, we get $grid(C) = \{$'3E', '4E', '5E', '2F', '3F', '4F', '5F', '2G', '3G', '4G', '5G', '2H', '3H', '4H', '5H'$\}$. By subsequently performing (2) to all the instances x^j in the dataset, we get the subgroup with the instances visible in Table 4.

index	$mammal\_species$	λ	φ	utm_ID	$features$	
0	rattus norvegicus	1.8	5.1	2F	...	...
2	rattus norvegicus	2.3	7.1	3H	...	...
3	mus domesticus	3.1	6.2	4G	...	...
5	mus domesticus	4.1	5.7	5F	...	...
6	alopex lagopus	4.2	4.8	5E	...	...

Table 4: All the instances in the subgroup with $C = \{(3.1, 5.2)\}$.

As mentioned in Section 2.1 the spatial constraint will be considered as an inductive constraint of the subgroup discovery task. The spatial constraint decides which instances are allowed in the subgroup. Therefore, the spatial constraint should be a part of the subgroup description. For that reason, the description contains not only the conditions but also the centroids of the constraint. Due to the fact that the subgroup keeps expanding each iteration by a maximum of the radius, the description can have multiple centroids. The algorithm will be further explained in Chapter 4. A definition of a spatial subgroup description a could be:

$$X_{max(temp(Nov))} > 17.16 \wedge X_{\mu(temp(Jun))} < 21.22 \ \wedge \ C = \{(2,3),(3,5),(4,5)\}$$

Symbol	Definition
D	Dataset.
$\mathbf{X}$	Dataset of explanatory variables.
X	An explanatory variable of $\mathbf{X}$.
x	The value of sample x for variable X.
$\lvert \cdot \rvert$	Number of elements in a set.
n	Number of examples in dataset D.
m	Number of explanatory variables.
M	Subgroup list model.
$\mathcal{M}$	Subgroup list model.
w	Number of subgroups in M.
s	A subgroup.
a	Description of a subgroup.
$p_{y\|c}$	Probability of category y given description a, i.e., $Pr(y\|a)$.
$\hat{\Theta}$	Maximum likelihood estimation of parameter Θ.
$f(\hat{\Theta}^a, \hat{\Theta}^d)$	Function of differences between distribution Θ^a and Θ^d.
KL_{Cat}	Kullback-Leibler divergence for categorical distributions.
WKL	Weighted Kullback-Leibler divergence general form.
$SWKL$	Sum of Weighted Kullback-Leibler divergences.
r	radius.
λ	x-coordinate (longitude) feature of dataset.
φ	y-coordinate (latitude) feature of dataset.
ϕ	ID of cell.
$corners(\phi)$	set of corners of cell ϕ.
C_i	set of centroids of subgroup i.
$dist$	length of a cell.
$grid(C_i)$	set of allowed cells of subgroup i.

Table 5: Notation table

2.3 Formal problem statement

Remember the problem statement defined in Chapter 1:

Find the optimal list of subgroups that 1) together explain the most relevant deviations in the data with respect to a given target variable and 2) that cohere to the imposed spatial constraint.

It consists of two components and therefore the formal problem statement will also consist of two components.

The first component of the problem statement refers to the definition of subgroup discovery. The deviation in the data with respect to a given target variable is calculated by a quality measure. This thesis makes use of the Weighted Kullback-Leibler divergence. The bigger this number, the more the subgroup deviates from the dataset and therefore the better the quality of the subgroup. The quality of a subgroup list is given by the SWKL. The goal is to find the model M with the highest SWKL out of all the possible models $\mathcal{M}$, because that implies the optimal list of subgroups. Therefore the first component of the problem statement can be formalised as:

$$1)\ arg\_max_{M \in \mathcal{M}}[\text{SWKL}(M)]$$

The second component of the problem statement refers to the spatial constraint. In Section 2.2 we defined when an instance is allowed to the formed subgroup. For all the instances in the subgroup i should hold that its utm_ID is in the set of allowed cells of that subgroup. This can be formalised to:

$$2)\ \forall_{s_i \in M} \forall_{x^j \in s_i} :\ x^j_\phi \in grid(C_i)$$

It says that for every subgroup i in the model M and for all the instances x^j in subgroup i the utm_ID value of the instance x^j should exist in the set of allowed cells ($grid(C_i)$).

Now that we have the two components we can define the formal problem statement as:

$$arg\_max_{M \in \mathcal{M}}[\text{SWKL}(M)],$$
$$\text{s.t. } \forall_{s_i \in M} \forall_{x^j \in s_i} :\ x^j_\phi \in grid(C_i)$$

The objective is thus to return the subgroup list M that maximizes the combined quality of the subgroups and that adheres to the spatial constraint.

3 Related Work

In this section the most relevant related work will be mentioned. Firstly, the most relevant papers about subgroup discovery will be mentioned. Then, some papers about spatial data mining and clustering will be shortly described. Finally, a paper in which classification is used for subgroup discovery will be mentioned

3.1 Subgroup discovery

The spatial subgroup discovery algorithm in this thesis will be based on subgroup discovery, also known as top-k subgroup mining. Atzmueller describes it as top-k subgroup mining because the applied subgroup discovery algorithm can return a result set containing those subgroups above a certain minimal quality threshold or only the top-k subgroups [13]. These top-k subgroups are decided by a quality measure. The top-k mining algorithm beam search will be used for the spatial subgroup discovery algorithm and will be further explained in Section 4.1.

Proença et al. uses the Minimum Description Length (MDL) principle to find the best subgroup list [18]. Besides that, the paper mentions that when the subgroup list only contains one subgroup, it corresponds to top-1 subgroup discovery with the Weighted Kullback-Leibler divergence as a quality measure. In this thesis we decided to work with the Kullback-Leibler divergence. Furthermore, Proença et al. is the first one to address two subgroup discovery challenges at the same time; subgroups should not only stand out with respect to the target attribute, but should also be statistically robust and non-redundant [18]. The paper tackles the problem of finding interesting subgroups arising out of coincidences. The subgroups should therefore be against multiple hypothesis testing and be statistically robust by themselves.

Van Leeuwen and Knobbe already researched the problem of the redundancy of subgroup sets mined; the fact that subsets with the highest deviation according to a certain quality measure tend to cover the same region of the dataset with slight variations in their description of the subset [21]. As a solution, subgroup set mining is proposed in which not individual subgroups are considered, but only sets of subgroups. This thesis compares the performance therefore on the quality of subgroup sets instead of on individual subgroups. Furthermore, in Section 4.3 a manner to reduce the redundancy in the spatial subgroup discovery algorithm is explained.

3.2 Spatial data mining

Anuradha et al. names the importance of spatial clustering towards the decision making process [3]. For example, in the public safety measures spatial clustering is able to identify urban activity centers in a dataset of a city that contains spatial information, etc.

Kolatch used an unsupervised data mining task to cluster spatial datasets [9]. Regular data mining differs from spatial data mining due to the existence of spatial data. This spatial data gives information about the space occupied by objects. In the case of this thesis we work with spatial data that consists of geometric discrete information. The paper says that “the attributes of a spatial object stored in a database may be affected by the attributes of the spatial neighbors of that object.

In addition, spatial location, and implicit information about the location of an object, may be exactly the information that can be extracted through spatial data mining" [9]. It would therefore be interesting to add spatial constraints to subgroup discovery to find out if better subgroups can be found.

The paper describes the different spatial clustering algorithms and compares them on six factors which are necessary for effective clustering of large spatial datasets. Identifying irregular shapes and handling data with higher dimensionality are examples of factors. Clustering tries to create a group of objects that is organized on some similarity among the members. In the perspective of spatial datasets, "clustering permits a generalization of the spatial component that allows for successful data mining" [9]. The best performing method according to the 6 factors is the bottom-up hierarchical clustering algorithm CURE.

3.3 Clustering

Van Leeuwen et al. introduces two algorithms for identifying the different distributions in a transaction database [22]. These different distributions can be defined as components. The paper shows that highly characteristic components are identified by using two MDL-based algorithms that follow orthogonal approaches [22].

The main reason why this paper is interesting for this thesis is due to the experiment with the *mammals* dataset. In Figure 1 the best found decomposition is visible. No spatial information was used for the decomposition. The instances of the *mammals* dataset represent the cell location of 50 x 50 km. Each instance has been assigned to one of the six components based on the mammals that are recorded at that location. As mentioned in Chapter 1 this thesis introduces spatial information to only allow cells in the subgroup that are geographically linked. Therefore, the unsupervised algorithm that is used in the paper is transformed to a supervised algorithm with spatial constraints.

3.4 Classification & subgroup discovery

Similarly to subgroup discovery, classification is a supervised data mining task. The goal is to "build a concise model of the distribution of class labels in terms of predictor features. The resulting classifier is then used to assign class labels to the testing instances where the values of the predictor features are known, but the value of the class label is unknown"[10]. In this thesis classification will be used but with a different purpose. The goal of the use of classification in this thesis is not to predict the target value of new instances, but to know which features are the most important ones to classify the instances.

A subgroup description contains conditions. Due to the fact that there will be worked with a dataset that has approximately 80 features, it is useful to know which features are important. That way more logical conditions can be defined. A measure for the importance of the features is the entropy. It measures the disorder of a feature and it is a number between 0 and 1 [1]. A high entropy (1) means a high level of disorder. In Figure 8 the proportion of data points belonging to the positive class versus the relative entropy is visible.

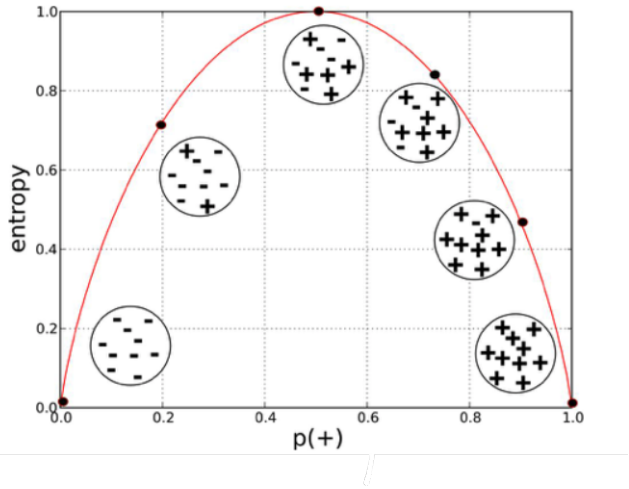

Figure 8: Proportion target variable versus entropy. The least amount of disorder can be found bottom left and bottom right of the figure. That corresponds to an entropy nearby 0 [19].

We want the feature to split the data with the least amount of disorder, or as 'pure' as possible. This results in the highest information gain; the additional information that the feature provides [1]. A decision tree is an example of a classifier and will be used as classifier in this thesis. Each node of the tree represents a feature. A decision tree is made by calculating the information gain of each feature. The feature with the highest information gain is chosen as node. Some features will have an higher information gain than others. This can differ each node. A way to express the combined relative information gain of the features is by the feature weights. This provides a ranking of the features that contribute most to the classification model [12]. The spatial subgroup discovery algorithm uses the features with the highest feature weights to set the conditions. This idea is based on the paper 'Maximal exceptions with minimal descriptions' [20] in which classification is used for description minimisation. In order to find subgroups that are both exceptional and interesting, the paper proposes two information-theoretic measures: one based on the Kullback–Leibler divergence (also used in this thesis), and the other on Krimp. The paper focuses on exception maximisation and description minimisation. The reason why this paper is relevant for this thesis is due to the use of classification for the description minimisation.

The model described in the paper starts with a candidate subgroup and improves it each iteration. One of the two steps that is done each iteration is a step for the description minimisation. In this step the subgroup descriptions need to be specified, how to find and minimise them. The paper explains that the mapping from the description data and a subgroup to {0,1}, can be regarded as a binary classification task. The classification model RIPPER is used in the paper.

4 Spatial Subgroup Discovery Algorithm

In this section the algorithm of the spatial subgroup discovery will be explained. The algorithm returns the subgroup list that maximizes the SWKL and adheres to the spatial constraint. The algorithm makes use of the idea behind beam search combined with separate and conquer. Therefore, beam search in the context of subgroup discovery will be firstly explained. Subsequently, the spatial subgroup discovery algorithm will be explained as in the steps mentioned in Section 4.1.

4.1 Beam search

In comparison with the best-first search algorithm, the beam search algorithm reduces the space complexity by expanding the most promising nodes in a limited set [14]. Beam search is therefore a heuristic algorithm and the limited set is decided by the beam width. When applied to subgroup discovery, the algorithm starts with candidate subgroups of size one and iteratively refine a subset by adding one more condition per iteration of those to subgroups to a larger length [18]. The refinements are decided by a quality measure. In our case it is the Weighted Kullback-Leibler divergence.

The spatial subgroup discovery algorithm makes use of three hyperparameters:

- Beam width: decides how many subgroups are allowed for further refinements **and** how many conditions are set.
- Maximum search depth: decides how many conditions are allowed as maximum to a subgroup.
- Radius: the *dist* plus the radius decides the cells that are allowed in $grid(C)$.

A visualization of the beam search algorithm can be seen in Figure 9:

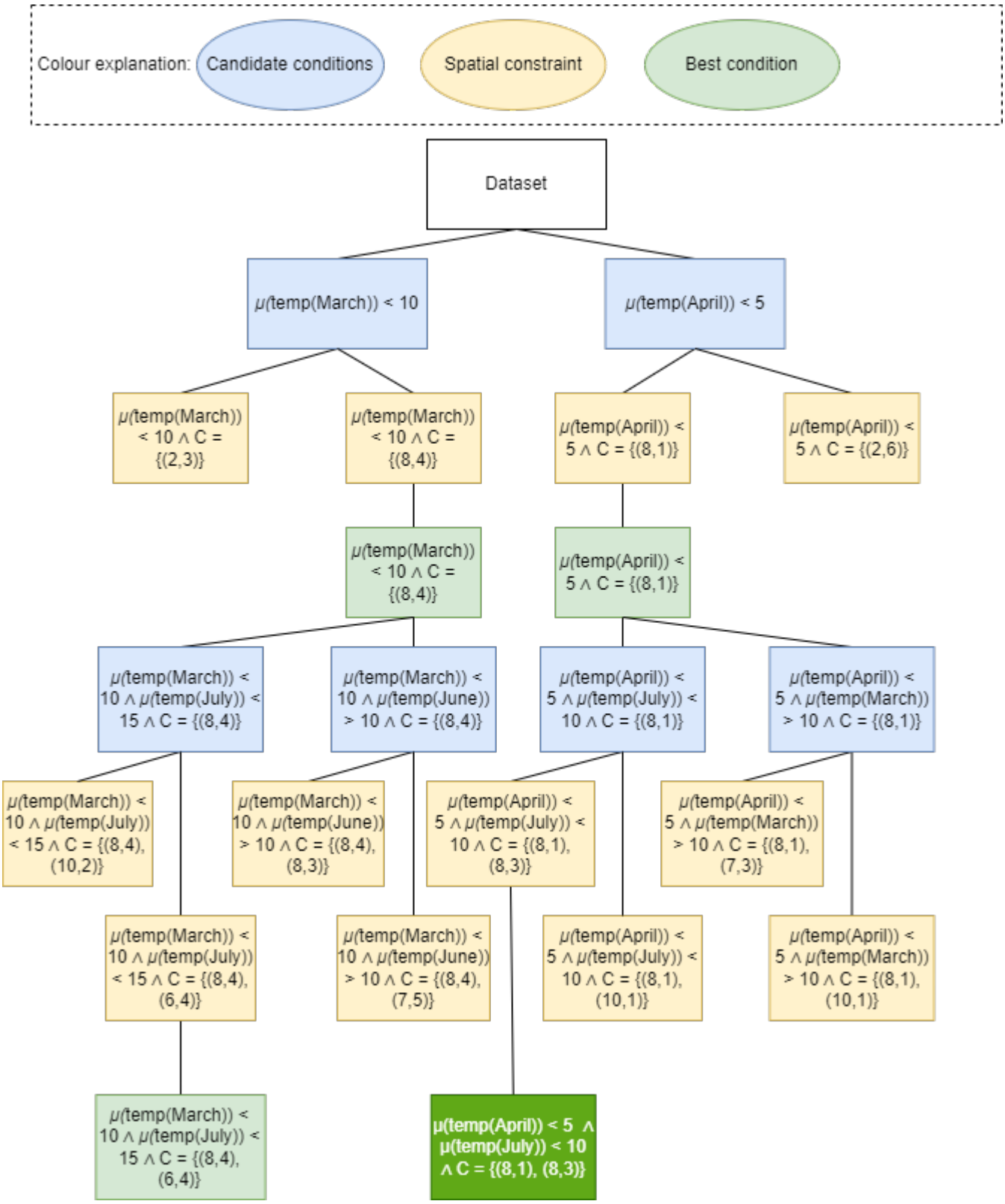

Figure 9: Visualization of the beam search algorithm. The best condition is selected based on quality. The beam width and maximum search depth are both set to 2.

The top level of Figure 9 represents the whole dataset. The first step (Section 4.2) is then defining the conditions. In the figure it is visible that two conditions are defined. A condition is for example: $\mu(temp(March)) < 10$ where $\mu(temp(March))$ is the average temperature recorded in March.

The second step (Section 4.3) is adding the spatial constraints on each one of the refinements. As mentioned in Section 2.2 a constraint consists of a centroid and a predefined radius. A random instance out of the data which covers the condition is chosen as centroid. All the instances that are in the cells, which are covered or partly covered by the radius, belong then to that subgroup. In Figure 9 two centroids are chosen and thus two subgroups are made out of each refinement.

The third step (Section 4.4) is measuring the quality of all the acquired subgroups. These subgroups will subsequently be ranked according to their quality. The predefined hyperparameter beam width decides how many of these subgroups are allowed for further refinements. The beam width in Figure 9 is two and therefore the first two subgroups in the ranked list are allowed for further refinements.

Then, a new iteration can start (Section 4.5) with the two subgroups found after the just mentioned step three. The algorithm keeps iterating until the max search depth is reached or the quality does not improve. In the figure the max search depth is two and therefore only two conditions are allowed. If the dark green subgroup has the best quality, it is returned.

4.2 Step 1: Feature selection & conditions

The first step of the spatial subgroup discovery algorithm is defining the conditions. Each condition is based on a feature. In order to make a division of the dataset we will use a classification algorithm to measure the importance of each feature.

As classifier the decision tree will be used. The features that contain spatial information are not considered for the classification. Instead of using the decision tree to make a model to predict the target variable on a new dataset, we are going to calculate the importance of each feature. This can be expressed by the weight of the feature. The higher the weight, the more important the feature is. After measuring the weight of each feature, they are ranked from highest weight to lowest weight. We have chosen that the parameter that decides how many conditions are made, has the same value as the beam width. Therefore, the beam width does not only decide the beam selection (Section 4.4), but also the number of conditions. If the beam width equals 5, the top 5 features of the ranking are taken for the conditions.

Now the features are decided on which the conditions will be set. The next step is deciding on which value of the feature the dataset will be split. In order to split the dataset in half of the instances the median value of the feature is calculated. The conditions are then alternately higher and lower than the median value of the feature. For example, if the beam width equals 5, there will be made 5 subgroups with 5 different conditions which are specified by alternately higher and lower than the mean value of the feature.

In Figure 10 an example of this step is visible. The dots represent the instances. The right figure is the made subgroup and is the result of adding a condition.

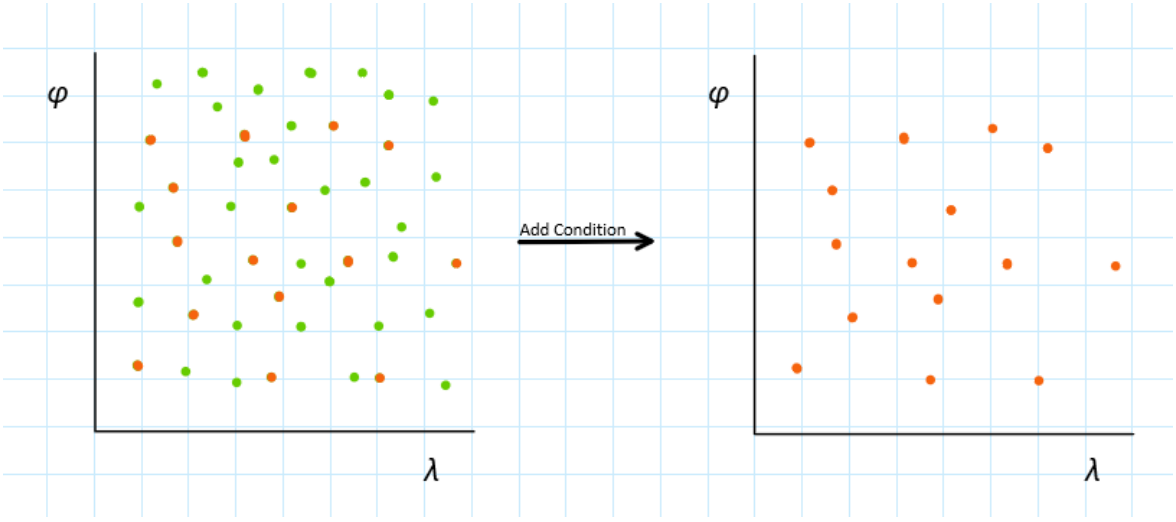

Figure 10: Example of *Step 1*. Assume that the green instances have a maximum temperature above 20 and the orange instances below 20 in November. If the condition $max(temp(Nov)) < 20$ is set, the orange dots are the resulting subgroup.

4.3 Step 2: Spatial constraints

The second step is imposing the spatial constraints on the just made subgroups. In order to reduce the runtime , the parameter *n_samples* decides how many instances will be sampled to impose the spatial constraint on. The sampling of the instances has one restriction: in order to prevent instances that are close to each other, and therefore create almost the same subgroups, only instances are allowed that have at least a distance of the predefined radius from the already chosen instances. This holds only for the first iteration, because after the first iteration the made subgroup is much smaller than the whole dataset. Finding instances that have at least a distance of the radius from each other in the made subgroup is impossible. Therefore, the distance is defined by a fraction of the radius.

Now that we have the sampled instances, it is time to impose the spatial constraint on them. A circle with the predefined radius is made around all the sampled instances. All the instances that are in the covered or partly covered cells by the circle are allowed in the subgroup. The subgroup has as description the condition and the centroid of the used instance. It is allowed that one instance belongs to multiple subgroups if it is covered by multiple radii. Figure 11 gives an example of *Step 2*.

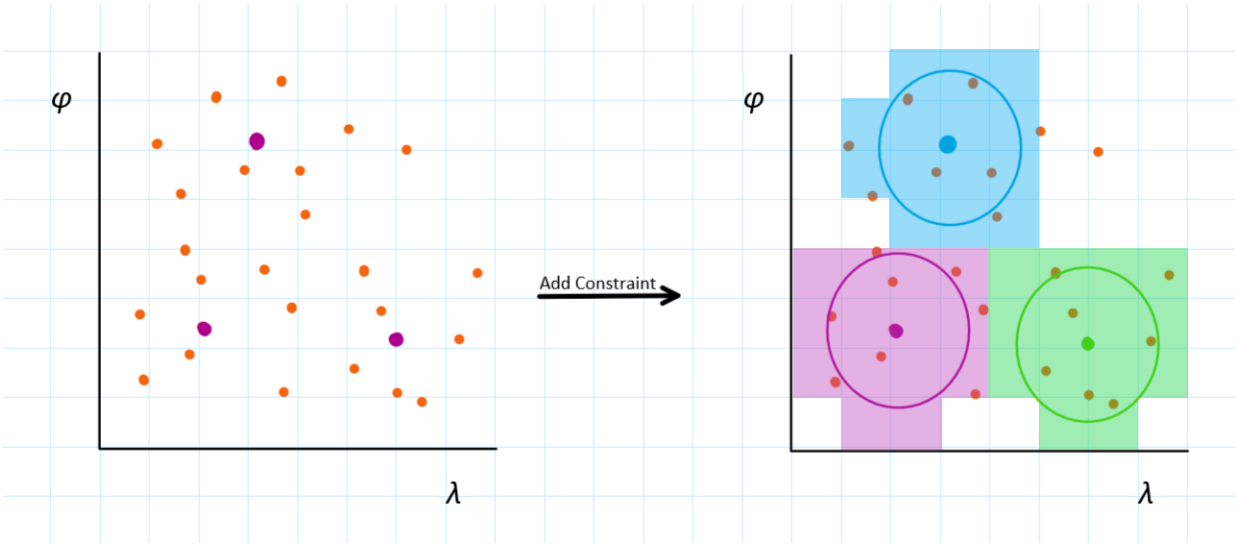

Figure 11: Example of *Step 2*. The three purple dots of the left figure are the three sampled instances. The figure on the right is the result of imposing the spatial constraint with a radius of 1.5. You can observe the three made subgroups. Remember that all the instances are allowed in the subgroup even if the circle only partly covers the grid. To make this clear all the grids belonging to the subgroup with their corresponding centroid are marked with the same colour. For example, all the instances that are in the pink marked grids belong to the subgroup with centroid (2,2.5). The description of a subgroup now contains one condition and one centroid.

4.4 Step 3: Quality measure

The third step is measuring the quality of the subgroups. This will be done with the Weighted Kullback-Leibler divergence. With the Weighted Kullback-Leibler divergence the distribution of the subgroup is compared with the distribution of the whole dataset and the quality of the subgroup is returned. The subgroups are subsequently ranked by their quality. The beam width hyperparameter decides how many subgroups are allowed for further refinements, which are explained in Section 4.5. Figure 12 illustrates this step with a beam width of 2.

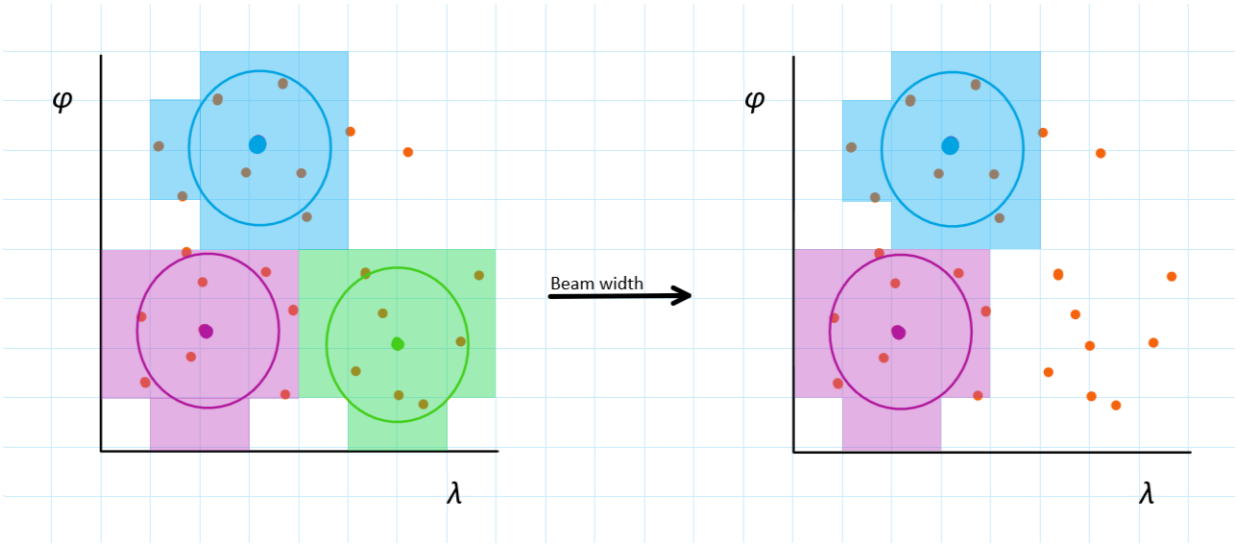

Figure 12: Example of *Step 3*. The left figure represents the three made subgroups after performing *Step 2*. Subsequently, the quality of these three subgroups is calculated. Assuming that the blue and pink subgroups have the highest quality, they are allowed for further refinements. The result of performing *Step 3* is the right figure.

4.5 Top-k subgroup selection

After performing steps 1, 2 and 3 of the algorithm, the top-k subgroups are returned. The top-k subgroups are allowed for further refinements. There exist two different refinements; performing steps 2 and 3, or performing steps 1, 2 and 3.

If the quality of a subgroup can be improved by only performing *Step* 2 and 3, this will be done. If that is not possible anymore, *Step 1* will be performed again. This implies that the number of centroids of a subgroup can be bigger than the number of conditions. After each refinement a centroid or a centroid and a condition are added to the subgroup description. Remember that only performing *Step 1* has influence on the maximum search depth.

Performing *Step 2* again requires an extra explanation. Instead of having one big dataset to set the conditions on as described in Section 4.2, there are now multiple smaller subgroups consisting of different conditions. An example of performing *Step 1* again is given in Figure 13.

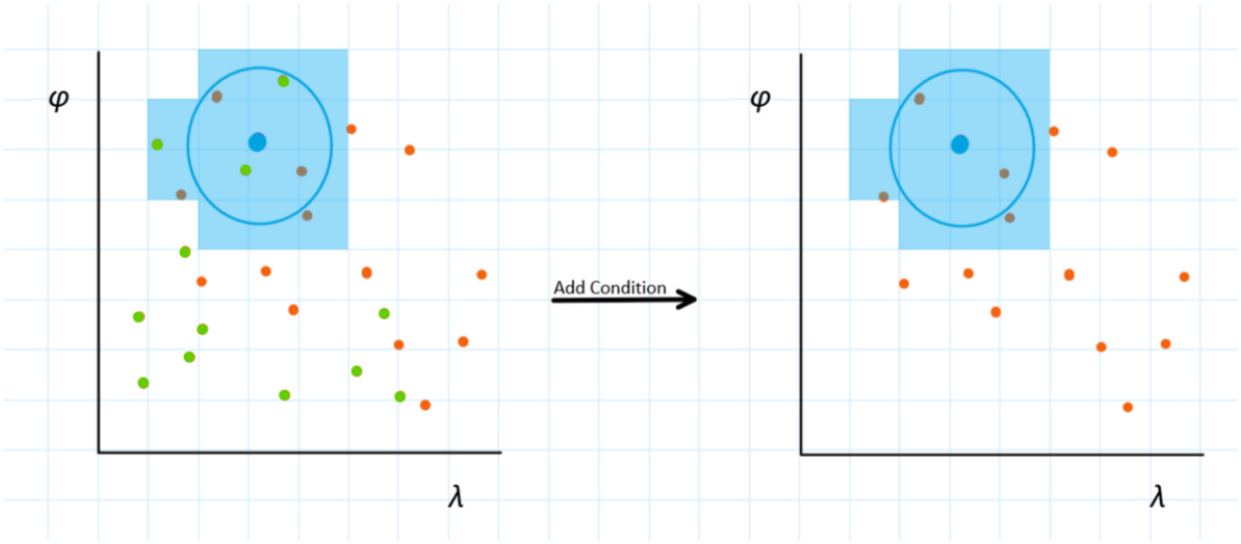

Figure 13: Example of *Step 1* in second iteration. A subgroup with a condition and centroid is seen on the left figure. Assume that the green instances have a mean temperature below 10 and the orange instances above 10 in June. If the condition $\mu(temp(Jun)) > 10$ is set, the blue marked dots in the right figure is the resulting subgroup.

After performing *Step 1* from the second refinement we are left with subgroups whose descriptions consist of two conditions and one centroid. To these subgroups we are adding a centroid by sampling an instance within that already existing subgroup. The sampled centroid has the same predefined radius. All the instances that are within the cells that are (partly) covered by the circle of the new centroid are added to the subgroup. Therefore, the subgroup is expanded. Figure 14 gives an example.

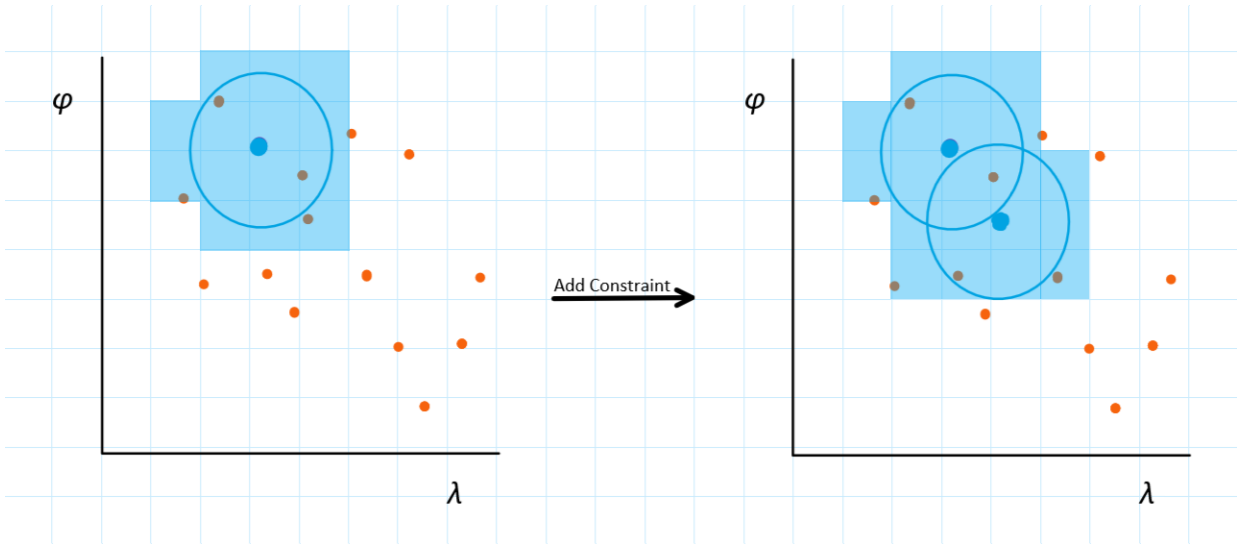

Figure 14: Example of *Step 2* in second iteration. The left figure represents the subgroup with two conditions and one centroid. To this subgroup another centroid is added (the other blue dot). The subgroup gets expanded due to this added centroid. This is visible by the added blue surface on the right figure.

Repeating *Step 3* of the refinement is the same as explained in Section 4.4. The top-k subgroups are returned after which another iteration can be done. If the maximum search depth is reached or the quality does not improve anymore the algorithm stops and it returns the subgroup with the best quality. This subgroup is subsequently added to the subgroup list M and separated from the dataset. Then, the spatial subgroup discovery algorithm can be repeated with the rest of the dataset. The result is that the subgroup list keeps expanding with the best found subgroups.

5 Experiments

In this section we analyze the pre-processing of the dataset and the setup of the experiments. In addition, we evaluate the performance of the spatial subgroup discovery method proposed in Chapter 4. The spatial subgroup discovery algorithm has three hyperparameters; the beam width, maximum search depth and the radius. On these three hyperparameters will be experimented. Furthermore, an anecdotal evidence and interpretation will be given. In order to understand the experiments, the dataset will be explained firstly in Section 5.1.

5.1 Dataset and pre-processing

The dataset that will be used for the experiments is the atlas of European *mammals* dataset [4]. It contains the presence of 194 mammal species within Europe. Each row of the dataset represents a grid cell. The feature of a grid cell is the utm_ID. The cell resolution of a grid is approximately 50 times 50 km, and the grid system is based on the Universal Transverse Mercator (UTM) projection and the Military Grid Reference System (MGRS). For each grid the latitude (φ) and longitude (λ) are defined. The 194 remaining attributes represent the different species. If the value is 1, it indicates presence of the mammal, 0 is absence. An example of the dataset is given in Table 6:

φ	λ	utm_ID	acomys minous	alces alces	alopex lagopus	...
39.52	-31.55	25SFD1	0	0	1	...
38.61	-29.01	26SLH1	1	1	0	...
...	...	...	...	...	...	...

Table 6: Mammals dataset

Due to the fact that we will make use of subgroup discovery, we need a target variable. The target variable of the *mammals* dataset is the mammals species. As you can observe in Table 6 there are multiple columns that define which mammals are in the cell. In order to have only one column as target attribute, the dataset is reorganized. Now each recorded mammal represents a row, see Table 7 for an example of this reorganization.

mammal	φ	λ	utm_ID
rattus norvegicus	38.61	-29.01	26SLH1
rattus rattus	38.61	-29.01	26SLH1
erinaceus europaeus	38.61	-28.44	26SLH3
mus domesticus	38.61	-28.44	26SLH3
...	...	...	...

Table 7: Mammals dataset with one target attribute

The spatial dataset that will be combined with the *mammals* dataset is the Worldclim[1] global climate dataset [8]. This dataset contains climate information about each cell. The attributes are

[1] http://www.worldclim.org/

therefore again the latitude, longitude and the utm_ID. The remaining attributes are the average monthly mean temperature, average monthly minimum temperature, average monthly maximum temperature, average monthly precipitation and the bioclimatic variables derived from the *tmean*, *tmin*, *tmax* and *precipitation*. The dataset gives the worldclim value and the utm value of each feature. For example, a column gives the worldclim mean temperature of March and another column gives the utm mean temperature of March. The utm value represents the average of the values of the worldclim squares that coincide within the UTM square. The worldclim value is the exact worldclim square value. Therefore, we only use the columns that give the utm value. To combine the *mammals* dataset with the worldclim dataset, we merge the datasets on the utm_ID. See an example of the new dataset in Table 8:

Descriptive			Spatial			Target
$\mu(temp(Jan))$	$\mu(temp(Feb))$	...	φ	λ	utm_ID	mammal
12.6	12.1	...	38.61	-29.01	26SLH1	rattus norvegicus
12.6	12.1	...	38.61	-29.01	26SLH1	rattus rattus
12.88	12.36	...	38.61	-28.44	26SLH3	erinaceus europaeus
12.88	12.36	...	38.61	-28.44	26SLH3	mus domesticus
...	...	...	...	...	...	...

Table 8: Mammals dataset merged with Worldclim dataset

However, for some of the utm_ID's in the *mammals* dataset there was no climate information available. The utm_ID's that have no climate information are therefore given the mean value of the utm_ID's in which the first five characters of that utm_ID occur. This applies to 159 out of 83586 recorded mammals. The first five characters are taken because these grids are near each other and have therefore similar climate information. Unfortunately, this rule doesn't cover all the grids that have no climate information. Of some of the grids there is also no climate information available if we only take the first five characters. For these grids the mean values of the first four characters are taken. This rule applies to 31 out of the 83586 recorded mammals. The grids that still have no climate information are from the dataset, this applies to 9 out of the 83586 recorded mammals. The map of all the instances is given in Figure 15.

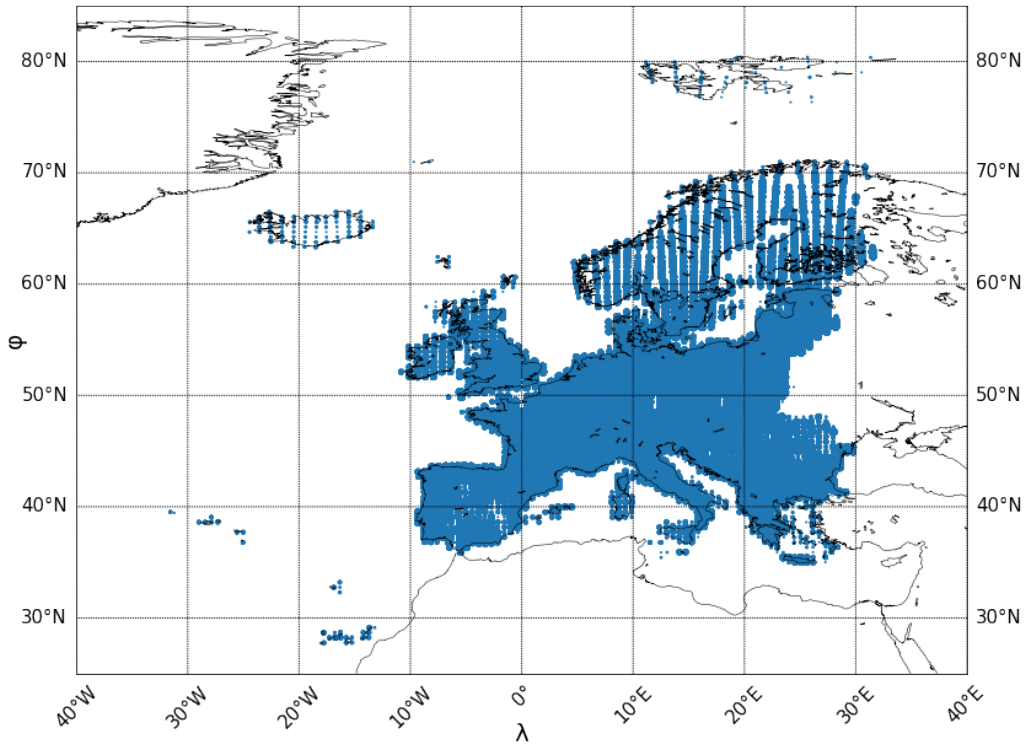

Figure 15: Map of Europe with the coordinate of each instance of the used dataset in blue.

This dataset has a limitation. In Section 4.3 we described that we want to sample the instances by selecting only the instances that are at least a fraction of the radius apart from each other. When trying this on the *mammals* dataset, it is not able to find instances. This is because a lot of the instances share the exact same location. Therefore, for this dataset the number of samples taken has a maximum of the different locations within the subgroup. If for example a subgroup has only three different locations, the maximum of taken samples is three.

5.2 Experimental evaluation

In this section the experiments will be shown. The effect of the spatial subgroup discovery hyperparameters on the discovered subgroup lists will be studied. The results will be evaluated by looking at the sum of Weighted Kullback-Leibler divergences (SWKL) of each subgroup list and the runtime. The hyperparameters of the algorithm are the beam width, maximum depth and the radius. We change the parameter of interest whilst two of the parameters stay fixed. The fixed values of the hyperparameters are:

- *beam_width* $= 5$ This parameter decides the number of made subgroups in *Step 1* (see Section 4.2) and the number of subgroups that are allowed for further refinements in *Step 3* (see Section 4.4). Proença et al. [18] has set this parameter to 100. However, this algorithm has at least the number of conditions times the number of taken samples more subgroups each iteration. It is "at least" because after a condition is set, multiple centroids can be added in step 2 (Section 4.3) to the subgroup. In order to reduce the runtime, this parameter is therefore set to 5.

- $max\_depth = 4$ The maximum depth decides the maximum amount of conditions that can be set on a subgroup. Proença et al. [18] has set this parameter to 5, but mentions that an average number of conditions above 4 barely occurs. Besides that, setting more than 4 conditions on a subgroup with a spatial constraint does not allow enough instances in a subgroup. Therefore, the maximum depth is set to 4.

- $radius = 5$ The radius defines the spatial constraint. If we want to find subgroups that are approximately the size of Spain, the radius should be 5. The radius should not be too big, because then the constraint does not make sense. The radius should also not be too small. The number of allowed instances in the subgroup is then not enough to allow a good quality.

The spatial subgroup discovery algorithm uses two more parameters. The first one is the number of samples taken. This parameter is used in *Step 2* (Section 4.3) of the algorithm. The higher the value of this parameter, the more subgroups will be found. However, increasing the value of this parameter does not work for this dataset. The result will be that the samples taken share exact the same location, which consequently results in the same found subgroups. Therefore, the number of samples taken is fixed at 3. The second parameter is the number of subgroups in the final list. In order to make a comparison possible with the clustering algorithm [22] visible in Figure 1 which was performed on the same dataset, the parameter is fixed at 6.

The next sections show the results of the experiments varying one hyperparameter each time. Each combination is run 5 times. The dots in the figures of the SWKL represent one run. Subsequently, in Section 5.2.4 we will look at individual subgroups found and give interpretation to them.

5.2.1 Beam width

The first hyperparameter that will be experimented on is the beam width. For the experiments we used $beam\_width = [1, 2, 3, 4, 5, 6, 7]$. The result of varying the beam width against the SWKL is shown in Figure 16 and against the runtime in Figure 17.

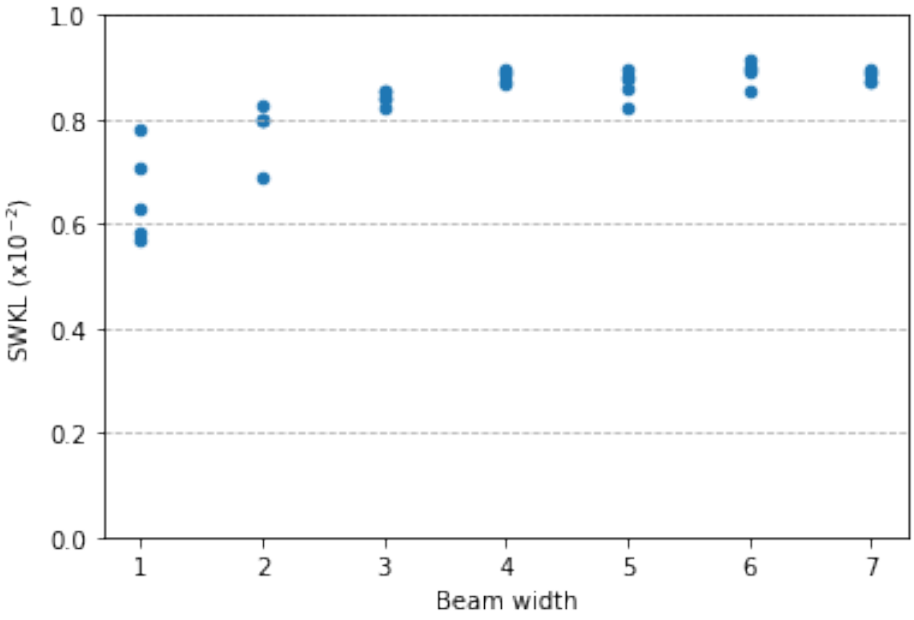

Figure 16: Experiment on beam width (SWKL). It can be observed that the quality of the subgroup list improves as the beam width increases. As a higher beam width provides more candidate subgroups, the chance of finding subgroups with a better quality is higher when compared to the chance of finding subgroups with a lower beam width. However, it can be observed that at a beam width of three the increase of the quality of the subgroup lists slows down. In *Step 1* of the algorithm the conditions are decided on the most important features. If the beam width increases, on more less important features candidate subgroups will be formed. The chance that the subgroups with a more important feature belong to the top-k subgroups of that iteration is bigger than the subgroups with a less important feature. Therefore, the increase of the quality of the subgroup list due to the increase in beam width slows down.

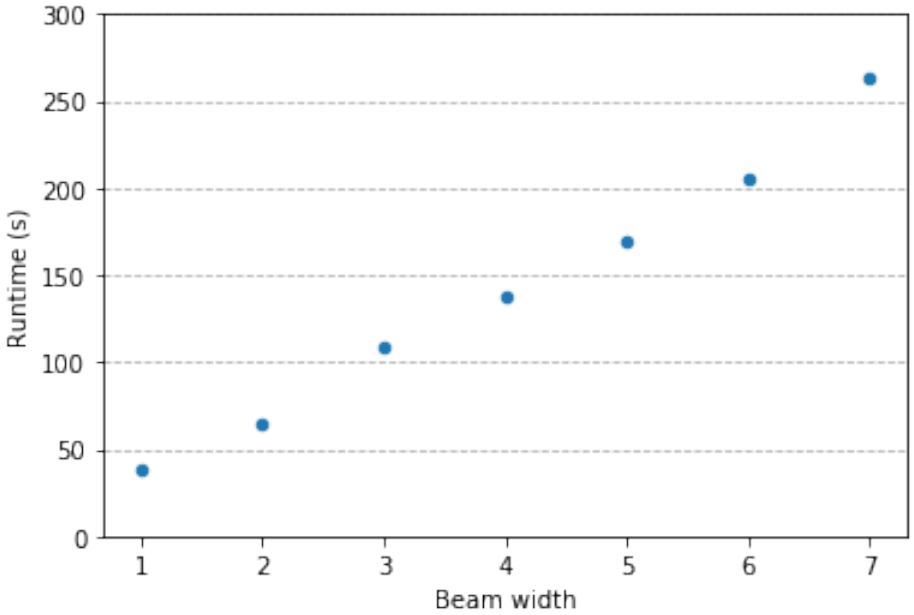

Figure 17: Experiment on beam width (average runtime). It can be observed that the increase in runtime is linear to the increase in beam width. If the value of the beam width increases by one, it means that one extra condition and one extra candidate subgroup will be formed which consequently leads to an higher runtime.

5.2.2 Maximum search depth

The second hyperparameter that will be experimented on is the maximum search depth. For the experiments we used $d\_max = [1, 2, 3, 4]$. The result of varying the maximum depth against the SWKL is shown in Figure 18 and against the runtime in Figure 19.

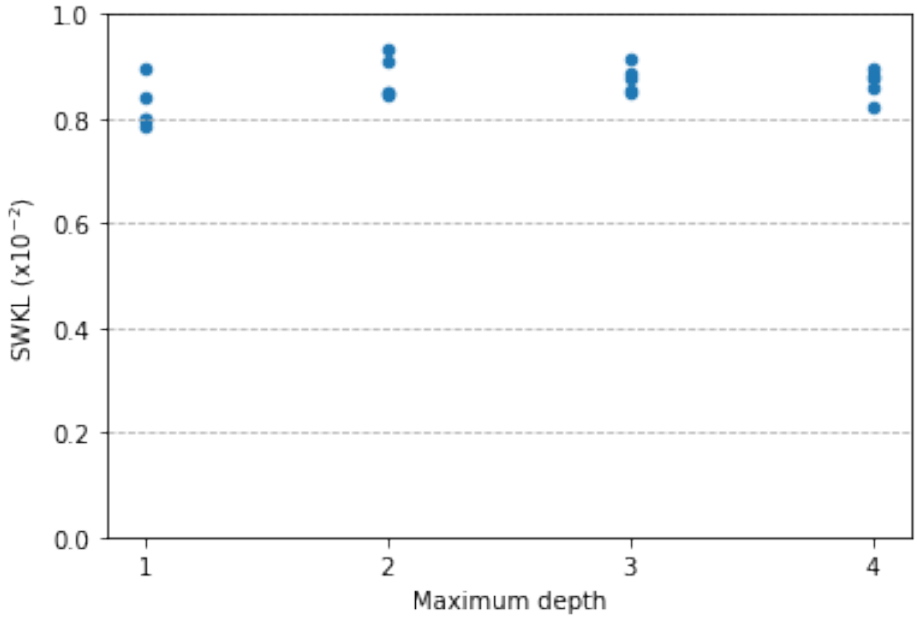

Figure 18: Experiment on maximum depth (SWKL). It can be noticed that as the maximum depth increases, the quality of the subgroup list does not increase.

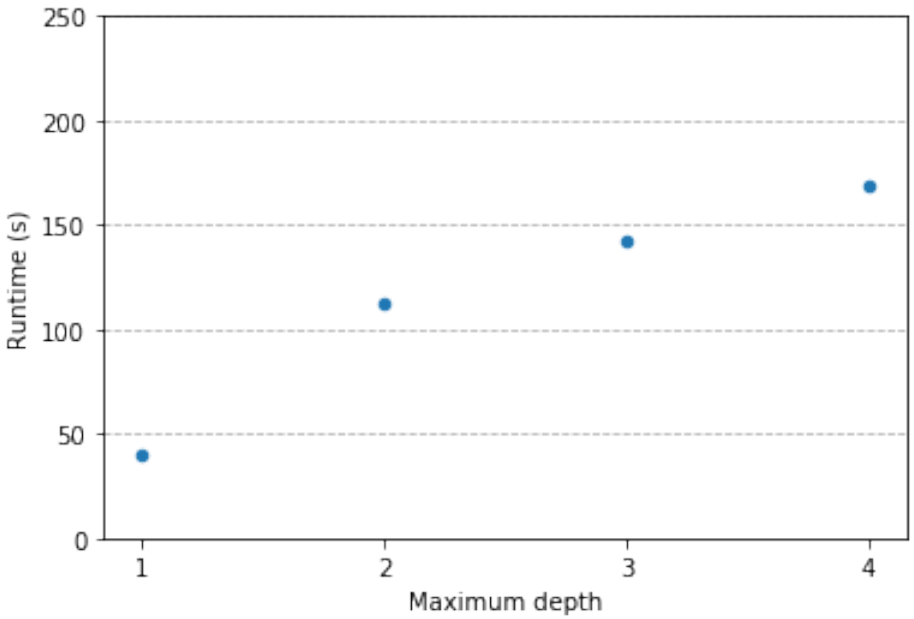

Figure 19: Experiment on maximum depth (average runtime). It can be observed that as the maximum depth increases, the runtime increases. However, the increment of the runtime from the maximum depth of 3 to 4 is less than the previous increments. The quality of some subgroups cannot be increased by adding a fourth condition, which results in a lower runtime

In Figure 18 it is remarkable that as the maximum depth increases the quality stays the same. You could think that the best quality is found when each subgroup has only one condition, and that therefore the quality stays the same after the value of the maximum depth is above 1. It is however not the case that each subgroup in the subgroup list has only one condition for each value of the maximum depth. An example of the subgroups in the subgroup list with a maximum depth of 4 is given in Table 9:

s	description
1	$max(temp(Nov)) > 17.16 \quad \wedge \quad prec(Aug) < 10.58$ $\wedge \quad prec(Jul) < 26.33 \quad \wedge \quad bioclim4 < 682.74$ $\wedge \quad C = \{(41.77, -4.5), (41.33, -8.7), (40.39, -6.32), (39.51, -6.96), (37.27, -8.72)\}$
2	$prec(Nov) > 67.83 \quad \wedge \quad min(temp(Jul)) > 14.77$ $\wedge \quad prec(Aug) < 34.75 \quad \wedge \quad bioclim4 < 685.32$ $\wedge \quad C = \{(43.1, -6.27), (42.2, -6.29), (42.23, -3.3), (37.27, -3.85), (39.07, -9.29)\}$
3	$bioclim15 > 57.85 \quad \wedge \quad prec(Jun) < 28.44$ $\wedge \quad prec(Aug) < 50.5 \quad \wedge \quad bioclim4 < 687.92$ $\wedge \quad C = \{(43.56, 17.17), (39.51, 17.04), (38.17, 15.86), (35.92, 14.72), (35.92, 14.17)\}$
4	$prec(Aug) < 38.22 \quad \wedge \quad prec(Oct) > 74.0)$ $\wedge \quad bioclim7 < 27.06$ $\wedge \quad C = \{(37.69, 23.63), (35.45, 25.07), (36.81, 22.4)\}$
5	$prec(Jul) > 68.5 \quad \wedge \quad C = \{(42.65, 0.28)\}$
6	$max(temp(Jan)) > 8.23 \quad \wedge \quad \mu(temp(Feb)) > 1.12$ $\wedge \quad max(temp(Sep)) > 18.66$ $\wedge \quad C = \{(42.22, 27.91), (43.57, 28.55), (37.72, 26.15), (39.97, 22.46), (40.39, 23.68)\}$

Table 9: This table shows the descriptions of the found subgroups after performing the spatial subgroup discovery algorithm.

Multiple subgroups have more than one condition. As the maximum depth increases the number of conditions increases too. The reason that the quality stays the same is therefore not that each subgroup only consists of one condition. It also depends on the number of centroids. Some subgroups are not able to reach an higher quality by adding another centroid, but other subgroups are and therefore have multiple centroids. This can be observed in Figure 20. The subgroup in the bottom center has for instance only one centroid, whilst the subgroup in the top right has multiple centroids.

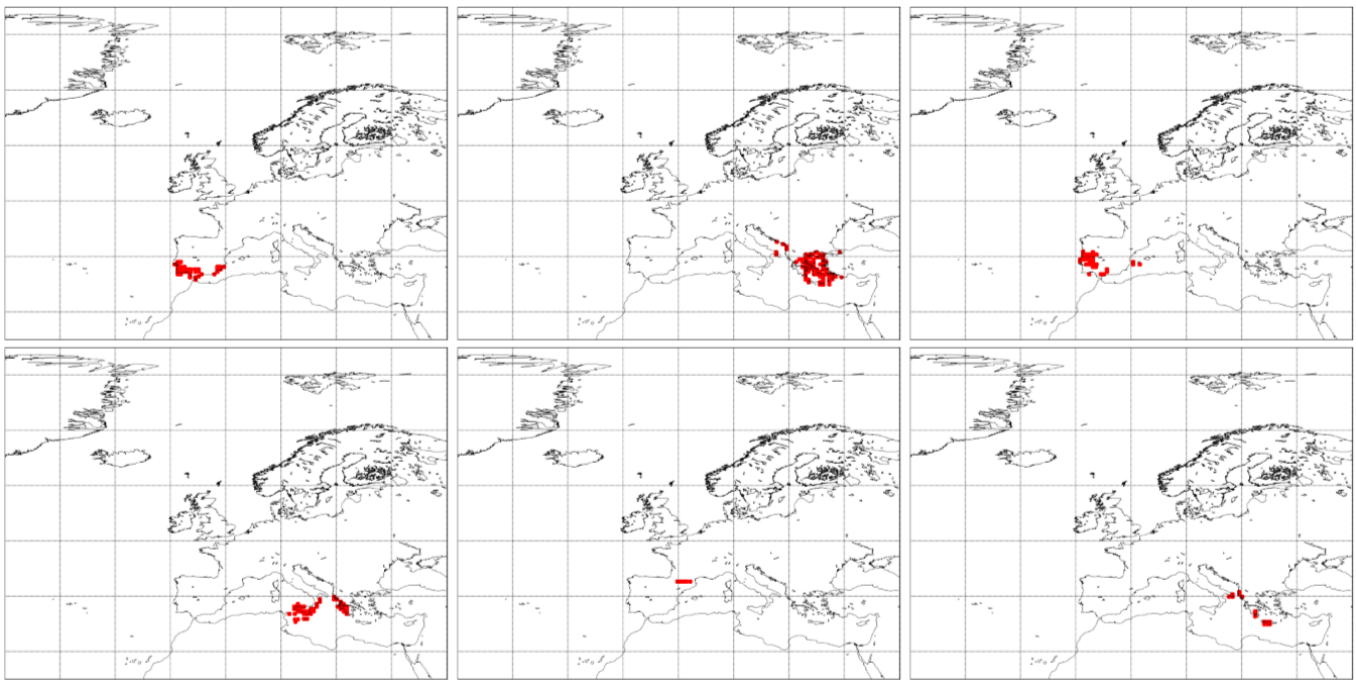

Figure 20: Maps of the subgroups found with $d\_max = 4$. In contrast to the maps in Figure 1, in which the cells are marked, the maps of this figure mark (in red) the coordinate of the mammals belonging to that subgroup. Although it looks like the subgroups are not spatially connected, they actually are. Within the $grid(C_i)$ of a subgroup there will also be cells of which no mammal belongs to the subgroup due to the set condition(s). These cells will not be marked, because only the mammals belonging to the subgroup are marked in the map. Therefore, it can happen that there are multiple red areas but that they are still spatially connected.

5.2.3 Radius

The last hyperparameter that will be experimented on is the radius. For the experiments we used $radius = [1, 2, 3, 4, 5, 6, 7, 8]$. The result of varying the maximum depth against the SWKL is shown in Figure 21 and against the runtime in Figure 22.

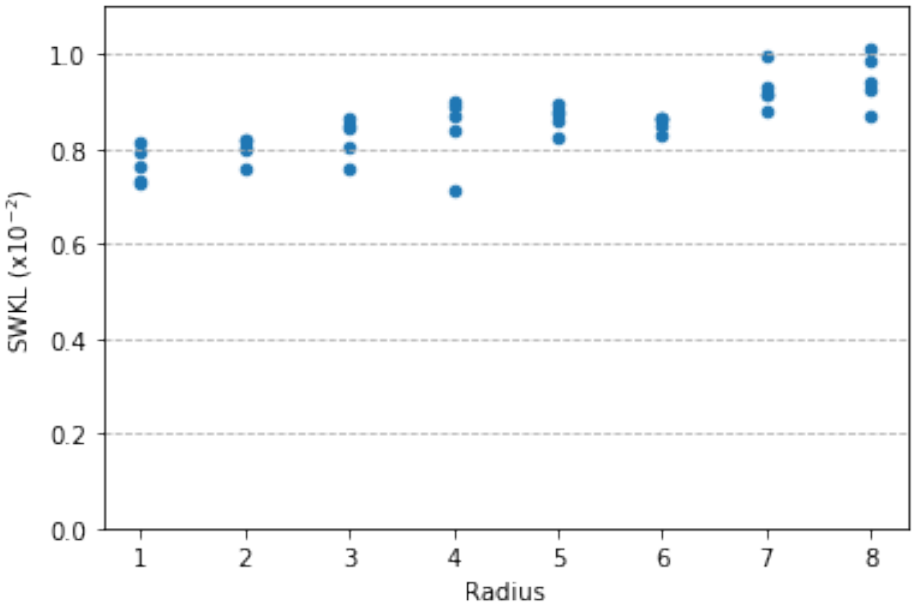

Figure 21: Experiment on radius (SWKL). It is visible that until the radius reaches 5 the quality increases slightly as the radius increases. The radius decides the instances that are allowed in the subgroup. If the radius is higher, more instances are allowed in the subgroup. Due to the fact that the coverage of the subgroup positively influences the quality measure, the quality of the subgroup increases as the radius increases. However, increasing the radius above 5 does not influence the quality as it did below 5. The quality measure also depends on the difference between the target variable distribution of the dataset and the subgroup. Having more instances in a subgroup makes it difficult to differentiate more from the target distribution of the dataset than having less instances.

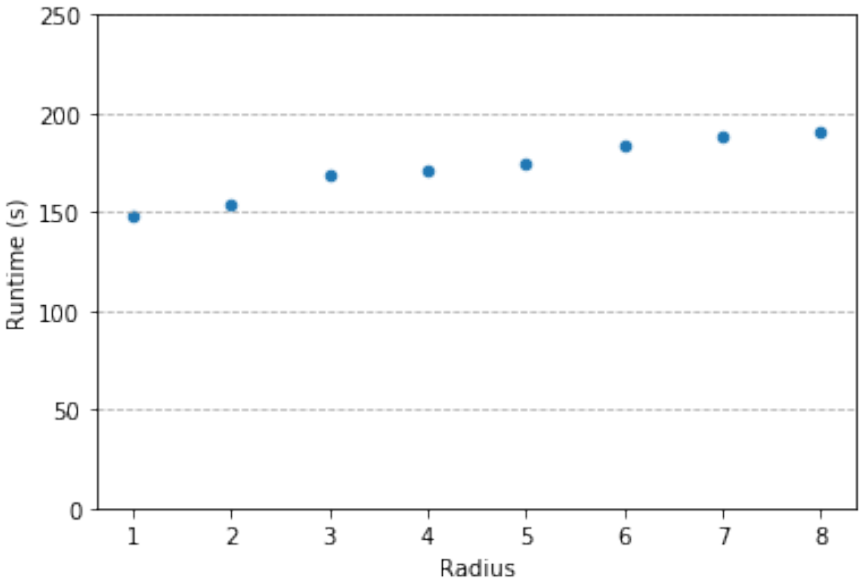

Figure 22: Experiment on radius (average runtime). It can be observed that the runtime increases slightly as the radius increases. A bigger radius corresponds with more instances and thus a bigger runtime.

In Figure 23 the maps of a subgroup list with a radius of 1 can be seen.

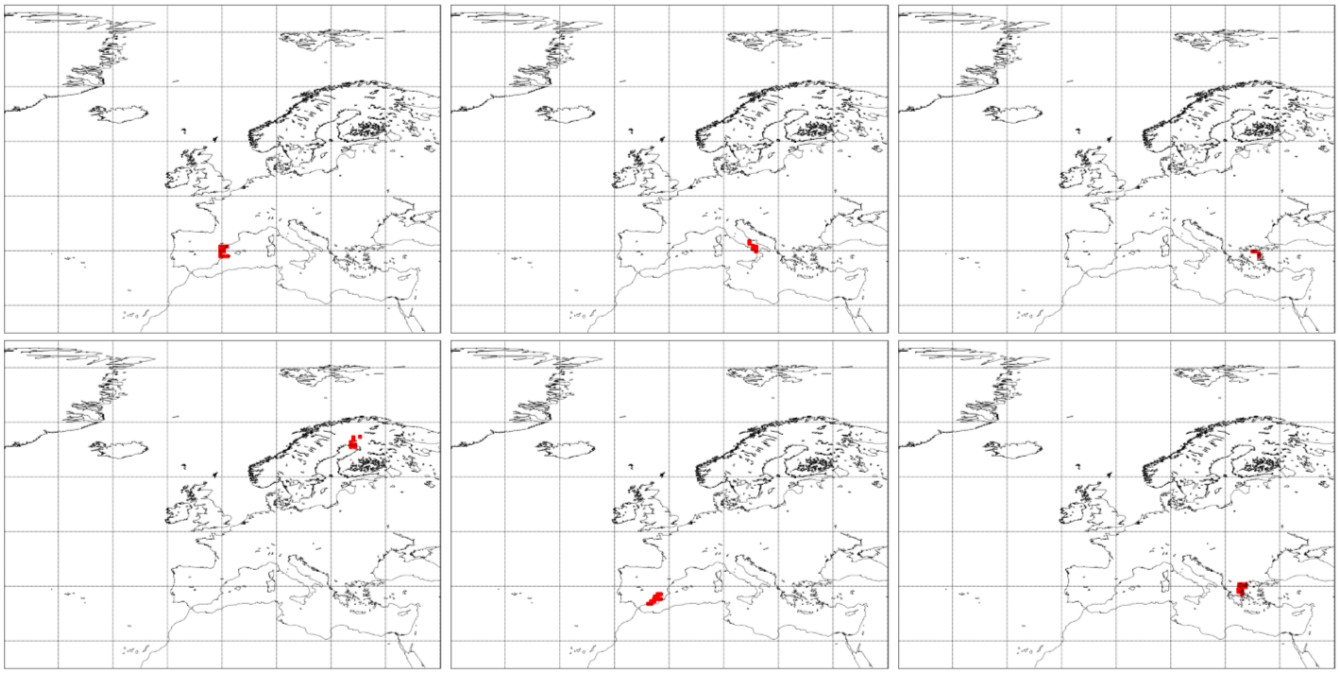

Figure 23: Subgroup list with $radius = 1$. Due to the small radius the subgroups aren't able to expand in further iterations because they are not able to find instances within the radius.

In Figure 24 the maps of a subgroup list with a radius of 8 can be seen.

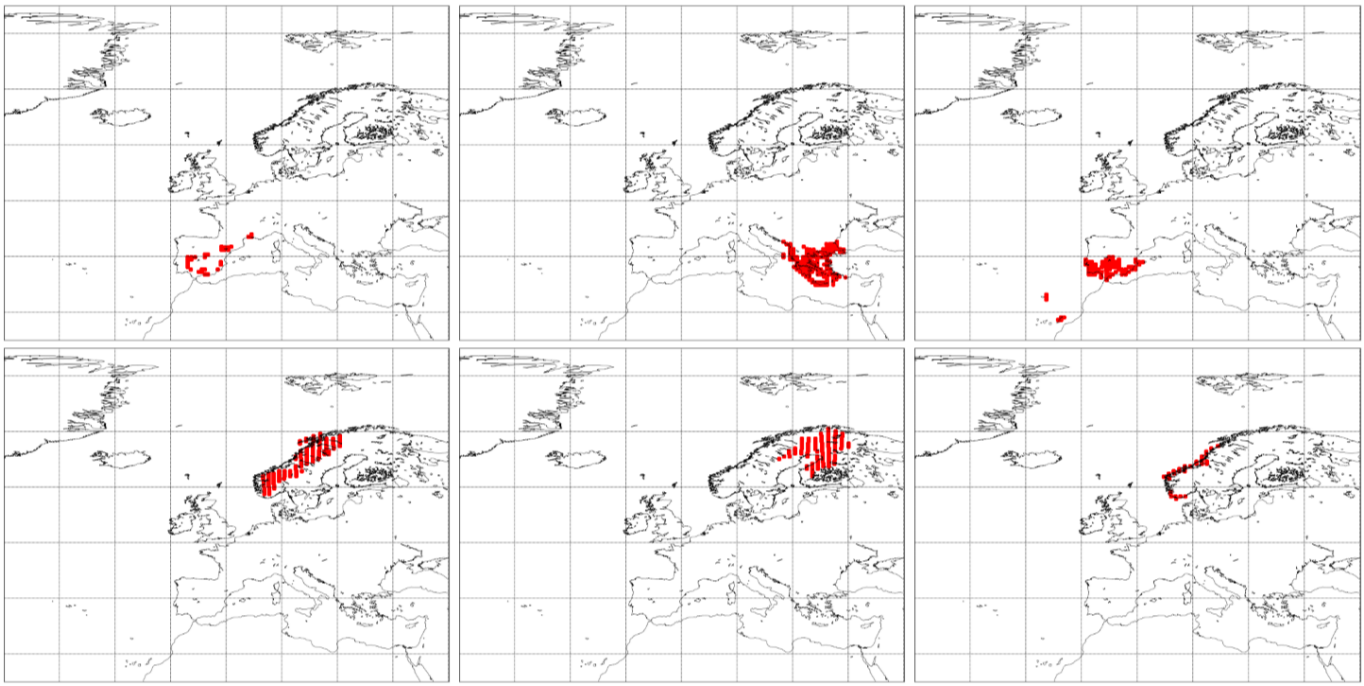

Figure 24: Subgroup list with $radius = 8$. As expected, most subgroups are bigger than the subgroups found in Figure 23. The subgroup in the top left shows for instance that a too big radius results in a subgroup that allows instances that are not near to each other.

5.2.4 Anecdotal evidence

Varying the hyperparameters has an effect on the quality of the found subgroup list and the runtime. However, all the found subgroups still have in common that they adhere to the spatial constraint. In contrast to the clustering methods used for Figure 1 and Figure 3, this thesis only allows spatially connected subgroups. This difference can also be observed in Figure 25.

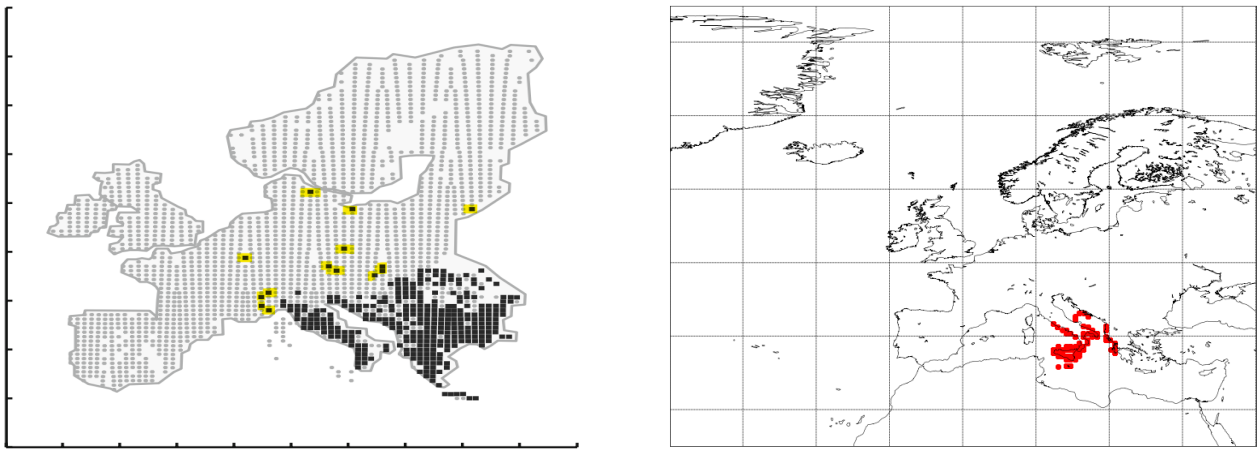

Figure 25: The left figure is one of the clusters found by a clustering algorithm [22]. The right figure is a subgroup found by the spatial subgroup discovery algorithm. The description of the subgroup is: $bioclim1 > 14.81 \wedge prec(Oct) > 61.5 \wedge C = \{(15.86, 38.62), (17.06, 40.41)\}$. It can be observed that the yellow marked instances in the left figure are not spatially connected to the cluster. Such instances are not visible in the right figure due to the imposed spatial constraint.

Unlike the clustering methods discussed in Chapter 1, the approach in this thesis does not only find patterns, but is also able to explain why these patterns exist. For example, the found subgroup with centroids (15.86,38.62) and (17.06,40.41) of Figure 25 can be explained by $bioclim1 > 14.81 \wedge prec(Oct) > 61.5$. Furthermore, because we do not want to make a complete decomposition of the dataset, we can decide what the size of the found subgroups approximately should be. If we want to find smaller subgroups, we give the radius a low value and vice versa. See Figure 26 for two subgroups with different radii.

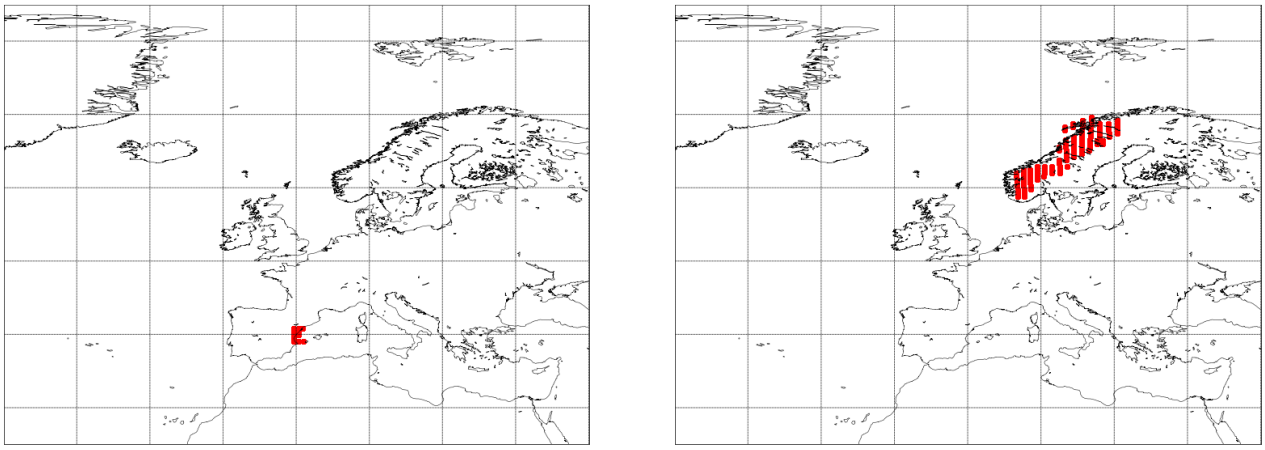

Figure 26: The left figure is the subgroup found with a radius of 1 and the right figure with a radius of 8. If we would like to know interesting patterns of small regions, the radius should be set to a relatively low value. However, if we would like to find interesting patterns of parts of Europe, the radius should be set to a relatively high value.

6 Conclusions & Further Research

We discover interesting subgroups in a spatial dataset with the constraint that the instances within a group should not be more than an acceptable radius away from each other. We therefore introduced the problem of finding the optimal set of subgroups that 1) together explain the most relevant deviations in the data with respect to a given target variable and 2) that adhere to the imposed spatial constraint. Two data mining concepts are combined; subgroup discovery and spatial data mining. In order to ensure the first component of the problem statement, the weighted Kullback-Leibler divergence is used. This calculates the average quality of each set of subgroups. The second component of the problem statement refers to the spatial constraint; all the instances within a subgroup should be in the set of allowed cells, defined by the centroids of the subgroup. All the cells that fall entirely in the surface created by the centroids and the radius plus the distance of a cell, are allowed to the set of allowed cells of a subgroup.

The proposed spatial subgroup discovery algorithm is a combination of beam search and separate and conquer method. After finding a subgroup with beam search, this subgroup is separated and the algorithm is performed again on the rest of the dataset. The algorithm has three hyperparameters: the beam width, the maximum search depth and the radius. Experiments on the *mammals* dataset have shown that the larger the beam width, the better the quality of the subgroup. That is also the case for the radius. However, making the radius too big will result in subgroups in which the instances are not near each other. The quality of the subgroup list, whilst varying the maximum depth, stays the same. In comparison with clustering, the results of this thesis show that due to the use of subgroup discovery in combination with spatial constraints, the algorithm is not only able to find patterns, but is also able to explain the patterns because of the subgroup description. This algorithm requires the dataset to be divided into cells, in which the recorded instances should have a variety of different locations. The algorithm had difficulties with imposing spatial constraints to subgroups in which the instances have the exact same location.

The spatial subgroup discovery algorithm has only been experimented on one dataset. For further research, the algorithm will be tested on multiple datasets containing spatial information. Furthermore, the step of setting the conditions should be refined. Instead of constantly using the median, cut points could be used. Besides that, the algorithm should also be able to allow nominal features.

References

[1] Entropy: How decision trees make decisions. https://towardsdatascience.com/entropy-how-decision-trees-make-decisions-2946b9c18c8, 2019. Accessed: 02-06-2022.

[2] R. Bembenik and G. Protaziuk. Mining spatial association rules. *Intelligent Information Processing and Web Mining*, 42:3–12, 2004.

[3] Anuradha.V.P Dr.Chandra.E. A survey on clustering algorithms for data in spatial database management systems. *International Journal of Computer Applications*, 24, 2011.

[4] Jussi Eronen Hannes Heikinheimo, Mikael Fortelius and Heikki Mannila. Biogeography of european land mammals shows environmentally distinct and spatially coherent clusters. *Journal of Biogeography, 34, 6, 1053-1064*, 2007.

[5] H. Heikinheimo, M. Fortelius, J. Eronen, and H. Mannila. Biogeography of european land mammals shows environmentally distinct and spatially coherent clusters. *Journal of Biogeography*, 34(6):1053–1064, 2007.

[6] Carmona C.J. González P. et al. Herrera, F. An overview on subgroup discovery: foundations and applications. *Knowl Inf Syst*, 29:495–525, 2011.

[7] Francisco Herrera, Cristóbal J. Carmona, Pedro González, and María José Del Jesus. An overview on subgroup discovery: Foundations and applications. *Knowledge and Information Systems*, 29:495–525, 12 2011.

[8] S.E. Cameron J.L. Parra P.G. Jones Hijmans, R.J. and A. Jarvis. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology 25: 1965-1978*, 2005.

[9] Erica Kolatch. Clustering algorithms for spatial databases: A survey. 04 2001.

[10] Sotiris B. Kotsiantis, Ioannis D. Zaharakis, and Panayiotis E. Pintelas. Machine learning: a review of classification and combining techniques. *Artif. Intell. Rev.*, 26:159–190, 2006.

[11] Valliappa Lakshmanan and Travis Smith. Data mining storm attributes from spatial grids. *Journal of Atmospheric and Oceanic Technology - J ATMOS OCEAN TECHNOL*, 26, 11 2009.

[12] lonalona 11922 bronze badges, Carlos MouganCarlos Mougan 5, and Bruno LubascherBruno Lubascher 3. Differences between feature weighting and feature selection. https://datascience.stackexchange.com/questions/65973/differences-between-feature-weighting-and-feature-selection, Jan 2020.

[13] Atzmueller M. Subgroup discovery. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 5(1)*, pages 35–49, 2015.

[14] Marvin Meeng and A. Knobbe. For real: a thorough look at numeric attributes in subgroup discovery. *Data Mining and Knowledge Discovery*, 35:158–212, 2021.

[15] ABDULLAHI SIDOW OSMAN. Data mining techniques: Review. *International Journal of Data Science Research*, 2, 2019.

[16] M. Dalla Mura P. Ghamisi and J. Benediktsson. A survey on spectralspatial classification techniques based on attribute profiles. *IEEE Transactions on Geoscience and Remote Sensing*, 53, 2015.

[17] Ioanna Papagianni. Spatial subgroup discovery. Thesis explanation.

[18] Hugo Manuel Proença, Thomas Bäck, and Matthijs van Leeuwen. Robust subgroup discovery. *arXiv preprint arXiv:2103.13686*, 2021.

[19] Foster Provost and Tom Fawcett. *Data Science for Business*. O'Reilly, Beijing, 2013.

[20] Matthijs van Leeuwen. Maximal exceptions with minimal descriptions. *Data Min. Knowl. Discov.*, 21:259–276, 2010.

[21] Matthijs van Leeuwen and Arno Knobbe. Non-redundant subgroup discovery in large and complex data. In Dimitrios Gunopulos, Thomas Hofmann, Donato Malerba, and Michalis Vazirgiannis, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 459–474, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[22] Matthijs van Leeuwen, Jilles Vreeken, and Arno Siebes. Identifying the components. In Wray Buntine, Marko Grobelnik, Dunja Mladenić, and John Shawe-Taylor, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 32–32, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[23] Knobbe A Van Leeuwen M. Non-redundant subgroup discovery in large and complex data. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 459–474, 2011.

[24] Liu B Wang K, Xu C. Clustering transactions using large items. *Proceedings of the CIKM'99*, pages 483–490, 1999.