Universiteit Leiden
The Netherlands

Opleiding Informatica

Comparing ways of finding patterns in flight delays

Ruben Klijn

Afstudeerrichting: Informatica & Economie

Supervisors:
Matthijs van Leeuwen & Hugo Proença

BACHELOR THESIS

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

08/07/2018

Abstract

The number of people traveling by airplane has been increasing over the last couple of years. Combine this with the effects of climate change, which increases unpredictable and more extreme weather, and it is easy to see that it is a difficult task for airlines to make sure their flights arrive according to schedule. Studies have been done on finding patterns in flight data and weather data, for example on finding causes of flight delays and comparing quality measures to use with subgroup discovery. This work focuses on finding the best way for finding patterns in flight data and weather data that help to tell something about flight delays.

The dataset used in this research consists of flight data with corresponding weather data of domestic flights from the United States in the year 2016. Only data of United Airlines was used from the airports of Denver, Tampa and San Diego. The Diverse Subgroup Set Discovery (DSSD) algorithm was used for doing subgroup discovery. With this algorithm, experiments were done in which two quality measures, two equal frequency discretization techniques, three different bin sizes and two search strategies were compared, resulting in 72 conducted experiments.

In conclusion, experiments showed that there are two ways particularly interesting for finding subgroups with high delays. The first way results in smaller subgroups with particularly high delays, which can, for example, be used for doing outlier detection. The second way results in bigger subgroups with lower (but still above average) delays. This way can, for example, be used for finding/capturing the effects of changes in strategy and policy made by the management of air carriers.

Contents

Chapter 1

Introduction

Over the last couple of years, popularity of traveling via airplanes has increased. According to the Bureau of Transportation Statistics [1], the number of airline passengers increased from 724 million in 2017 to 749 million in 2018 in the United States, which is an increase of 3.5% in just one year. Increases like this are common and tend to happen every year, which makes the task of ensuring that flights arrive according to schedule difficult for air carriers.

According to Mirza [2], climate change is currently happening and increased occurrences of extreme weather conditions will have an effect on every day life. Combine this fact with the yearly increasing number of airline passengers, and it becomes clear that these phenomena can impact flight schedules in a negative way. It is, for example, possible that unpredictable weather events and crowded airplanes can contribute to increased aircraft delays.

Several datamining techniques can be used to analyze data in order to combat aircraft delays. Sternberg et al. [3] tried to find patterns in flight delays by using data indexing techniques combined with association rules, while Post [4] did research on finding patterns in flight delays with subgroup discovery in order to find possible causes of arrival delays.

In his work, Post focused on experimenting with multiple quality measures while making use of different search strategies. However, the effects of using different search strategies or different discretization techniques were not studied in detail in his work. Therefore, this work will extend the work of Post by specifically focusing on comparing the effects of different search strategies and different discretization techniques with subgroup discovery. This will result in experimenting with multiple ways for discovering patterns in flight data, which leads to the following research question:

How can subgroup discovery be used in the best way to discover interesting subgroup sets, with flight delays as target, on a large dataset of flight data and weather data?

This research question will be answered by doing experiments with two different search strategies (standard beam search and diverse beam search) and two types of discretization (a priori binning and on the fly binning).

Both types of discretization are a variation of equal frequency binning, where also the effect of the number of bins will be studied by experimenting with three numbers of bins. Besides, two different quality measures will be compared, namely Weighted Relative Accuracy and the Mean Test, which will determine the interestingness of subgroups.

The term "interesting" appears in the research question and will reappear throughout this bachelor thesis, where interesting refers to a subgroup that tells something about the target, which is the arrival delay. Generally speaking, subgroups that show big deviations in arrival delays are considered to be interesting, and subgroups that show smaller deviations in arrival delays are considered to be less interesting. However, the exact definition for interestingness is given by the quality functions.

1.1 Thesis overview

Firstly, the related work will be discussed, followed by a chapter about the dataset. After that, methods and techniques used in this work will be explained, which is needed in order to have a better understanding about the next chapter, in which the experiments and results are shown. The research is then concluded with an answer to the research question, discussion, and possible future work.

Chapter 2

Related Work

Research on flight delays is an interesting topic from a business perspective, as these delays can cost businesses a lot of money. According to statistics from Airlines for America [5], the costs of flight delays in the year 2017 are estimated to be $26.6 billion in the United States alone. Therefore, research on this topic that is useful for this particular study is already available. This chapter will start with related work on flight delays, after which related work on subgroup discovery will be discussed.

2.1 Flight delays

Sternberg et al. [3] have done research on flight delays on Brazilian airlines. This research focused on finding patterns in flight delays by using association rules and indexing techniques. These patterns were then used to answer questions on causes of flight delays, relations between departure and arrival delays and relations between Brazilian airports and delays. Not only do answers to these questions contribute to understanding flight delays in this research, but the techniques used to build the dataset (like concept hierarchy, binning and temporal aggregation) are also applied in this research on a similar dataset of flight data.

Besides analysis of Brazilian flight data, analysis on domestic United Airlines flights in the United States was done by Post [4]. The goal of this work was to find causes of flight delays in 2016 by experimenting with different search strategies and quality measures. Post used the same dataset that will be used in this research, which means that his work contributes to a better understanding of the dataset, binning techniques and experiments done in this research. However, Post did not focus on the effects of search strategies and discretization techniques on results, which is what this work is trying to achieve.

Research on other aspects of delays has also been done on similar datasets. For example, Deshpande et al. [6] have worked on analyzing the impact flight schedules on flight delays. This work shows what the effects are of man-made decisions on the costs and development of flight delays, which can also be used to reflect on the policies and strategies used by the air carrier used in this work, which is United Airlines.

2.2 Subgroup discovery

Finding patterns in data can be done in multiple ways, with subgroup discovery being one of them. Atzmueller [7] provides an overview of subgroup discovery techniques, terms, and real-world applications that help in understanding subgroup discovery. It is a paper that talks about subgroup discovery with a focus on explanation and understanding of subgroup discovery and related topics such as quality functions, algorithms and subgroup set selection. Knowledge found in this paper is applicable to this research, because subgroup discovery will be used to analyze flight delays.

Subgroup discovery is implemented in a variety of ways, making use of different methods and tools. Helal [8] analyzes different subgroup discovery methods extensively in order to improve understanding these methods. The analyzes are done with multiple datasets, each with different properties, to see how the methods react to these kinds of inputs. The DSSD algorithm, which is used in this research, is also analyzed by Helal. This provides additional knowledge on how DSSD performs on different datasets as compared to other subgroup discovery algorithms.

The Diverse Subgroup Set Discovery (DSSD) algorithm [9] is used in this research and was proposed by Van Leeuwen and Knobbe. Their work explains the algorithm in detail, with additional results on sample datasets. DSSD is able to give diverse subgroups, has different quality functions implemented and can be used for complex tasks as well, which means it is able to perform the tasks done in this work.

The target in this research is arrival delays, which is a numeric attribute. This means that not all quality functions can be used for subgroup discovery. Lemmerich et al. [10] give additional information on evaluation of results, quality functions and subgroup discovery for numeric targets. So challenges and methods for working with numeric targets in subgroup discovery are addressed, which is useful for this work as well, as the target (arrival delay) is numeric.

Subgroup discovery is also used for finding patterns in other organizations. Lavrac et al. [11] tried to compare different subgroup discovery techniques in order to use them as supportive tools for making decisions. Not only did they apply subgroup discovery to two marketing case studies, they also applied subgroup discovery to a medical case study as well. This resulted in a list of possible subgroup discovery risks/problems and solutions, which is useful for this work, as subgroup discovery is used in this work as well.

Chapter 3

Data

According to the "garbage in, garbage out" principle [12], if the input data is of bad quality, the results/output of research will be of poor quality as well in most cases. The data needs to be from reliable sources, after which only the useful data is kept for further research. This chapter will firstly explain the source of the dataset, after which a simple descriptive analysis of single attributes will be discussed. Then, this chapter will conclude with a section on preprocessing and an overview of the resulting datasets that will be used in the experiments.

3.1 Source

The data used in this research is composed of two parts: one part consists of flight data, while the second part consists of weather data that matches the other part. The complete dataset contains data from all months in 2016. Reason for this being that this was the last year in which the data was completely logged before changing to other formats. It is the same dataset used in the research that was previously done by Post [4].

The dataset contains more information than needed for this research, as the flight data contains data of all carriers. In order to get accurate results, a scope needs to be defined that limits certain parts of the data. Therefore, only flight data of United Airlines was used. This decision was made because United Airlines has kept its policy/strategy relatively constant over the last years. This consistency in management is supported by the consistency in delays over the years and contributes to data that represents daily operations in regular circumstances, which reduces outliers and unexpected behaviour in the dataset.

The flight data was gathered from the United States Bureau of Transportation Statistics [1], while the weather data was gathered from the United States National Oceanic and Atmospheric Administrations [13]. The flight data contains information of all domestic flights in the United States across different carriers. The weather data consists of various weather and climate conditions as observed in the United States. The advantage of these two sources of data is that they provide consistent and relatively complete data, which contributes to accurate results.

The entire dataset contains 914495 flights of United Airlines in the United States and 157 attributes, with an average arrival delay of 5.1 minutes. The way in which this dataset is reduced is given in Chapter 3.3.

3.2 Single attribute statistical analysis

In order to get a better understanding of the dataset, a simple descriptive analysis is done consisting of a correlation matrix and a look at the target attribute, which is arrival delays. Figure 3.1 shows a correlation matrix of all raw flight data (excluding weather data) for United Airlines. At first glance, it becomes clear that most variables are not correlated and probably are linearly independent. The correlations that are present are obvious ones: CRS_ELAPSED_TIME (the time it takes a flight to travel from airport A to airport B) shows a high correlation with the DISTANCE and DISTANCE_GROUP (a representation of the distance of the flight). This is to be expected, as flights that travel longer distances take longer to arrive at the destination. The departure delays (DEP_DELAY) and arrival delays (ARR_DELAY) also are strongly correlated. This implies that flights that depart later than expected, also will arrive later than expected. For this reason, the DEP_DELAY attribute will be removed from the dataset, as it can influence the results with information that is not relevant for answering the research question. Another reason for leaving out this attribute is that this research only wants to use data that was available before departure, either through forecasts (for example with weather data) or schedules and information recorded before departure.

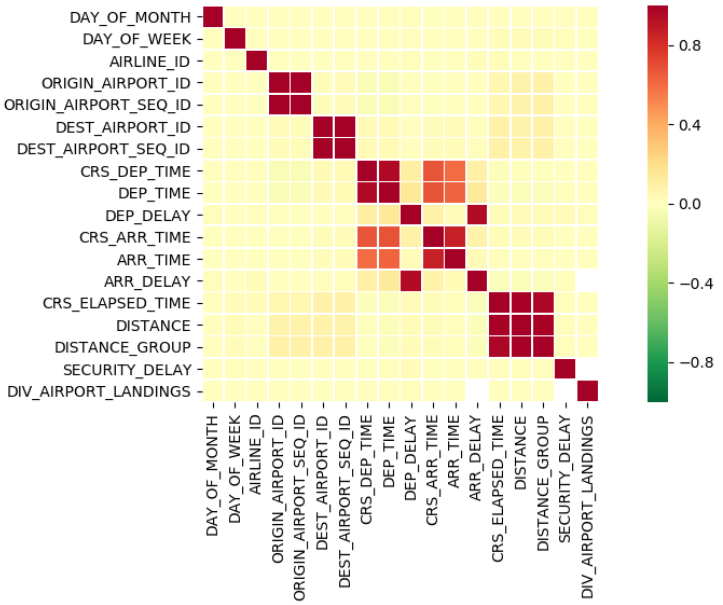


Figure 3.1: A correlation matrix of the variables in the raw domestic flight data of United Airlines in 2016.

When looking at average arrival delays in 2016 (see Figure 3.2), two sudden increases in the average arrival delays happen throughout the year, namely during summer and December. One possible explanation for this can be the weather conditions: summer and December are, respectively, very hot and cold periods of the year, which can affect arrival delays. Another possible explanation could be holidays. During summer and

December, a lot of people go on vacation. So these two periods can be, besides extreme in weather conditions, also the busiest times of the year. All of this can contribute to peaks in average arrival delays. These two possible explanations were suggestions, as it is difficult to find the exact explanations. This research is focused on finding an answer to the research question, not on finding causes for behavior otherwise seen in the data. In some cases, though, suggestions for possible explanations for causes will be given.

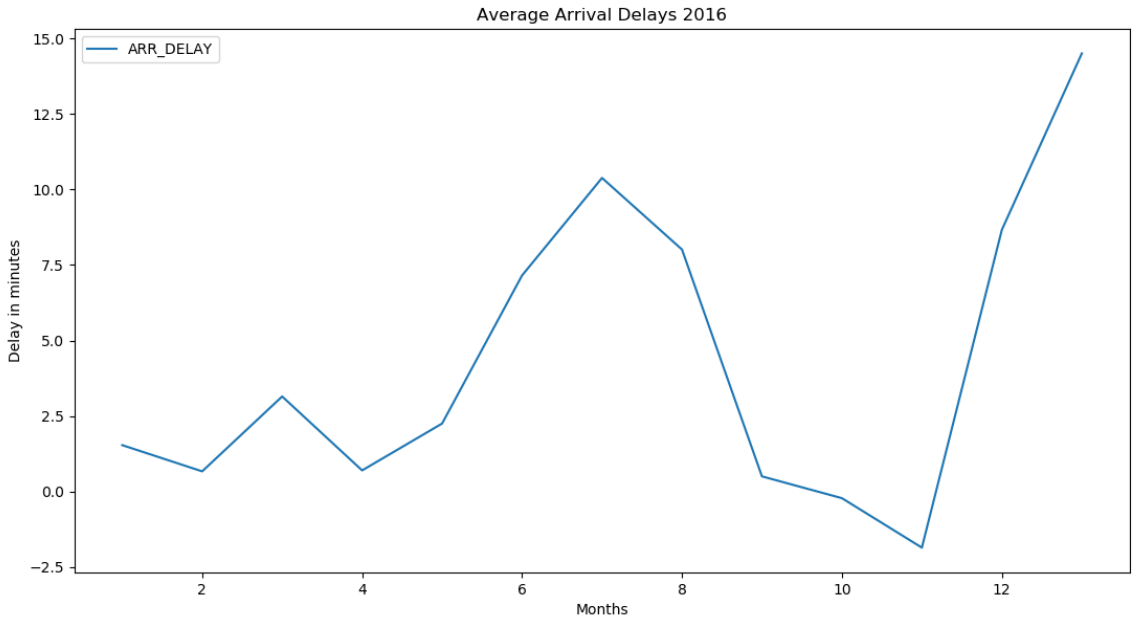


Figure 3.2: A plot of the average arrival delays in the year 2016 regarding domestic flights in the United States from United Airlines.

3.3 Selecting specific airports

A dataset that contains all domestic flights of United Airlines in the year 2016, complete with weather data, is too big given the available resources (time, computing power, etc.). Therefore, a smaller selection of the data must be made for further research. The data will be reduced to data of three specific airports, namely Tampa, Denver and San Diego. Each airport has data on at least 18000 flights. This includes incoming as well as outgoing flights from these airports.

These airports have been selected based on geographical location. The United States have different climates and weather conditions across the country. Therefore, airports that are not near each other will face different weather types and climate conditions (according to the Köppen Climate Classification [14]) throughout the whole year. The airports that are used in the dataset are shown in Table 3.1.

Furthermore, for each of these airports, cancelled and diverted flights will also be removed. In theory, a cancelled flight could have an infinite arrival delay or no arrival delay value at all. This makes it impossible to work with cancelled flights, as the target variable is the arrival delay. The same can be said for diverted flights,

City	State	Flights	Climate	Mean ARR-DELAY
Tampa	Florida	24498	Humid, Subtropical	4.7 minutes
Denver	Colorado	20978	Semi-Arid, Continental	6.6 minutes
San Diego	California	18486	Semi-Arid, Mediterranean	2.9 minutes

Table 3.1: An overview of the selected airports. These airports are spread out across the United States and each airport has a different climate, so each airport endures different weather conditions to base the results upon. This table shows the location of the airport (city and state), the number of domestic flights per airport during 2016, the climate of the airport and the average arrival delays per airport in 2016.

as these flights will never reach the original destinations.

3.4 Removing sparse attributes

In this context, an attribute is said to be sparse when an attribute has many missing values, which can give less accurate results. So attributes that contain too many missing values will be removed. To determine whether an attribute contains enough non-missing values or not, a threshold will be used. This threshold is relative to the size of the airport datasets (e.g. the number of flights). Iff an attribute contains an amount of non-empty values of at least 3% of the total amount of flights, that specific attribute will not be removed from the dataset. The threshold is put at 3% because a higher threshold results in no extreme weather condition attributes, as extreme weather conditions do not appear often. A threshold of lower than 3% would result in columns that contain too few non-empty values. See the third paragraph of the Discussion section in Chapter 6 for more information on the extreme weather attributes.

3.5 Overview of attributes after preprocessing

Table 3.2 gives summary information about the dataset that will be used in the experiments. Table 3.3 shows all the attributes that are used in this research, together with a short description on the attributes. Some attributes in the table appear with a *. These attributes have three different prefixes regarding weather conditions: LATEARRIVAL, ARRIVAL, or DEPART. The DEPART prefix indicates that the attribute tells something about the weather at the departure station at the scheduled time of departure and the ARRIVAL prefix indicates that the attribute tells something about the weather at the arrival station at the time of departure; and the LATEARRIVAL prefix indicates that the attribute tells something about the weather at the arrival station at the scheduled time of arrival. When looking at Table 3.3, for example, Visibility* is present. This means that LATEARRIVAL_Visibility, ARRIVAL_Visibility and DEPART_Visibility are all different attributes in the dataset.

City	State	Flights	Climate	Mean ARR-DELAY
Tampa	Florida	24234	Humid, Subtropical	4.7 minutes
Denver	Colorado	20635	Semi-Arid, Continental	6.6 minutes
San Diego	California	18297	Semi-Arid, Mediterranean	2.9 minutes

Table 3.2: An overview of the selected airports after removing sparse attributes and diverted/cancelled flights. These airports are spread out across the United States and each airport has a different climate, so each airport endures different weather conditions to base the results upon. This table shows the location of the airport (city and state), the number of domestic flights per airport during 2016, the climate of the airport and the average arrival delays per airport in 2016.

Attribute	Description
QUARTER	The quarter of the year.
MONTH	The month of the year.
DAY_OF_MONTH	The day of the month.
DAY_OF_WEEK	The day of the week.
ORIGIN_CITY_NAME	The name of the origin airport of the flight.
DES_CITY_NAME	The name of the destination airport of the flight.
CRS_DEP_TIME	The originally scheduled departure time.
DEP_TIME_BLK	The hour in which the departure time falls.
CRS_ARR_TIME	The originally scheduled arrival time.
ARR_TIME	The actual arrival time.
ARR_DELAY	The arrival delay.
ARR_TIME_BLK	The hour in which the arrival time falls.
DISTANCE_GROUP	Indicator for the distance of a flight.
Visibility*	Whether or not the pilots can see outside or have to rely on systems.
DryBulbCelsius*	Measure for the outside temperature.
RelativeHumidity*	The humidity of the air outside.
WindSpeed*	The speed of the wind.
WindDirection*	The wind direction.
ValueForWindCharacter*	Describes the character/aggressiveness of the wind.
StationPressure*	Air pressure at the weather station.
SeaLevelPressure*	Air pressure at sea.
HourlyPrecip*	The amount of precipitation per hour.
SKYCONDITION1	Sky condition 1.
SKYCONDITION2	Sky condition 2.
SKYCONDITION3	Sky condition 3.
DEPART_FOG	Whether or not there was fog at departure.
ARRIVAL_FOG	Whether or not there was fog at arrival.
SEASON	The season.

Table 3.3: An overview of the remaining attributes that will be used in the experiments. A * means that an attribute has three different variations: a DEPART, ARRIVAL and LATEARRIVAL variation. The DEPART prefix indicates that the attribute tells something about the weather at the departure station at the scheduled time of departure; the ARRIVAL prefix indicates that the attribute tells something about the weather at the arrival station at the time of departure and the LATEARRIVAL prefix indicates that the attribute tells something about the weather at the arrival station at the scheduled time of arrival.

Chapter 4

Methods

This chapter will provide information on the techniques and methods used in this research. Firstly, a brief introduction to the notations used in this chapter will be given. This is followed by an explanation of subgroup discovery, after which four subgroup discovery properties (the target attribute, subgroup description language, quality functions and search strategies) will be explained. After that, the binning techniques relevant for this research are discussed, before ending with a short explanation of the DSSD algorithm.

4.1 Data notation

Let S denote the entire dataset, which consists of a set of attributes A. A contains x description attributes D (with $x \geq 1$) and exactly 1 target attribute T. This means that the dataset S is a collection of tuples over the set of attributes $A = \{D_1, ..., D_x, T\}$.

A subgroup G is then a collection of tuples $G \subseteq S$, with $|G|$ being the size or coverage of the subgroup and $|S|$ being the size of the entire dataset. The average value of the target attribute T over the collection of tuples G is denoted with μ^G, while μ^S denotes the average value of T over the collection of tuples S.

4.2 Subgroup discovery

Subgroup discovery is a technique for finding interesting subgroups for a specific target attribute. A subgroup is described by a rule, which can look like this: YEAR = 2016 $\wedge$ DAY = Monday. This means that all rows/cases in the data where YEAR is equal to 2016 and DAY is equal to Monday are interesting in telling something about the target attribute (for example, cases for which the target attribute has high or low values), where interestingness is determined by a quality function. According to Atzmueller [15], four properties are important for subgroup discovery, namely: the target attribute, the subgroup description language, the quality function, and the search strategy. Each of these four properties will be discussed next.

4.2.1 The target attribute

Target attributes can come in a variety of types, of which ordinal, binary, and numeric targets are all possibilities. Each type of target can give different insights and answer different questions. It may sometimes be necessary to change the type of the target. For example, numeric targets may need to be discretized because of a certain quality function that cannot work with numeric data. So the type of the target attribute is important for the setup of the research. In case of this research, the target attribute is the arrival delay, which is a numeric attribute.

4.2.2 Subgroup description language

As said before, a subgroup is a subset of the data that complies with a specific description and there are multiple ways to build these descriptions. A description consists of conditions. In this research, these conditions can either have the following operands: =, $\neq$, $>$ and $<$. So for example, WEATHER = Sunny and TIME $<$1200 are valid conditions. A description consists of multiple individual conditions. In this research, these conditions can only form a description if the conditions are connected with the $\wedge$ operand, i.e., conjunction. When looking at the previous example, WEATHER = Sunny $\wedge$ TIME $<$1200 could be an example of a description that is allowed, but WEATHER = Sunny $\vee$ TIME $<$1200 is an example of a description that is not allowed within this research.

4.2.3 Quality function

Subgroups are prioritized based on "interestingness", which can be made explicit by using quality functions. A quality function generates a score based on the characteristics of the subgroup and/or the entire dataset. This score then indicates the interestingness of the subgroups, with a higher score implying more interestingness. For this research, two quality functions are relevant, namely the Weighted Relative Accuracy for numeric attributes and the Mean Test. These two quality functions are chosen because they work with numeric attributes, so no discretization of the target attribute is needed.

Weighted Relative Accuracy

The Weighted Relative Accuracy (WRAcc) for numeric attributes is defined as follows:

$$WRAcc(G) = \frac{|G|}{|S|}(\mu^G - \mu^S)$$

Where G stands for the subgroup and S stands for the entire dataset. So it first divides the size of the subgroup by the size of the entire dataset, and then multiplies this with the difference between the average value of the target attribute within the subgroup and the average value of the target attribute within the whole dataset. This means that this quality function does not only care for subgroups with a high deviation in the target attribute, but it takes the size of the subgroup relative to the size of the entire dataset into consideration as well.

So the biggest subgroups with the highest deviations in the target attributes get the highest scores according to this quality function.

Mean Test

The Mean Test (MT) is defined as follows:

$$MT(G) = \sqrt{|G|}(\mu^G - \mu^S)$$

Where, again, G stands for the subgroup and S stands for the entire dataset. It looks the same as the WRAcc quality function, with one big difference: instead of dividing the size of the subgroup by the size of the entire dataset, the MT uses the square root of the size of the subgroup. This means that the MT only cares for subgroups with the highest deviations of the target, not taking the size of the subgroups into account. So only subgroups with the highest deviations get the highest scores, which means it is possible for relatively small subgroups to make it to the top of the ranking based on just their score.

4.2.4 Search strategy

An exhaustive method of searching the entire search space gives optimal results. Such a search strategy, however, is often not feasible because of the size of the search space. Therefore, to reduce the search space, heuristic strategies can be used. For this particular research, beam search was used, which selects only the most promising attribute subsets, but truncates after a certain number of subsets is reached. This maximal number is called the beam width. The process of selecting and adding attributes is done level-by-level, where in each level, the search space is reduced by adding one promising attribute to the subset of attributes. In other words, subgroups get refined in each level, where the quality function ranks the (refined) candidate subgroup sets according to which subgroup sets are most promising. In this way, the search space gets smaller with each level, making beam search more efficient for use with larger datasets.

4.3 Equal frequency binning

Equal frequency binning (or equal frequency discretization) refers to categorizing numeric attributes. Numeric attributes can potentially have high cardinalities, making the search space large. This large search space can affect run times of algorithms in a negative way by increasing it. Therefore, discretization is necessary in order to improve run times and efficiency of algorithms.

There are two types of binning (discretization) strategies: equal interval binning and equal frequency binning. With equal interval binning, the domain of values is divided into intervals of equal width. However, some intervals may contain more values than others. Equal frequency binning divides the domain into intervals that contain the same number of values. The width of the intervals can vary with equal frequency binning. For this

research, only equal frequency binning will be used. The number of bins will vary based on the distributions of the attributes.

4.3.1 A priori binning

One way of applying a binning technique is to do the binning before the data is used as input for an algorithm. This is called a priori binning. For this research, the number of bins will be determined by the distributions of the attributes and the maximum number of splits, where a maximum of 3 splits uses the fewest bins, a maximum of 7 splits uses the most bins and a maximum of 5 splits uses in between 3 and 7 splits. For each numeric attribute, the distribution was manually inspected in order to determine the appropriate number of bins, which will now be discussed.

Normal and skewed distributions

If the distribution of an attribute is (close to) normal or skewed, three different bins will be used in the first setup. This is done because a bin is created in the area where the data is most dense, namely in the middle of the distribution. The number of bins of attributes will increase with two bins, which makes sure that the bin in the middle (where the data is dense) is kept. The binning setups are visualized in Figure 4.1.

(A) (B) (C)

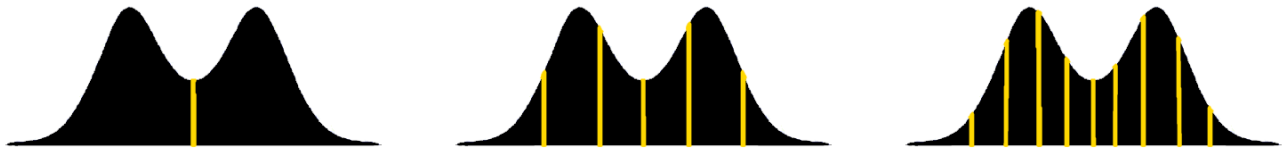
Figure 4.1: Schematic view of bins per setup for normal and skewed distributions. (A) corresponds to maximal 3 splits, (B) corresponds to maximal 5 splits and (C) corresponds to maximal 7 splits.

Waved distributions

If the distribution of an attribute appears to have at least two waves (at least two peaks), the following formula will be used to determine the fewest number of splits:

$$number\ of\ splits = number\ of\ peaks - 1$$

After that, each wave gets treated like a normal distribution. This means that per wave, a total of 3 bins will be used. An example with a bimodal distribution is given in Figure 4.2. This distribution has two peaks. The formula states that $2 - 1 = 1$, so the smallest number of bins will be 2. Since there are two waves, the second biggest number of bins will be 6 bins. After that, 4 bins will be added in the largest number of bins, which results in 10 bins.



(A) (B) (C)

Figure 4.2: Schematic view of bins per setup for bimodal distributions. (A) corresponds to the smallest number of bins, (B) corresponds to the second biggest number of bins and (C) corresponds to the largest number of bins.

Constant distributions

If the distribution is (close to) being constant, the smallest number of bins will be equal to 2. After that, the number of bins is increased by 1. So the second biggest number of bins will be 3 bins, while the biggest number of bins will be 4. Figure 4.3 shows this graphically. This is done because the density of data is approximately the same everywhere along the distribution. So there is no part where the data is most dense to isolate, which is why two initial bins are chosen as the smallest number of bins.

(A) (B) (C)

Figure 4.3: Schematic view of bins per setup for constant distributions. (A) corresponds to the smallest number of bins, (B) corresponds to the second biggest number of bins and (C) corresponds to the largest number of bins.

4.3.2 On the fly binning

As said earlier, a priori binning is a way of adding bins before the data is used as input by the algorithm. DSSD offers "on the fly" (OTF) binning. With OTF binning, unbinned data is the input of DSSD. DSSD will then apply its own binning technique when the algorithm runs, in which the values of a numeric attribute that occur within a subgroup are split into a number of six bins. The number of bins must be determined before the algorithm starts and cannot be changed during or after finishing the algorithm. After that, the $\{<, >\}$-conditions for each given cut point will be calculated. This process is done after each refinement step in the algorithm, where refinement is the process of adding a condition to subgroups that are generated, reducing the search space in a level-wise fashion.

In order to compare a priori binning with OTF binning, the number of bins needs to be comparable. Therefore, DSSD will be run with three different numbers of splits:

- a maximum of 3 splits to compare with the fewest number of bins with a priori binning
- a maximum of 5 splits to compare with the second biggest number of bins with a priori binning
- a maximum of 7 splits to compare with the biggest number of bins with a priori binning

An overview of all numeric attributes with their respective bins is given Appendix A. These numbers of bins are obtained based on the distributions of the attributes, as explained in this chapter.

4.4 Diverse Subgroup Set Discovery

Diverse Subgroup Set Discovery [9] (DSSD) is the algorithm that is going to be used to discover subgroups. It offers multiple quality functions (WRAcc, Mean Test, ChiSquared Test and (Weighted) Kullback Leibler Measure), of which the WRAcc and the Mean Test are going to be used in this research. DSSD also has four beam strategies, of which three are diverse beam search strategies and one standard beam search strategy. In theory, the diverse strategies should find more diverse subgroups than the standard strategy. For this research, only the standard beam strategy (named "quality" in DSSD) and the cover based diverse strategy are relevant.

4.4.1 Subgroup selection

DSSD offers four different strategies for making selections of subgroups, called quality, description, cover and compression in DSSD. Quality refers to standard top-k beam search, while the other strategies are DSSD strategies. Only the cover-based selection strategy was used in this research, which is computationally more demanding than description-based beam search. With the description-based selection strategy, subgroups are selected based on their descriptions, while the cover-based search strategy also takes the coverages of the subgroups into account.

After subgroups are found, dominance pruning is applied to improve the quality of subgroups. This can potentially improve the quality of the subgroup if, for example, the quality of the subgroup would be higher if a specific condition were to be removed. This pruning is done based on the dominance of the subgroup. Let us have two subgroups, G_y and G_z. G_y dominates G_z iff the quality of G_y, as generated by the quality function, is equal to or greater than the quality of G_z and the conditions that appear in the description of G_y are a strict subset of the conditions that appear in the description of G_z.

Chapter 5

Experiments

This chapter will show the results obtained from the experiments, which will be used to support an answer to the following research question:

How can subgroup discovery be used in the best way to discover interesting subgroup sets, with flight delays as target, on a large dataset of flight data and weather data?

Firstly, an overview of the 72 conducted experiments will be given. This is followed by an in-depth analysis of the results, in which the following categories of results will be discussed: diversity, quality measures, and obtained flight delays. Each category will answer a different question:

- Diversity
  - Which parameters give the largest diversity?
- Quality Measures
  - Do quality measures give higher rankings to certain experiments?
  - What parameters give the largest coverage?
- Obtained Flight Delays
  - What parameters result in the biggest delays?

This means that the results of the experiments will be judged based on diversity, quality measures and flight delays. The diversity gives information on how diverse the obtained subgroups are, which depends on the number of the same conditions that appear in the subgroup descriptions. For example, if the same conditions appear many times over across the different subgroup descriptions, results are not diverse. Ideally, results need to be as diverse as possible, which is why diversity is one of the three evaluation criteria. The second evaluation criteria is the quality measure, of which there are two in this research. Each quality measure works differently and gives different results. Therefore, it is interesting to see what these quality measures do in terms of the subgroup sizes they generate and whether some experiments get higher scores than other experiments. The third evaluation criterion is flight delays. Flight delays are important in this research, as the goal is to find subgroups that give interesting results regarding flight delays.

5.1 Setup of experiments

The experiments vary in the number of bins, the type of discretization, the search strategy, quality measures and the datasets used (since every airport utilizes a different part of the data). An overview is given in Table 5.1, of which each of the columns will be discussed next.

Number of Splits	Type of Discretization	Beam Search Strategy	Quality Measure
3	AP	Diverse	WRAcc
5	OTF	Standard	Mean Test
7			

Table 5.1: An overview of the experiments. Each experiment has four different parameters, namely the maximum number of splits for binning (exact number of bins can differ), the type of binning (a priori (AP) or on the fly (OTF)), the type of beam search (standard or diverse) and the quality measure (WRAcc or Mean Test). These experiments need to be repeated three times, once for each airport, giving a total of 72 experiments.

5.1.1 Number of splits

Three different variations of the data will be used, with each variation differing in the number of bins used when binning numeric attributes. The first variation of the data will consist of numeric attributes binned with at most three splits, while the second data variation consists of numeric attributes binned with at most five splits. The third and last data variation consists of at most seven splits in the binning of numeric attributes. See Chapter 4 for more information on binning.

5.1.2 Type of discretization

Two ways of discretization are compared, namely a priori binning (denoted with AP in tables) and on the fly binning (denoted with OTF in tables). With a priori binning, Python will be used to apply the number of bins before the DSSD algorithm is used. DSSD offers an "on the fly" binning option as well, which will be used as the second way of discretization of data. When using on the fly binning, DSSD will use its built-in discretization method to bin the numeric attributes (see Chapter 4 for more information on DSSD on the fly binning).

5.1.3 Beam search strategy

DSSD offers standard beam search (denoted with BS) as well as diverse beam search (denoted with DBS). In order to discover what effects these two different search strategies have on the results, the two beam search strategies will be compared to each other.

5.1.4 Quality measures

Post [4] compared different quality measures to each other in detail on the same dataset. Therefore, only two quality measures will be compared in this research, namely Weighted Relative Accuracy (WRAcc) for numeric data and the Mean Test.

5.1.5 DSSD parameter settings

Although the DSSD algorithm has many parameters to tune, the parameters discussed in this part were the only parameters that were tuned for this research. All other parameters were left at the default settings. With regards to the top-k setting, which relates to the number of results to keep during the initial search phase of the algorithm, the value was put at 1000. A top-k setting that was higher than 1000 did not result in more interesting subgroups (where the exact amount of interestingness is determined by the quality measures), which is why this number was chosen. The same can be said for the beam width, which is put at 100. Lastly, the maximum number of conditions that could appear in a rule is 3. This is done because a number greater than 3 resulted in extra attributes that did not contribute to better results, as these attributes repeatedly added the same conditions to multiple subgroups.

5.2 Diversity of the results

5.2.1 Unique Conditions as a metric for diversity

For both the top-10 and top-100 subgroups generated with each experiment, the structure of the rules was analyzed to describe diversity. The number of unique conditions is used as a metric for measuring diversity. The terms “unique conditions” and ”diverse“ are explained with the following example shown in Table 5.2. A description (the rule of the subgroup) consists of at most three conditions. The top-10 subgroups shown in Table 5.2 contain fourteen unique conditions in total, which is the result of counting the occurrence of each individual condition before removing duplicate conditions. A result is said to be diverse when the amount of unique conditions is relatively high. So the example from Table 5.2 is not diverse, as many conditions across multiple subgroups are the same.

5.2.2 Results

The experiments showed that, regardless of the number of splits, type of discretization and quality measure, diverse beam search resulted in more unique conditions (so more diverse subgroups) as compared to standard beam search. This was true for both the top-10 (see Table 5.3 and Table 5.4, where DBS scores higher than BS in almost all experiments) and top-100 results per result file. This shows that diverse beam search indeed results in more diverse subgroups compared to standard beam search.

Sg	Quality	#Conditions	Description
1	253	3	CRS_ARR_TIME=Late ∧ ARR_TIME≠Late ∧ ARR_TIME_BLK≠1700-1759
2	42	3	CRS_DEP_TIME=Late ∧ ARR_TIME≠Late ∧ DEST_CITY_NAME=Dallas-TX
3	113	3	CRS_ARR_TIME=Late ∧ ARR_TIME=Early ∧ ARR_TIME_BLK≠2300-2359
4	252	3	CRS_ARR_TIME≠Early ∧ ARR_TIME=Early ∧ ARR_TIME_BLK≠1200-1259
5	135	3	CRS_ARR_TIME=Late ∧ ARR_TIME=Early ∧ CRS_DEP_TIME≠Early
6	133	3	CRS_ARR_TIME=Late ∧ ARR_TIME=Early ∧ DISTANCE_GROUP≠F
7	282	3	CRS_ARR_TIME=Late ∧ ARR_TIME≠Late ∧ DEP_TIME_BLK≠1400-1459
8	111	3	CRS_ARR_TIME=Late ∧ ARR_TIME=Early ∧ DEP_TIME_BLK≠2200-2259
9	266	3	CRS_ARR_TIME≠Early ∧ ARR_TIME=Early ∧ DEP_TIME_BLK≠0800-0859
10	281	3	CRS_ARR_TIME=Late ∧ ARR_TIME≠Late ∧ CRS_DEP_TIME≠Middle

Table 5.2: An example of what the top-10 results of a single results file look like. The columns show respectively the index of the subgroup, the score of the quality measure for the subgroup, the size of the subgroup, the number of conditions of the rule that describes the subgroup and the description of the subgroup itself.

Another finding was that in general, on the fly binning resulted in more diverse subgroups compared to a priori binning. This was again true for both the top-10 and top-100 results. This is also shown in Table 5.3 and Table 5.4, as the OTF columns score better than the AP columns. This can possibly be explained by the fact that DSSD on the fly binning can apply binning in a more efficient and precise way than a priori binning.

The top-100 results with the WRAcc quality measure showed that the number of bins seems to influence the number of unique conditions positively: as the number of bins increases, the number of unique conditions also increases. This is not the case when looking at the top-10 results, where the number of bins does not seem to affect the number of unique conditions.

The next observation is about the Mean Test quality measure in both the top-10 (see Table 5.4) and top-100 results. When looking at the number of unique conditions found with on the fly binning, the number of unique conditions increased as the number of bins increased. So the number of bins influences the number of unique conditions when using the Mean Test quality measure with on the fly binning.

Regarding the unique conditions, the quality measures showed something interesting. With the top-10 results, both quality measures showed roughly the same number of unique conditions. When looking at the top-100 results, however, the Mean Test resulted in more unique conditions compared to the WRAcc quality measure. So when looking at all results, the Mean Test outperformed the WRAcc quality measure when it comes to unique conditions.

	Max. Splits: 3				**Max. Splits: 5**				**Max. Splits: 7**			
	AP		**OTF**		**AP**		**OTF**		**AP**		**OTF**	
	BS	**DBS**	**BS**	**DBS**	**BS**	**DBS**	**BS**	**DBS**	**BS**	**DBS**	**BS**	**DBS**
Denver	**11**	8	15	**22**	4	**14**	19	**20**	4	**12**	18	**19**
Tampa	**10**	**10**	18	**23**	8	**12**	19	**23**	4	**12**	15	**25**
San Diego	12	**17**	18	**24**	**11**	**11**	17	**23**	10	**14**	12	**22**
Average	11	**11.7**	17	**23**	7.7	**12.3**	18.3	**22**	6	**12.7**	15	**22**

Table 5.3: The amount of unique conditions in the top-10 results with WRAcc. The columns represent the maximum number of splits in binning, standard beam search (BS) and diverse beam search (DBS), a priori binning (AP) and on the fly binning (OTF).

	Max. Splits: 3				Max. Splits: 5				Max. Splits: 7			
	AP		OTF		AP		OTF		AP		OTF	
	BS	DBS	BS	DBS	BS	DBS	BS	DBS	BS	DBS	BS	DBS
Denver	**15**	12	18	**20**	**18**	16	21	**22**	15	**20**	**20**	19
Tampa	13	**14**	**17**	**17**	**12**	**12**	**18**	**18**	12	**14**	18	**20**
San Diego	13	**14**	15	**18**	15	**19**	15	**21**	11	**16**	19	**21**
Average	**13.7**	13.3	16.7	**18.3**	15	**15.7**	18	**20.3**	12.7	**16.7**	19	**20**

Table 5.4: The amount of unique conditions in the top-10 results with Mean Test. The columns represent the maximum number of splits in binning, standard beam search (BS) and diverse beam search (DBS), a priori binning (AP) and on the fly binning (OTF).

5.2.3 A closer look at a diversity outlier

Table 5.3 shows that a result for the Denver airport is different from the rest: the experiment with max. splits 3, a priori binning and standard beam search resulted 11 unique conditions, whereas diverse beam search resulted in 8 unique conditions. This is the only case in the table where standard beam search scores higher than diverse beam search with WRAcc. Both of these cases, with standard beam search and diverse beam search, are shown in Table 5.5 and Table 5.6 respectively.

These tables do not appear to be special. As stated before, the rules in Table 5.5 are less diverse than in Table 5.6 due to the different beam search strategies. Also, in both tables, the average delays are higher than the average delay for the whole airport of Denver, which is 6.6 minutes, and the amount of delays tend to decrease as the ranking (based on the scores) of the subgroups decrease. These remarks make sense, as the WRAcc quality measure (as well as the Mean Test) give the highest scores to subgroups that tend to deviate the strongest from the total average delay of Denver airport. Unfortunately, these tables do not appear to show anomalies, which means that based on these tables, it is not possible to explain why BS outperformed DBS with Denver, max. splits 3 and a priori binning.

5.2.4 Which parameters give the biggest diversity?

Based on the results, the experiments with generally the biggest diversity had the following parameters:

- Quality Measure: Mean Test
- Type of Binning: OTF
- Beam Search Strategy: Diverse Beam Search
- Maximum Number of Splits: 7

	size	AvgDelay	Condition 1	Condition 2	Condition 3
0	6561	19.1	CRS_ARR_TIME≠Early	ARR_TIME≠Middle	MONTH≠Nov
1	6609	18.9	CRS_ARR_TIME≠Early	ARR_TIME≠Middle	MONTH≠Feb
2	7038	18.2	CRS_ARR_TIME≠Early	ARR_TIME≠Middle	ARR_TIME_BLK≠1200-1259
3	7053	18.1	CRS_ARR_TIME≠Early	ARR_TIME≠Middle	DEP_TIME_BLK≠0800-0859
4	6934	18.2	CRS_ARR_TIME≠Early	ARR_TIME≠Middle	DEP_TIME_BLK≠2100-2159
5	7091	17.9	CRS_ARR_TIME≠Early	ARR_TIME≠Middle	ARRIVAL_WindDirection≠Varied
6	7098	17.9	CRS_ARR_TIME≠Early	ARR_TIME≠Middle	-
7	7311	16.9	CRS_DEP_TIME≠Early	ARR_TIME≠Middle	ARR_TIME_BLK≠0001-0559
8	7745	16.4	CRS_DEP_TIME≠Early	ARR_TIME≠Middle	DEP_TIME_BLK≠2300-2359
9	7512	16.5	CRS_DEP_TIME≠Early	ARR_TIME≠Middle	MONTH≠Nov

Table 5.5: Top-10 results of experiment Denver, max. splits 3, with WRAcc, a priori binning and standard beam search with average delays for the subgroups. This table does not show any unexpected results and does not appear to be diverse.

	size	AvgDelay	Condition 1	Condition 2	Condition 3
0	6561	19.1	CRS_ARR_TIME≠Early	ARR_TIME≠Middle	MONTH≠Nov
1	7525	16.4	CRS_DEP_TIME≠Early	ARR_TIME≠Middle	MONTH≠Feb
2	10868	11.1	DEPART_SeaLevelPressure≠High	QUARTER≠Fourth	MONTH≠Sep
3	9474	12.2	CRS_ARR_TIME≠Early	SEASON≠WINTER	MONTH≠Nov
4	8818	12.2	LATEARRIVAL_RelativeHumidity≠Low	DEST_CITY_NAME≠Charlotte-NC	DEPART_SeaLevelPressure≠High
5	10606	11.3	ARR_TIME≠Middle	ARR_TIME_BLK≠1100-1159	DEP_TIME_BLK≠0700-0759
6	11599	10.9	CRS_DEP_TIME≠Early	MONTH≠Apr	MONTH≠Feb
7	7776	13.3	ARR_TIME≠Middle	ARR_TIME_BLK≠1100-1159	DEPART_DryBulbCelsius≠Low
8	9129	12.0	CRS_ARR_TIME≠Early	MONTH≠Sep	QUARTER≠Fourth
9	9373	11.8	CRS_ARR_TIME≠Early	SEASON≠WINTER	MONTH≠Sep

Table 5.6: Top-10 results of experiment Denver, max. splits 3, with WRAcc, a priori binning and diverse beam search with average delays for the subgroups. This table does not show any unexpected results, but it does show diversity

5.3 The influence of quality measures on coverage

5.3.1 Coverage as a metric for quality measures

The metric with which the impact of the quality measures of the results is judged, is the coverage. Each subgroup has a size attribute in the result files, which refers to the number of flights that fulfill the rule belonging to that subgroup. Size is also known as coverage and coverages of all experiments have been compared, resulting in some interesting findings.

5.3.2 Results

The first noticeable observation in the data is with regard to the different quality measures. The WRAcc quality measure gives bigger coverages as compared to the Mean Test. This is caused by the way in which the quality measures work. The Mean Test does not take the size of the subgroup compared to the size of the total dataset into account, while the WRAcc quality measure does take this into account, thus caring more for bigger subgroups. This is supported by Table 5.7 and Table 5.8: the coverages in Table 5.7 are way bigger.

With regard to the WRAcc quality measure, a priori binning seems to give the biggest coverages. Table 5.7 supports this with higher coverages in general for the AP columns as compared to the OTF columns. This is, however, not the case with the Mean Test, where on the fly binning appears to give the biggest coverages as seen in Table 5.8. So the quality measures can influence which type of discretization gives bigger coverages.

The number of bins appear to be influencing coverage. Results showed that the number of bins and the coverage appear to have a positive relation: as the number of splits increases, the coverage also increases. This is also shown in Table 5.7. The Mean Test, however, does not show this positive relation, as the coverages in Table 5.8 do not appear to increase or decrease as the maximum number of splits increases.

Lastly, diverse beam search gives bigger coverages than standard beam search with both quality measures. This is to be expected, since diverse beam search allows for subgroups to me more diverse. So with diverse beam search, it is possible to look at more promising, bigger subgroups than with standard beam search, which generally explores a smaller search space due to lack of diversity.

	Max. Splits: 3				**Max. Splits: 5**				**Max. Splits: 7**			
	BS		**DBS**		**BS**		**DBS**		**BS**		**DBS**	
	AP	**OTF**	**AP**	**OTF**	**AP**	**OTF**	**AP**	**OTF**	**AP**	**OTF**	**AP**	**OTF**
Denver	**7095**	6769	9173	**9232**	**8513**	7137	**10568**	8878	**12067**	6668	**12610**	9374
Tampa	7738	**8688**	11121	**11412**	**10837**	8224	12114	**11218**	**14304**	8547	**13832**	11046
San Diego	5351	**6941**	**8166**	8140	**8970**	6714	**9015**	7458	**11133**	7283	**9912**	7082
Average	6728	**7466**	9486	**9594**	**9440**	7358	**10565**	9184	**12501**	7499	**12118**	9167

Table 5.7: The average coverages of the top-10 results generated by the WRAcc quality measure. The columns represent the maximum number of splits in binning, standard beam search (BS) and diverse beam search (DBS), a priori binning (AP) and on the fly binning (OTF).

	Max. Splits: 3				Max. Splits: 5				Max. Splits: 7			
	BS		DBS		BS		DBS		BS		DBS	
	AP	OTF	AP	OTF	AP	OTF	AP	OTF	AP	OTF	AP	OTF
Denver	187	**218**	321	**625**	**315**	224	212	**358**	**279**	189	290	**303**
Tampa	264	**296**	256	**1146**	**283**	143	370	**508**	**261**	165	332	**484**
San Diego	142	**263**	167	**507**	232	**340**	215	**434**	**420**	244	307	**323**
Average	198	**259**	248	**759**	**277**	236	266	**433**	**320**	199	309	**370**

Table 5.8: The average coverages of the top-10 results generated by the Mean Test quality measure. The columns represent the maximum number of splits in binning, standard beam search (BS) and diverse beam search (DBS), a priori binning (AP) and on the fly binning (OTF).

5.3.3 What parameters give the biggest coverage?

The quality measures have the biggest impact on coverage, with the WRAcc resulting in the biggest coverages. Other parameter settings that resulted in the highest coverage were the highest number of splits, AP binning and DBS.

So based on the results, the experiments with the biggest coverage in general had the following parameters:

- Quality Measure: WRAcc
- Type of Binning: AP
- Beam Search Strategy: Diverse Beam Search
- Maximum Number of Splits: 7

5.4 The behaviour of quality measures on experiments

5.4.1 Evaluation based on rankings

Each quality measure gives scores that determine how interesting a subgroup is: generally, more interesting subgroups get higher scores than less interesting subgroups. Of each experiment, the average scores for the top-10 subgroups are shown in Table 5.9 for WRAcc and Table 5.10 for the Mean Test. These tables provide a ranking, as experiments with higher average scores are more interesting, and are therefore ranked higher than experiments with lower average scores. So with these tables, it is possible to see whether or not quality measures give higher scores (and therefore rank experiments higher) to certain experiments.

5.4.2 Results

The top-10 results show that both the Mean Test and WRAcc give higher rankings to experiments done with beam search compared to diverse beam search. This can also be observed in Table 5.9 and Table 5.10, where the BS columns get higher scores (on average) than the DBS columns.

With regard to the maximum number of splits (and the number of bins), both quality measures do not appear to give higher scores to certain experiments and therefore, do not rank certain experiments higher than other experiments. Table 5.9 and Table 5.10 show this with scores that do not get noticeably higher or lower as the maximum number of splits increases.

When taking a look at the AP columns in Table 5.9, it becomes clear that, in general, WRAcc ranks AP experiments higher than OTF experiments. With the Mean Test, this is only the case when looking at a maximum number of splits of three, as seen in Table 5.10. With a higher number of maximum splits, OTF experiments get ranked higher than AP experiments.

	Max. Splits: 3				**Max. Splits: 5**				**Max. Splits: 7**			
	BS		**DBS**		**BS**		**DBS**		**BS**		**DBS**	
	AP	**OTF**	**AP**	**OTF**	**AP**	**OTF**	**AP**	**OTF**	**AP**	**OTF**	**AP**	**OTF**
Denver	**3.8**	2.8	**2.7**	2.4	**3.5**	2.9	**2.6**	**2.6**	**3.1**	2.9	2.4	**2.6**
Tampa	**3.0**	2.9	**2.4**	2.3	**3.2**	2.8	**2.5**	2.4	**2.8**	**2.8**	**2.4**	**2.4**
San Diego	**2.6**	2.2	**2.0**	1.9	**2.4**	2.2	**2.0**	**2.0**	**2.4**	2.3	**2.1**	2.0
Average	**3.1**	2.6	**2.4**	2.2	**3.0**	2.7	**2.4**	2.3	**2.7**	**2.7**	**2.3**	**2.3**

Table 5.9: The average scores of the top-10 results generated by the WRAcc quality measure, where higher average scores are said to be ranked higher than lower average scores. The columns represent the maximum number of splits in binning, standard beam search (BS) and diverse beam search (DBS), a priori binning (AP) and on the fly binning (OTF).

	Max. Splits: 3				**Max. Splits: 5**				**Max. Splits: 7**			
	BS		**DBS**		**BS**		**DBS**		**BS**		**DBS**	
	AP	**OTF**	**AP**	**OTF**	**AP**	**OTF**	**AP**	**OTF**	**AP**	**OTF**	**AP**	**OTF**
Denver	**2115.5**	1997.9	**1527.1**	1333.3	1819.5	**2004.9**	1561.3	**1588.5**	1740.8	**1940.3**	**1576.2**	1543.2
Tampa	**1704.3**	1676.3	**1438.7**	1177.7	1760.9	**2090.6**	1494.2	**1530.4**	1794.2	**1802.0**	**1513.7**	1485.1
San Diego	**1560.4**	1415.0	**1387.3**	1112.7	1476.6	**1547.1**	**1265.0**	1112.9	1480.4	**1584.3**	1245.4	**1309.3**
Average	**1793.4**	1696.4	**1451.1**	1207.9	1685.7	**1880.9**	**1440.2**	1410.6	1671.8	**1775.5**	1445.1	**1445.9**

Table 5.10: The average scores of the top-10 results generated by the Mean Test quality measure, where higher average scores are said to be ranked higher than lower average scores. The columns represent the maximum number of splits in binning, standard beam search (BS) and diverse beam search (DBS), a priori binning (AP) and on the fly binning (OTF).

5.4.3 Do quality measures give higher rankings to certain experiments?

Standard beam search experiments get higher rankings than diverse beam search experiments with both quality measures. When using WRAcc, experiments with a priori binning get higher ranks than experiments with on the fly binning. With the Mean Test, this is only true for experiments with a maximum number of splits of three. With more splits, on the fly binning experiments get ranked higher. The maximum number of splits in itself does not seem to influence rankings.

5.5 The best strategy for finding subgroups with high delays

In the end, the goal of this work is to find interesting subgroups, so subgroups with a high deviation regarding arrival delays. This is done by looking at the average delays, which is calculated over all subgroups found per experiment. The average delays are shown in Table 5.11 and Table 5.12 for both WRAcc and Mean Test respectively.

5.5.1 Results

An important difference is seen in the quality measure. With the Mean Test quality measure, the average delays tend to be around 80 minutes, with some delays going well over 100 minutes. This is observed in Table 5.12. The Mean Test does, however, result in smaller subgroups (generally with a coverage below 1000). The WRAcc quality measure gives relatively lower delays, as seen in Table 5.11. The subgroups found with WRAcc are bigger than the subgroups found with the Mean Test. The smaller subgroups found with the WRAcc quality measure tend to have a coverage of +/- 5000, while the biggest coverages can go over 13000.

In Table 5.11 can also be seen that the OTF columns contain higher values than the AP columns, indicating that on the fly binning generally results in higher flight delays compared to a priori binning. This is true for the WRAcc quality measure. For the Mean Test, the opposite is true: Table 5.12 shows that AP columns tend to have bigger values than the OTF columns. This means that when using the Mean Test, a priori binning results in higher delays than on the fly binning in most cases.

The WRAcc quality measure and the Mean Test have in common that with both quality measures in general, standard beam search tends to give higher flight delays than diverse beam search. This is also noticeable in Table 5.11 and Table 5.12. The number of bins used does not appear to affect the flight delays found with the Mean Test. With the WRAcc quality measure, however, there seems to be a negative relation between the number of bins and the delays. Table 5.11 shows that, as the maximal number of splits increases, the delays tend to be lower.

	Max. Splits: 3				**Max. Splits: 5**				**Max. Splits: 7**			
	BS		**DBS**		**BS**		**DBS**		**BS**		**DBS**	
	AP	**OTF**	**AP**	**OTF**	**AP**	**OTF**	**AP**	**OTF**	**AP**	**OTF**	**AP**	**OTF**
Denver	**17.2**	15.8	**12.3**	10.8	7.9	**10.3**	6.9	**8.3**	6.6	**10.7**	6.2	**8.1**
Tampa	**8.8**	8.1	6.4	**6.6**	7.0	**8.7**	6.2	**6.8**	6.0	**8.6**	5.7	**6.7**
San Diego	8.1	**8.9**	**7.3**	**7.3**	**9.2**	8.8	6.6	**8.1**	8.0	**8.7**	6.7	**7.8**
Average	10.3	**11.0**	**8.7**	8.3	8.1	**9.3**	6.5	**7.7**	6.9	**9.3**	6.2	**7.5**

Table 5.11: The average flight delays for all experiments done with the WRAcc quality measure. The columns represent the maximum number of splits in binning, standard beam search (BS) and diverse beam search (DBS), a priori binning (AP) and on the fly binning (OTF).

	Max. Splits: 3				Max. Splits: 5				Max. Splits: 7			
	BS		DBS		BS		DBS		BS		DBS	
	AP	OTF	AP	OTF	AP	OTF	AP	OTF	AP	OTF	AP	OTF
Denver	**174.4**	95.7	**106.9**	93.4	127.7	**129.6**	102.5	**157.8**	**154.3**	110.5	**85.5**	65.9
Tampa	111.6	**119.3**	**100.5**	79.2	101.7	**128.1**	**92.6**	72.8	**105.4**	104.5	**104.8**	63.5
San Diego	**121.3**	97.1	**113.2**	59.1	**89.8**	63.8	**121.2**	40.5	**109.1**	89.3	**82.0**	69.0
Average	**138.0**	104.0	**106.0**	77.2	**110.5**	107.2	**105.4**	90.4	**130.3**	101.4	**89.2**	66.1

Table 5.12: The average delays for each experiment according to the Mean Test quality measure. The columns represent the maximum number of splits in binning, standard beam search (BS) and diverse beam search (DBS), a priori binning (AP) and on the fly binning (OTF).

5.5.2 What parameters result in the biggest delays?

Based on observations in the data, a trade-off occurs when it comes to finding subgroups with the biggest delays. The Mean Test results in the biggest delays, but gives smaller coverages, while WRAcc gives smaller delays, while giving larger coverage.

If the Mean Test is chosen, a priori binning combined with standard beam search results in the biggest delays. The maximum number of splits does not seem to influence the delays, so the largest number of bins is prefered as it had the least amount of information loss. So in this case, the following parameters are chosen:

- Quality Measure: Mean Test
- Type of Binning: AP
- Beam Search Strategy: Standard Beam Search
- Maximum Number of Splits: 7

This combination of parameters results in smaller subgroups with the highest delays and less diversity due to the AP binning. The fact that these parameters give smaller subgroups with the highest delays can mean that these parameters are, for example, suited for outlier detection.

If WRAcc is chosen, on the fly binning combined with standard beam search with a maximum number of splits of 3 will result in the biggest delays. This is represented with the following parameters:

- Quality Measure: WRAcc
- Type of Binning: OTF
- Beam Search Strategy: Standard Beam Search
- Maximum Number of Splits: 3

This combination of parameters does result in bigger subgroups and more diversity due to OTF binning, but with lower average delays. The relatively larger subgroups with lower delays can make that these parameters are suited for finding the effects of policy/strategy changes made by the management of carriers, as these changes can increase (or decrease) delays for specific subgroups.

5.5.3 Subgroups with highest delays

WRAcc

Table 5.13 takes a closer look at the top-10 rules for the experiment WRAcc, Denver, max. splits 3, BS and OTF binning. Note that this table shows many rules that contain "MONTH≠Nov", "MONTH≠Feb" and some rules contain "SEASON≠WINTER". A possible explanation for these rules showing up in the top-10 could be that these variables indicate the coldest period of the year. This cold weather, in combination with the holidays, different schedules and possibly less flights, could possibly explain the higher delays. Besides, Table 5.13 also shows many rules implying arrival and departure times later than 12:00 hours in the afternoon. This could imply that there are fewer delays in the morning. As these delays increase in amount, they can stack, resulting in growing delays as the day progresses.

When looking at the top-10 rules for the Tampa and San Diego airports with the same parameters (see Table 5.14 and Table 5.15 respectively), the same conclusions can be made as with the Denver airport. However, Table 5.14 shows some rules containing "LATEARRIVAL_RelativeHumidity>54.00000", while Table 5.15 has some rules that contain "DEPART_SeaLevelPressure<30.02000". It is most likely that these conditions apply to the different climates in which the airports operate as, for example, San Diego is positioned near the west coast of the United States. This can make that the DEPART_SeaLevelPressure variable influences the aircraft delays.

	size	μ^{Delay}	**Condition1**	**Condition2**	**Condition3**
0	8366	13.7	CRS_ARR_TIME>1546.00000	MONTH≠Nov	-
1	4639	19.3	CRS_ARR_TIME>1736.00000	MONTH≠Nov	SEASON≠WINTER
2	8286	13.7	CRS_ARR_TIME>1203.00000	CRS_DEP_TIME>1240.00000	MONTH≠Nov
3	5700	16.9	CRS_ARR_TIME>1736.00000	MONTH≠Nov	MONTH≠Feb
4	5700	16.9	MONTH≠Nov	CRS_ARR_TIME>1736.00000	MONTH≠Feb
5	5700	16.9	CRS_ARR_TIME>1736.00000	MONTH≠Feb	MONTH≠Nov
6	5700	16.9	MONTH≠Feb	CRS_ARR_TIME>1736.00000	MONTH≠Nov
7	8272	13.6	CRS_ARR_TIME>1203.00000	CRS_DEP_TIME>1240.00000	MONTH≠Feb
8	6902	15.0	CRS_ARR_TIME>1546.00000	SEASON≠WINTER	-
9	8422	13.5	CRS_DEP_TIME>1340.00000	MONTH≠Nov	-

Table 5.13: The top-10 rules and average delays for the experiment WRAcc, Denver, max. splits 3, BS and OTF binning. Note that these descriptions describe mainly the colder periods of the year and arrival/departure times scheduled later than 12:00 hours.

	size	μ^{Delay}	Condition1	Condition2	Condition3
0	8083	13.3	QUARTER≠Fourth	CRS_ARR_TIME>1254.00000	LATEARRIVAL_RelativeHumidity>54.00000
1	9856	11.7	MONTH≠Nov	CRS_ARR_TIME>1251.00000	LATEARRIVAL_RelativeHumidity>54.00000
2	8100	13.2	CRS_ARR_TIME>1250.00000	LATEARRIVAL_RelativeHumidity>54.00000	QUARTER≠Fourth
3	9864	11.7	CRS_ARR_TIME>1250.00000	LATEARRIVAL_RelativeHumidity>54.00000	MONTH≠Nov
4	8110	13.2	LATEARRIVAL_RelativeHumidity>54.00000	CRS_ARR_TIME>1249.00000	QUARTER≠Fourth
5	9878	11.7	LATEARRIVAL_RelativeHumidity>54.00000	CRS_ARR_TIME>1249.00000	MONTH≠Nov
6	8125	13.2	LATEARRIVAL_RelativeHumidity>54.00000	QUARTER≠Fourth	CRS_ARR_TIME>1246.00000
7	9160	12.2	LATEARRIVAL_RelativeHumidity>54.00000	CRS_DEP_TIME>1179.50000	MONTH≠Nov
8	7547	13.8	LATEARRIVAL_RelativeHumidity>54.00000	CRS_DEP_TIME>1179.50000	QUARTER≠Fourth
9	8162	13.1	CRS_DEP_TIME>1025.00000	LATEARRIVAL_RelativeHumidity>54.00000	QUARTER≠Fourth

Table 5.14: The top-10 rules and average delays for the experiment WRAcc, Tampa, max. splits 3, BS and OTF binning.

	size	μ^{Delay}	Condition1	Condition2	Condition3
0	5327	10.7	ARR_TIME>1643.00000	CRS_ARR_TIME<2057.00000	-
1	7144	8.6	DEPART_SeaLevelPressure<30.01000	CRS_ARR_TIME>1404.00000	-
2	7956	8.0	CRS_ARR_TIME>1179.50000	DEPART_SeaLevelPressure<30.02000	ARR_TIME_BLK≠1300-1359
3	8075	7.9	CRS_ARR_TIME>1179.50000	DEPART_SeaLevelPressure<30.02000	MONTH≠Feb
4	7375	8.4	CRS_ARR_TIME>1179.50000	DEPART_SeaLevelPressure<30.02000	ARRIVAL_SKYCONDITION1≠Clear
5	7375	8.4	CRS_ARR_TIME>1179.50000	DEPART_SeaLevelPressure<30.02000	ARRIVAL_SKYCONDITION2≠Clear
6	7375	8.4	CRS_ARR_TIME>1179.50000	DEPART_SeaLevelPressure<30.02000	ARRIVAL_SKYCONDITION3≠Clear
7	6949	8.7	DEPART_SeaLevelPressure<30.01000	DEP_TIME_BLK≠0700-0759	CRS_ARR_TIME>1403.00000
8	6815	8.7	CRS_ARR_TIME>1410.00000	DEPART_SeaLevelPressure<30.00000	-
9	5019	10.8	ARRIVAL_SeaLevelPressure<30.01000	CRS_ARR_TIME>1346.00000	DEPART_SeaLevelPressure<30.00000

Table 5.15: The top-10 rules and average delays for the experiment WRAcc, San Diego, max. splits 3, BS and OTF binning.

Mean Test

The top-10 rules for Denver, Tampa and San Diego are shown, respectively, in Table 5.16, Table 5.17 and Table 5.18 for the experiments with Mean Test, max. splits 7, BS and AP binning. These tables contain rules with many common conditions, which means that these results are not diverse. For example, the conditions "CRS_ARR_TIME=Very-Late" and "ARR_TIME≠Very-Late" appear together in rules many times. This does not provide any information, as it basically means that the flight did not arrive at the time it was supposed to arrive. The same can be said for the conditions "CRS_ARR_TIME=Very-Early" and "ARR_TIME≠Very-Early", which appear many times in Table 5.17.

	size	μ^{Delay}	Condition1	Condition2	Condition3
0	234	131	CRS_ARR_TIME=Very-Late	ARR_TIME≠Very-Late	ARR_TIME≠Late
1	331	105	ARR_TIME=Very-Late	CRS_ARR_TIME≠Very-Late	DEP_TIME_BLK≠2200-2259
2	331	105	ARR_TIME=Very-Late	CRS_ARR_TIME≠Very-Late	ARR_TIME_BLK≠0001-0559
3	328	105	ARR_TIME=Very-Late	CRS_ARR_TIME≠Very-Late	CRS_DEP_TIME≠Very-Late
4	330	101	CRS_ARR_TIME=Late	ARR_TIME≠Late	ARR_TIME≠Middle-Late
5	329	100	ARR_TIME=Very-Late	CRS_ARR_TIME≠Very-Late	CRS_ARR_TIME≠Very-Early
6	166	138	CRS_ARR_TIME=Very-Late	ARR_TIME≠Very-Late	DISTANCE_GROUP≠F
7	295	105	ARR_TIME=Very-Late	CRS_ARR_TIME≠Very-Late	ORIGIN_CITY_NAME≠Charlotte-NC
8	324	100	ARR_TIME=Very-Late	CRS_DEP_TIME≠Very-Late	ARR_TIME_BLK≠2100-2159
9	121	159	CRS_ARR_TIME=Very-Late	ARR_TIME≠Very-Late	DISTANCE_GROUP=C

Table 5.16: The top-10 rules and average delays for the experiment Mean Test, Denver, max. splits 7, BS and AP binning.

	size	μ^{Delay}	Condition1	Condition2	Condition3
0	153	156	CRS_ARR_TIME≠Very-Early	ARR_TIME=Very-Early	ARR_TIME_BLK≠2300-2359
1	293	112	CRS_ARR_TIME≠Very-Early	ARR_TIME=Very-Early	ARR_TIME_BLK≠0900-0959
2	213	129	CRS_ARR_TIME≠Very-Early	ARR_TIME=Very-Early	ORIGIN_CITY_NAME≠Charlotte-NC
3	212	129	CRS_ARR_TIME≠Very-Early	ARR_TIME=Very-Early	DISTANCE_GROUP≠C
4	292	110	CRS_ARR_TIME≠Very-Early	ARR_TIME=Very-Early	CRS_ARR_TIME≠Early
5	302	108	CRS_ARR_TIME≠Very-Early	ARR_TIME=Very-Early	DEP_TIME_BLK≠0700-0759
6	302	107	CRS_ARR_TIME≠Very-Early	ARR_TIME=Very-Early	CRS_DEP_TIME≠Early
7	230	122	CRS_ARR_TIME≠Very-Early	ARR_TIME=Very-Early	DEP_TIME_BLK≠2200-2259
8	293	106	CRS_ARR_TIME≠Very-Early	ARR_TIME=Very-Early	DEPART_SeaLevelPressure≠Very-High
9	317	102	CRS_ARR_TIME≠Very-Early	ARR_TIME=Very-Early	CRS_DEP_TIME≠Very-Early

Table 5.17: The top-10 rules and average delays for the experiment Mean Test, Tampa, max. splits 7, BS and AP binning.

	size	μ^{Delay}	Condition1	Condition2	Condition3
0	450	73	ARR_TIME=Very-Late	CRS_ARR_TIME≠Very-Late	CRS_ARR_TIME≠Very-Early
1	450	73	ARR_TIME=Very-Late	CRS_ARR_TIME≠Very-Late	ARR_TIME_BLK≠0001-0559
2	450	73	ARR_TIME=Very-Late	CRS_ARR_TIME≠Very-Early	CRS_ARR_TIME≠Very-Late
3	450	73	CRS_ARR_TIME≠Very-Early	ARR_TIME=Very-Late	CRS_ARR_TIME≠Very-Late
4	452	73	ARR_TIME=Very-Late	CRS_ARR_TIME≠Very-Late	CRS_DEP_TIME≠Very-Late
5	151	123	CRS_ARR_TIME≠Very-Early	ARR_TIME=Very-Early	CRS_ARR_TIME≠Early
6	461	71	ARR_TIME=Very-Late	CRS_ARR_TIME≠Very-Late	DEP_TIME_BLK≠2100-2159
7	433	73	ARR_TIME=Very-Late	CRS_ARR_TIME≠Very-Late	ARR_TIME_BLK≠2100-2159
8	441	73	ARR_TIME=Very-Late	CRS_ARR_TIME≠Very-Late	ARRIVAL_SeaLevelPressure≠Very-High
9	463	71	ARR_TIME=Very-Late	CRS_ARR_TIME≠Very-Late	DEP_TIME_BLK≠2200-2259

Table 5.18: The top-10 rules and average delays for the experiment Mean Test, San Diego, max. splits 7, BS and AP binning.

Chapter 6

Conclusions

This chapter will provide an answer to the research question, after which this work will be discussed and aspects of this work that can be improved upon will be given.

6.1 Answering the research question

The goal of this work was to find the best way to find interesting subgroups with regard to arrival delays. This was done by doing 72 experiments, with each experiment differing in binning techniques (a priori binning vs. on the fly binning), the maximum number of splits (3, 5 and 7 splits), the beam search strategy (standard beam search vs. diverse beam search) and quality measures (WRAcc vs. Mean Test). These experiments were done on flights from three different sized airports with different weather conditions and climates. By judging the results of experiments on diversity, coverage and flight delays, the following research question can be answered:

How can subgroup discovery be used in the best way to discover interesting subgroup sets, with flight delays as target, on a large dataset of flight data and weather data?

Based on the experiments, two answers can be given to this question and they are summarized in Table 6.1. Each of these answers has advantages and disadvantages. The first answer in Table 6.1 is finding subgroups with the Mean Test, a priori binning, standard beam search and maximum 7 splits, while the second answer consists of finding subgroups with WRAcc, on the fly binning, standard beam search and a maximum of 3 splits. The first answer results in finding subgroups with higher delays than the second answer, but the first answer also results in the smallest coverages as compared to the second answer. When looking at diversity, the second answer scores better with a higher diversity than the first answer.

So, when finding subgroups with the highest delays has priority, doing subgroup discovery with the Mean Test, a priori binning, standard beam search and a maximum of 7 splits provides the best answer to the

research question. When diversity and coverage are important, doing subgroup discovery with WRAcc, on the fly binning, standard beam search and a maximum of 3 splits answers the research question best.

This results in a trade-off between results that give smaller, less diverse subgroups with high average delays (see first answer in Table 6.1); and results that give larger, more diverse subgroups with lower average delays (see second answer in Table 6.1). The properties of the first answer make the combination of parameters well-suited for tasks like outlier detection, as the small number of flights per subgroup with relatively large delays can possibly be classified as outliers. When the goal is to analyze larger subgroups with more diversity, the second answer will give the best results. This combination of parameters can, for example, be used for finding effects of policy or strategy changes, as the coverages of subgroups are relatively large with lower delays as compared to the first answer. These changes can then possibly increase or decrease the arrival delays for specific subgroups, which can be found with the second answer from Table 6.1.

#	Quality Measure	Type of Binning	Beam Search Strategy	Max. Splits
1	Mean Test	AP	Standard	7
2	WRAcc	OTF	Standard	3

Table 6.1: The two possible answers to the research question.

6.2 Discussion

In Chapter 4, equal frequency binning techniques used in this work were explained. The number of bins is chosen based on the distribution of the attributes in order to capture areas where data is most dense. This is, however, an approximation, as it is not certain that the data will be split perfectly as presented in Chapter 4 due to the fact that equal frequency binning results in bins of equal size, not equal width.

Furthermore, Chapter 5 describes how the number of unique conditions is used as a metric for diversity. However, it could be the case that with some experiments, rules were selected with more patterns than other experiments. This means that experiments where this is the case have less unique conditions per definition, in which case it can be unfair to compare results of these experiments.

Regarding the preprocessing step, a lot of the extreme weather conditions were left out of the dataset because the threshold was put at 3% (see Chapter 3). Flights that did not endure extreme weather conditions had empty values in the dataset, and only a very small number of flights endured extreme weather conditions. This means that a lot of the extreme weather condition attributes were removed from the dataset, which had an advantage and a disadvantage. During test runs of the algorithm with the extreme weather condition attributes still included in the dataset, the majority of the subgroups found had descriptions that showed conditions implying that extreme weather did not occur. This was not useful, as they did not add much information and took up the space of at least one condition in the descriptions of subgroups. So the advantage of removing such attributes is that there was more room for conditions that could potentially add more useful information. However, the disadvantage is that removing the majority of the extreme weather condition attributes means

that the few cases where the extreme weather conditions did appear in the descriptions, were not found during this study.

Lastly, one mistake was made during the selection of attributes. The ARR_TIME and CRS_ARR_TIME attributes were left in the dataset, but should have been removed during preprocessing. These attributes represent the time of arrival and scheduled time of arrival, respectively. In some experiments, these two attributes appeared frequently, representing that a flight did not arrive at the scheduled arrival time and thus was delayed. However, this did not add any new information, as subgroups that showed up in the results of experiments always had flights that were delayed. Therefore, these attributes should have been removed.

6.3 Future work

Several aspects of this work can be improved upon in the future. First of all, the size of the dataset used is a relatively small part of all available data, since only domestic flights of three airports were used. The use of more airports will give a better understanding of the effects of weather and climate, airport sizes and other factors on arrival delays. Furthermore, almost no extreme weather conditions appeared in the preprocessed dataset, as there were too few occurrences of such weather conditions with just three airports. Increasing the size of the dataset will also result in more extreme weather conditions to use for further research.

Secondly, this work is limited to only using cover based diverse beam search. Besides this diverse beam search strategies, DSSD offers other two other variants of diverse beam search, known as "description" and "compression" in the DSSD settings file. Future works could include comparisons between the description based and compression based beam search strategies on the same type of dataset, to see what different results they give.

Thirdly, only the domestic flights of one carrier (United Airlines) were used. This work could be extended by comparing flights of multiple airlines in order to compare delays per carrier. This can be used to study the effects of strategies and policies of individual carriers on delays and competitiveness.

Lastly, as described in Chapter 4, the number of splits for numeric attributes was determined by manually inspecting the distributions of attributes. Future work could include a way of doing this automatically, which will save time and be more accurate as compared to the manual inspections.

Bibliography

[1] B. of Transportation Statistics, "Flight data - on time performance,"

[2] M. Q. Mirza, "Climate change and extreme weather events: can developing countries adapt?," *Climate Policy*, vol. 3, no. 3, pp. 233 – 248, 2003.

[3] A. Sternberg, D. Carvalho, L. Murta, J. Soares, and E. Ogasawara, "An analysis of brazilian flight delays based on frequent patterns," *Transportation Research Part E: Logistics and Transportation Review*, vol. 95, pp. 282 – 298, 2016.

[4] M. J. Post, "Understanding flight delays,"

[5] A. for America, "U.s. passenger carrier delay costs,"

[6] V. Deshpande and M. Arkan, "The impact of airline flight schedules on flight delays," *Manufacturing & Service Operations Management*, vol. 14, no. 3, pp. 423–440, 2012.

[7] A. Martin, "Subgroup discovery," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 5, no. 1, pp. 35–49.

[8] S. Helal, "Subgroup discovery algorithms: A survey and empirical evaluation," *Journal of Computer Science and Technology*, vol. 31, pp. 561–576, May 2016.

[9] M. van Leeuwen and A. Knobbe, "Diverse subgroup set discovery," *Data Mining and Knowledge Discovery*, vol. 25, pp. 208–242, Sep 2012.

[10] F. Lemmerich, M. Atzmueller, and F. Puppe, "Fast exhaustive subgroup discovery with numerical target concepts," *Data Mining and Knowledge Discovery*, vol. 30, pp. 711–762, May 2016.

[11] N. Lavrač, B. Cestnik, D. Gamberger, and P. Flach, "Decision support through subgroup discovery: Three case studies and the lessons learned," *Machine Learning*, vol. 57, pp. 115–143, Oct 2004.

[12] D. Seland, "GARBAGE IN GARBAGE OUT," *Quality*, vol. 57, pp. 439–473, May 2018.

[13] N. Oceanic and A. Administrations, "Domestic weather data,"

[14] M. C. Peel, B. L. Finlayson, and T. A. Mcmahon, "Updated world map of the Köppen-Geiger climate classification," *Hydrology and Earth System Sciences Discussions*, vol. 4, pp. 439–473, Mar. 2007.

[15] M. Atzmueller, "Subgroup discovery," *Hydrology and Earth System Sciences Discussions*, Mar. 2005.

Appendices

Appendix A

Overview of bins per attribute

Attribute	Max. 3 splits	Max. 5 splits	Max. 7 splits
CRS_DEP_TIME	3	5	7
CRS_ARR_TIME	3	5	7
ARR_TIME	3	5	7
ARR_DELAY	-	-	-
DryBulbCelcius*	3	5	7
RelativeHumidity*	3	5	7
WindSpeed*	3	5	7
ValueForWindCharacter*	3	5	7
StationPressure*	2	6	10
SeaLevelPressure*	3	5	7
HourlyPrecip*	2	3	4

Table A.1: An overview of the numeric attributes and their number of bins. ARR_DELAY was not binned as it was the target attribute. A * means that an attribute has three different variations: a DEPART, ARRIVAL and LATEARRIVAL variation. The DEPART prefix indicates that the attribute tells something about the weather at the departure station at the scheduled time of departure; the ARRIVAL prefix indicates that the attribute tells something about the weather at the arrival station at the time of departure and the LATEARRIVAL prefix indicates that the attribute tells something about the weather at the arrival station at the scheduled time of arrival.

Attribute	Bins	Number of bins
QUARTER	First, Second, Third, Fourth	4
MONTH	Jan, Feb, ... , Nov, Dec	12
DAY_OF_MONTH	Weekday, Weekend	2
DAY_OF_WEEK	Monday, Tuesday, ... , Saturday, Sunday	7
DISTANCE_GROUP	A, B, C, D, E, F, G, H, I, J	10
SEASON	Summer, Autumn, Winter, Spring	4
Visabiltity*	IFR, VFR	2
WindDirection*	All 8 wind directions plus a value for when there is no wind	9

Table A.2: An overview of the numeric attributes and their binning that stayed constant across experiments. A * means that an attribute has three different variations: a DEPART, ARRIVAL and LATEARRIVAL variation. The DEPART prefix indicates that the attribute tells something about the weather at the departure station at the scheduled time of departure; the ARRIVAL prefix indicates that the attribute tells something about the weather at the arrival station at the time of departure and the LATEARRIVAL prefix indicates that the attribute tells something about the weather at the arrival station at the scheduled time of arrival.

The SKYCONDITION attributes, together with all weather attributes not listed in this appendix, were used in exactly the same way as Post [4] did.