Identifying flight delay patterns using diverse subgroup discovery

Hugo M. Proença
LIACS
Leiden University
Leiden, Netherlands
h.manuel.proenca@liacs.leidenuniv.nl

Ruben Klijn
LIACS
Leiden University
Leiden, Netherlands
r.a.klijn@umail.leidenuniv.nl

Thomas Bäck
LIACS
Leiden University
Leiden, Netherlands
T.H.W.Baeck@liacs.leidenuniv.nl

Matthijs van Leeuwen
LIACS
Leiden University
Leiden, Netherlands
m.van.leeuwen@liacs.leidenuniv.nl

***Abstract*—Flight delay is a common hassle that affects around one of each four flights and has been a major concern for airlines for decades. As a consequence, an increasing amount of research was done on this topic in recent years. Notably, the fields of machine learning and data mining have proposed various solutions for the prediction of flight delays, typically hours to days before departure. However, the most important decisions made by airlines that could benefit from such predictions, i.e., those on scheduled block time and crew schedules, are made between *two to six months* prior to departure. Consequently, late delay predictions are useless for these scheduling tasks.**

As accurately predicting delays for individual flights a long time in advance is practically infeasible, we instead propose to search for circumstances that tend to lead to large delays. For this we propose to use diverse subgroup discovery, a data mining technique that allows to discover subgroups, i.e., subsets of the data that 1) deviate from the overall data with regard to some target variable, and 2) can be described by a simple conjunctive query on the other variables.

We apply diverse subgroup discovery to historic flight data and mine subgroups of flights that, on average, have a large delay. We show that this approach gives subgroups that can be easily understood by experts, despite the fact that non-trivial relations between multiple variables can be discovered. We show that using diverse subgroup discovery gives less redundant results than standard top-k subgroup discovery and demonstrate that even in situations where inferring an accurate predictive model is infeasible, local deviations can be effectively captured and described by local patterns, potentially providing valuable insights for, e.g., airline scheduling problems.

***Index Terms*—subgroup discovery, flight delays, data mining**

I. Introduction

In 2017, the airline industry carried more than 500 billion passengers worldwide [1]. Even relatively rare flight delays can thus affect millions of persons every year, which not only decreases airline revenues and consumer satisfaction, but also produces unnecessary carbon emissions. In 2017, 19% of the (more than) 1.8 million flights in the US alone were delayed [2]. Taking into account that the number of passengers is predicted to double over the next 20 years, while airport capacity is not expected to increase that much, it is highly likely that these numbers will increase [1]. This corroborates the need for effective solutions for reducing flight delays in the near future.

Research on the topic of flight delays has seen a fourfold increase between 2010 and 2016 [3]. The topic has been approached from the fields of machine learning, econometrics, operations research, statistics, and probability, which has resulted in methods that aim at, e.g., prediction of flight delays [4]–[6], network delay propagation [7]–[9], and cancellation policies of airlines [10]. Interestingly, most of this research has been restricted to the last hours or days prior to departure, when data sources are abundant. For example, weather data and information on the state of the airline network, both strongly associated with delay, are typically used. Until now, there has not been much interest in long term prediction of (chronic) delays, probably because this is a very hard task, if at all feasible: information that is available long in advance, such as airplane schedules and origin–destination airports pairs, is typically only very weakly associated with flight delays. Regardless, we argue that even if it is infeasible to infer accurate, global models for the prediction of individual flight delays, historic flight data still contains relevant information that can provide valuable insight into temporal and spatial aspects of flight delay—already at the crucial moment when airline schedules need to be decided.

Approach and contributions. To identify flight delay patterns from data, we propose to use *subgroup discovery*, a data mining technique that allows to discover subgroups, i.e., subsets of the data that 1) deviate from the overall data with regard to some target variable, and 2) can be described by a simple conjunctive query on the remaining variables. Given a dataset, a query language, and a quality measure, subgroup discovery algorithms find queries describing large subsets of the data having large deviations in the target variable.

More specifically, we use *diverse subgroup discovery*, which improves over basic subgroup discovery by identifying sets of subgroups that are less redundant [11].

We apply diverse subgroup discovery to historic flight data and mine subgroups of flights that, on average, have a large delay. The purpose of our experiments is to demonstrate what kind of insights an airline team in charge of scheduled block times and/or crew assignment could use when making its decisions, which typically happens six to two months prior to take-off [12]. We therefore solely use data that is available six months prior to departure. As an example, the best subgroup obtained using arrival delay as target variable is described by the following query: $date \in [26/03, 31/08] \wedge dep. \in [11:30, 23:59] \wedge arr. \in [15:40, 3:50]$. This query can be read as: flights departing from the 26^{th} of March to the end of August, with departure times between $11:30$ and $23:59$, arrival times between $15:40$ and $3:50$. These flights constitute 21% of the dataset and have an average delay of 20 minutes, compared to an overall average of 7 minutes.

The main contributions can be summarized as follows: 1) we propose the use of subgroup discovery for the identification of flight delay patterns; 2) we demonstrate the benefits of using diverse subgroup discovery, as opposed to standard subgroup discovery, on a real-world problem; and 3) we show that part of the chronic flight delays can be described using only variables that are available 6 months prior to departure, thereby contributing to an understudied problem in the airline industry. By being able to take into account circumstances that structurally lead to delays, the use of historical data will ultimately enable airlines to make their operation more robust.

II. Related work

The research on the topic of flight delays can be divided in four parts [3]: problem; scope; data; and method. Our main contributions are in the realm of the *scope*: we study flight delays with variables available far in advance; and the *method*: we use subgroup discovery to identify regions of interest in the dataset. For this reason we will restrict our related work to: 1) machine learning and data mining applications to flight delays; and 2) subgroup discovery.

A. Flight delays

With the recent affluence of large and cheap storage facilities, and substantial computational power, data science techniques have started to play an important role in the study of flight delays [3]. Frequent pattern mining, the unsupervised counterpart of subgroup discovery, was used to analyze flight delays in the Brazilian network [13]. This work has some similarity to ours as it also uses patterns to study regions of the dataset that are more prone to delays, but four key differences distinguish our work. First, they analyze flight delays as a binary target, either delayed or not, while we analyze flight delays as numeric targets, taking into account the length of delays. Second, their analysis is guided by asking a question and fixing the respective variables, while we only choose an interestingness measure and obtain all our results automatically without manual selection. Thirdly, they need to make an *a priori* discretization of their continuous variables, while we allow our algorithm to automatically find the best intervals for our variables during search. Lastly, their work also takes into account weather data and delay state of the airports, while we only focus on scheduled dates and origin-destination airports pairs.

Machine learning algorithms have been able to harness the information available in historical data to accurately predict flight delays in different contexts. To mention some interesting applications: random forests were applied to predict departure delays 2–24 hours in advance in the 100 most delayed links of the US system [5]; the predictions of an adaptive fuzzy network were used as input in a decision support system for arrivals in JFK airport [6]; and reinforcement learning algorithms were used for the prediction of taxi-out times in Tampa international airport [4]. Even though they also use historical data, their concern is to produce a *global* model, that can perform predictions, where we focus on *locally* describing delays in an human interpretable way.

In general, regression and classification also aim at finding the relationship between the explanatory and response variable through historic data, but they are concerned with finding a *global* model of how the variables interact with each other. Subgroup discovery, on the other hand, finds *local* models that describe the behavior in specific subsets of the dataset.

B. Subgroup discovery

Subgroup discovery was first introduced in the $90s$ [14], with algorithms such as Opus [15] and Explora [16]. It has been applied to a vast array of applications [17], such as identifying the properties of materials [18] and unusual consumption patterns in smart grids [19]. Recently, for the specific case of numerical targets, an efficient exhaustive search algorithm was proposed to reduce the need to resort to heuristics [20], and a novel way of taking into account the dispersion of the target variable of the groups was introduced [21].

Common approaches to reduce the redundancy among subgroups encompass supervised pattern set mining [22] and methods based on relevance [23] and diversity [11], [24] (as used in this work). Unlike diversity-based methods, the supervised pattern set mining objective is to find a fixed number of patterns, which has to be chosen in advance, and relevance is limited to non-numeric targets.

III. Data acquisition and preprocessing

Historic data of flights in the United States is freely available through the Bureau of Transportation Statistics (BTS) [2].

We restrict our analysis to a single airline, a single year, and the busiest operating airports for that airline. The reason for this is that airlines have different strategies when choosing the schedule time for their routes, and these can also change with a periodicity of six months [12]. Also, the results should reveal which schedules are associated with more delays so that an airline can improve, and using multiple airlines with different strategies can muddle these insights. The restriction

to the busiest airports comes from the fact that they are more prone to chronic delays than smaller ones, while also covering most flights of an airline. To make the analysis more relevant, the year of 2017 was selected, a normal year of airline operation when compared with the years of 2007 and 2008, where the recession affected the operation of some airports [2]. After that, the major US airline with most delays during this period was chosen: American Airlines (AA), with circa 20% of flights delayed [2]. Lastly, we restricted our study to the top-3 departure airports and their top-3 arrival airports, as they cover about 5% of all the flights of AA and account for most of delays.

After all canceled and diverged flights have been removed the acquired data totals $36\,149$ samples; the used variables can be seen in Table I. The original features downloaded from the BTS are: date; CRS (schedule local time) departure time; CRS arrival time; CRS elapsed time; distance group; origin airport; destination airport; arrival delay; and departure delay. The date was transformed to six variables: the metereological season, the respective month, the day of the year, day of the week, if it is a week day (or weekend), and if it is a national holiday. The scheduled local (CRS) arrival and departure times were kept as they were and duplicated to departure and arrival time 12, which are rotations of the variables starting at $12:00$ instead of $00:00$. This will take into account the fact that time is a circular variable that jumps from $23:59$ to $00:00$ and allow for complementary intervals such as: $22:00 < dep.\,time < 04:00$, which otherwise would be impossible. Elapsed time, distance group, origin, and destination were kept unchanged.

TABLE I: Description of the variables of the dataset taken from the US Federal Bureau of Statistics, for the top-3 busiest airports and the top-3 busiest destinations of American Airlines in 2017. For each variable, given are its type $\{binary, categorical, numeric\}$, its range (numeric) or its number of unique values (binary/categorical), and its usage in the model: either to describe the subset (explanatory) or to define the distribution of interest (target).

Variable	Type	Range/cardinality	Usage
Season	categorical	4	explanatory
Month	numeric	$[1, 12]$	
Day of the year	numeric	$[0, 365]$	
Day of week	numeric	$[1, 7]$	
Week day	binary	2	
Holiday	binary	2	
Departure time	numeric	$[00:00, 23:59]$	
Departure time 12	numeric	$[12:00, 11:59]$	
Arrival time	numeric	$[00:00, 23:59]$	
Arrival time 12	numeric	$[12:00, 11:59]$	
Elapsed time	numeric	$[101, 229]min.$	
Origin	categorical	3	
Destination	categorical	5	
Arrival delay	numeric	$[-58, 1079]min.$	target
Departure delay	numeric	$[-30, 1099]min.$	target

IV. Subgroup Discovery

Subgroup discovery is an exploratory data mining technique used to find interesting subsets of a dataset [17]. A subset is described by a conjunctive query on the explanatory variables, which we will call a pattern. In the case of flight delays a pattern could take the form of $X = \{Origin = San\ Francisco \wedge Season = Summer\}$, with $\wedge$ being the logical AND operator. Combined with a target variable, e.g., "Arrival Delay", the pattern forms a rule, which describes the distribution of arrival (or departure) delay of the subset formed by that description. Note that only one target variable can be used at a time. As the number of possible combinations of variables in a dataset grows exponentially with the number of variables, their interestingness has to be measured to select the best descriptions. This interestingness is evaluated by means of a quality function that, typically, quantifies to what extent a subset differs from the overall dataset distribution [17]. In the flight industry, the population distribution of the arrival and departure times is given by the mixture of distributions of all flight schedules as they were arranged by their respective airlines. Our objective is to find the most interesting patterns that deviate from these "normal" schedules.

The software used in our experiments to find subgroups is the Diversity Subgroup Set Discovery (DSSD) tool [11], and it was chosen for its diverse subgroup set selection and fast implementation.

The rest of this section is organized as follows: 1) query language: the language used to describe the subsets and target variable; 2) quality measure: how the interestingness of subsets is quantified; and 3) diverse subgroup selection: how diverse search is established.

A. Notation and query language

Consider a dataset D composed of one flight $(\mathbf{x}, y)$ per row, with a vector of explanatory variables $\mathbf{x}$ and a numeric target variable y. Explanatory variables are used to explain the target. The vector $\mathbf{x}$ is composed of n variables, with their respective values of the form: $\mathbf{x} = (x_1, x_2, ..., x_n)$. These variables can be of three types: numeric, categorical, and binary. Each type of variable can be queried according to different logical operators: numeric with greater or less than $\{>, <\}$; binary with equal to $\{=\}$, categorical with equal or not equal to $\{=, \neq\}$. The query formed by the conjunction of specific variable values or intervals is called a pattern X, e.g., $X = \{\text{Origin}^{categorical} \neq \text{San Francisco} \wedge \text{Week day}^{binary} = \text{yes} \wedge \text{Departure Time}^{numeric} \geqslant 11:00\}$, where the superscript denotes the type of variable. A pattern X over dataset D implies a subgroup, formally given by

$$D_X = \{(\mathbf{x}, y) \in D \mid X(\mathbf{x}) = true\}, \tag{1}$$

where $X(\mathbf{x})$ is a predicate that returns positive if the conditions of pattern X are satisfied by tuple $\mathbf{x}$, and $|D_X|$ is the

coverage of the subgroup, i.e., the number of samples in the dataset for which $X(\mathbf{x})$ returns positive.

The empirical probability distribution of the numeric target y over the whole dataset D will be referred to as $\hat{p}_D(y)$, and over a subgroup D_X as $\hat{p}_X(y)$. Thus a pattern X implies a rule from query to numeric target of the form:

$$X \rightarrow \hat{p}_X(y) \tag{2}$$

Importantly, we consider *only the means* of the empirical distributions just mentioned. That is, we assume that all deviations between $\hat{p}_X(y)$ and $\hat{p}_D(y)$ that are of interest can be observed from the difference between their respective means.

B. Quality measure

To assess the quality (or interestingness) of a pattern X, a measure that scores and ranks subsets D_X needs to be chosen. The measures used in subgroup discovery vary depending on the target and/or application [17]. In the case of flight delays, either arrival or departure delay can be the numeric target. A quality measure for a numeric target usually takes the following form

$$q(X) = |D_X|^a(\mu_X - \mu_D),\ a \in [0, 1], \tag{3}$$

where $|D_X|$ is the number of samples covered by the pattern X, μ_X and μ_D are the mean of the empirical distribution of the subgroup, $\hat{p}_X(y)$, and the dataset, $\hat{p}_D(y)$, respectively. Parameter a is a coefficient that allows to control the trade-off between coverage and the difference of the means. The separation of the quality measure into two parts helps to find generalizable patterns: if only the difference of means $(\mu_X - \mu_D)$ would be used, its maximization would be trivially achieved by selecting the pattern with highest difference, independently of how much data it covers. Given our interest in patterns that have a larger delay, we selected the *Mean Test* [11], with $a = 0.5$, giving relatively large importance to the difference of the means:

$$q_M(X) = \sqrt{|D_X|}(\mu_X - \mu_D) \tag{4}$$

C. Diverse subgroup set selection

Classically, subgroups are ranked by their qualities and then the top-k patterns are presented, with k being a value selected by the user [17]. The problem with this approach is that the top-k results tend to be redundant, especially when the number of variables and samples is large [11].

This is especially important in the case where explanatory variables are redundant, as it is in our case, where for example the month of January can be described by $day\,of\,year < 32$ or $month < 2$, and the same is true for other cases such as departure, arrival, and elapsed time, or day of week and week day, and so forth.

Without diversity the top-k results would overlap and be variations of the same pattern. For this reason, we use diversity-based selection. In particular, DSSD offers three different approaches: cover-based; compression-based; and description-based selection. Compression-based selection is immediately excluded as it can only be applied to information-theoretic measures, such as Kullback-Leibler divergence. Description-based does not help solving our redundancy problem, as our redundancy does not stem from the repetition of variables in a description but from the associations between our variables. The diversity measure left that fits our goal is cover-based selection, where subgroup qualities are penalized when they cover the same region of the dataset. We will proceed to describe cover-based selection.

Cover-based beam search first selects the highest-scoring subgroup based on the quality measure (in our case the M-score (4)), and then updates the score of subsequent candidates by introducing a weight if they cover the same part of the space of previous selected subgroups. The weighting scheme used is based on multiplicative weighted covering [25]:

$$\Omega(X, Sel) = \frac{1}{|D_X|} \sum_{(\mathbf{x}) \in D_X} \alpha^{c(\mathbf{x}, Sel)},\ \alpha \in]0, 1] \tag{5}$$

where Sel are the subgroups already selected, $c(\mathbf{x}, Sel)$ is the number of subgroups that already cover this sample and α is a weight parameter. The new selection score for ranking the subgroups is the product of the chosen quality measure and the diversity score:

$$q(X) = \Omega(X, Sel) \cdot q_M(X) \tag{6}$$

The candidates for the first subgroup will have $\Omega(X_1, \varnothing) = 1$, as there are no subgroups selected yet. After the first subgroup is selected, the second subgroup will have a weight of $\Omega(C_i, X_1)$, which will be updated for each candidate C_i depending on their overlap with X_1 according to (6).

D. DSSD algorithm

The DSSD algorithm is a heuristic that allows to find a good approximation of the best subgroups given a quality measure [11]. A high-level description is given in Algorithm 1. The algorithm can be divided in three distinct phases that we will briefly explain next; for an in-depth characterization of each phase, please refer to Van Leeuwen and Knobbe [11].

The first phase (line 1) consists of a beam-search of fixed (or varying) width b to find the top-j subgroups, where j should be a number much larger than the final number of subgroups wanted. The beam search varies depending on the type of diversity exploration the user selects (parameter S in Alg. 1): none (standard beam search); description; cover; or compression. The generation of candidates proceeds top-down, where new candidates are generated by adding one condition to subgroups in the beam at each iteration, after which the beam is updated with new candidates, until the maximum depth selected by the user is reached. The generation of new candidates by adding conditions is called refinement and depends on the type of variable: binary, categorical, or numeric. For the binary case, only the operator $=$ is used to construct candidate conditions, e.g., $weekday = yes$. For the categorical case, $=, \neq$ can be used, e.g., $season \neq spring$. For the numeric case, $<, >$ are used, e.g., $departure\,time < 11{:}00$. For numeric variables DSSD has an "on-the-fly" binning

technique that generates up to a maximum number of equal-sized bins specified by the user for each subgroup where the variable occurs, allowing for a more fine-grained discretization of the variables. After the maximum depth search is achieved, the top-j subgroups are returned to enter the next phase.

In the second phase (line 2) dominance-based pruning is applied to the top-j subgroups obtained from phase one. Dominance-based pruning tests the existence of subgroups that are strict subsets of subgroups of phase one that have a higher quality than the original, and in that case these subsets replace their supersets.

In the third phase (line 3) k subgroups, with $k \ll j$, are selected based on the strategy S chosen for the beam search: none, description, cover, or compression; to remove any redundancy that was left after the previous phases.

Algorithm 1 The DSSD algorithm

Input: Dataset D, quality measure q, number of top subgroups j and k, minimum coverage $minc$, maximum size of subgroups $maxd$, and specific selection strategy parameters S

Output: $\mathcal{R}$, a diverse set of subgroups with high quality

1: $\mathcal{R} \leftarrow$ Diverse-Beam-Search$(D, q, j, minc, maxd, S)$
2: $\mathcal{R} \leftarrow$ Dominance-Pruning$(\mathcal{R})$
3: $\mathcal{R} \leftarrow$ Subgroup-Selection$(\mathcal{R}, q, k, S)$
4: **return** $\mathcal{R}$

V. Experiments

In this section we present the results of our experiments, obtained using the dataset as described in Section IV-D.

First, a single variable analysis of arrival and departure delays will be shown. Then a baseline subgroup discovery algorithm—DSSD without diversity strategy—will be applied to departure delays, to compare against its diverse counterpart. Finally, diverse subgroup discovery will be applied with both departure and arrival delay as target.

We present only the best ten subgroups found by the different methods, as this is also how domain experts would typically inspect the results in a real-world scenario.

DSSD parameters. The DSSD algorithm, described in Section IV-D, has several parameters that need to be set. The parameters chosen for our experiments were the following: type of search: *beam*; *fixed* beam width 100; quality measure: *mean test* of eq. (4); diversity strategy: *none* for non-diverse analysis of Section V-B, and *cover* for the diverse analysis of Sections V-C and V-D; *weighted covering* and *multiplicative covering* with *cover multiplier* $\alpha = 0.9$ for the multiplicative cover-based selection of eq. 5; maximum description length 5; minimum coverage 10; $j = 10\,000$ (top-j of the first phase); $k = 100$ (top-k of the third phase). Most parameters were selected similarly to the recommendations by Van Leeuwen and Knobbe [11], with the exception of j, which was chosen large enough as to accommodate the larger dataset of 30 000 samples.

Description of the subgroups. The variables of Table III have some overlap and can represent the same description with different variables, which can be hard to interpret. For this reason we chose to translate the results obtained with DSSD to a less redundant structure, as follows:

- $date$ – the days and months of the year that make the subgroup interval, in dd/mm format, where dd is the day of the month and mm is the month of the year.
- $dep.$ and $arr.$ – the departure and arrival time, respectively, in $hh\!:\!mm$ format, where hh stands for hours and mm minutes.
- day_month, day_week – the day of the month and week, respectively, in numeric format $[1, 12]$ and $[1, 7]$, where the 1st day of the week is Monday.
- $orig.$ and $dest.$ – origin and destination airport respectively.
- $time$ – the flight scheduled duration in minutes.

A. Single variable analysis of arrival and departure delays

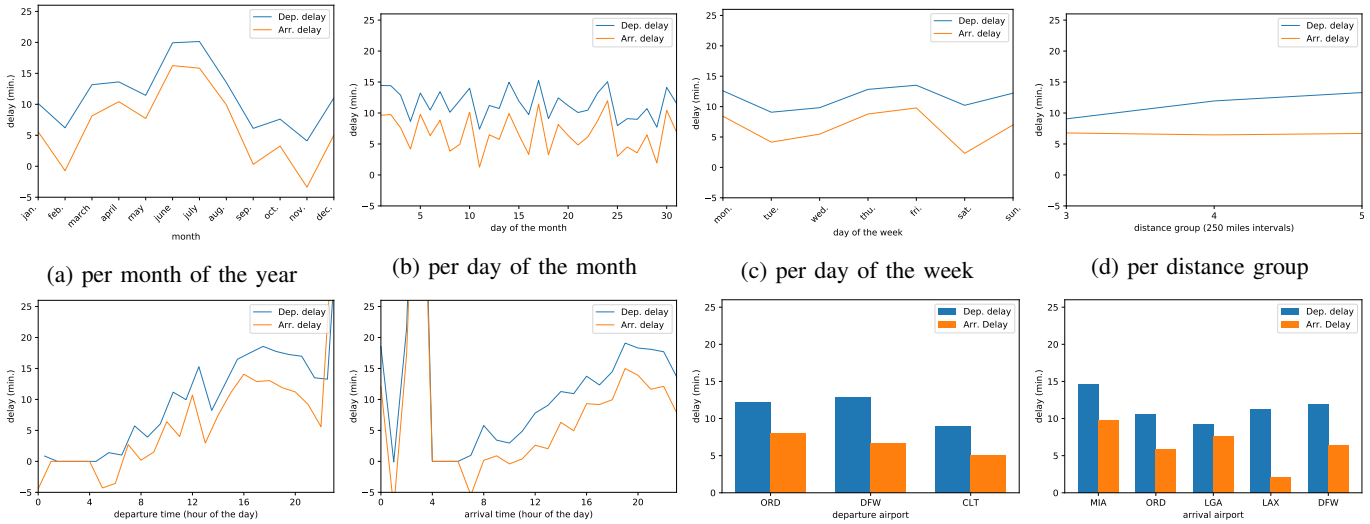
The goal of this analysis is to show to what extent it is possible to find informative subgroups using a single variable analysis. Average arrival and departure delays were averaged over: month of the year; day of the month; day of the week; distance group; departure time; arrival time; origin airport; and arrival airport. Other variables, such as season, day of the year, week day, and elapsed time, were not plotted as their structure can be seen through the eight selected variables.

The results can be seen in Figure 1. In general, the summer months, June to August, seem to imply longer delays, while fall and winter months seem to imply shorter delays. Also late departure/arrivals seem to be associated with longer delays. In the case of arrival it is worth noticing that there is a delay spike around $03\!:\!00$. This happens because at this hour there are only 10 flights and one of them suffers from a major delay, up to 10 hours. LAX (Los Angeles International Airport) seems to have relatively short arrival delays: 2 minutes compared to a dataset average of 7 minutes. For the other variables, i.e., day of the month, distance group, and departure airport, there seems to be no visible relation with delay as their averages remain similar throughout their respective ranges.

From this point onwards the single variable analysis could only be extended to subgroups with more than one condition by manually combining visually promising variables, which would require an intensive 'trial and error' process until manually finding the right intervals of the variables. Even then, this slow approach might not result in the best possible subgroups, hence the need for an automated approach to subgroup discovery.

B. Top-10 non-diverse subgroups for departure delay

For this experiment departure delays were used for non-diverse subgroup discovery, to have a comparison baseline against the diverse search. It should be noted that DSSD without diverse selection is similar to most standard subgroup discovery approaches and represents a baseline algorithm to compare to. The top-10 subgroups can be seen in Table II.



(a) per month of the year (b) per day of the month (c) per day of the week (d) per distance group

(e) per departure hour (f) per arrival hour (g) per origin airport (h) per destination airport

Fig. 1: Average arrival and departure delay for different temporal and spatial variables

After the subgroups were translated to the format described in the beginning of this section, it turned out that most descriptions were the same although written with different variables, and that only three distinct subgroups were actually found: $1-4$, $5-6$ and $7-10$. After closer inspection it can be observed that all subgroups are a variation of the first one, with all the ten subgroups having the same date range. The only changes are the usage of departure instead of arrival time in the subgroups $5-6$, and by not selecting the 31^{st} day of the month and including the holidays in the subgroups $7-10$. This shows that for a dataset such as ours—with redundant variables—enforcing diversity is necessary. (Note that it would also be possible to do feature selection to remove seemingly redundant variables, but it is usually not clear a priori which ones should be removed to allow for clear and concise patterns.)

C. Diverse subgroup sets for departure delay

In the next experiment, we use DSSD with cover-based diversity and departure delay as target variable. The diverse subgroup set, consisting of ten subgroups found through iterative scoring and selection, can be seen in Table III. On average they cover 22% of the dataset and have a delay of 22 minutes, compared to a dataset average of 12 minutes.

The first departure delay subgroup covers 23% of the dataset and has an average delay of 23 minutes. It thus affects almost one fourth of the dataset and has a delay that is almost double the average. This is clearly an interesting part of the data that should be taken into consideration by schedulers. On the other hand, the subgroup with the highest delay, 30 minutes, covers 8% of the dataset, and has as condition that the day of the month should be smaller than 24^{th}. This is especially interesting as in the single variable analysis of Figure 1 there seems to be no clear relationship between delays and the day of the month.

Compared with the non-diverse results of Table II, it can be seen that the diversity measure helps to obtain subgroups that cover different regions of the dataset. However, it should be said that most subgroups tend to focus around the central months of the year of June to August, and late departures/arrivals in general. This is interesting as even penalized by the diversity measure there is still an overlap, which tells us that this central region of the dataset is particularly prone to large departure delays.

D. Diverse subgroup set for arrival delay

In this experiment we used DSSD with cover-based diversity and arrival delay as target variable. The best ten subgroups obtained can be seen in Table IV. On average they cover 22% of the dataset, and have a delay of 19 minutes, compared to a dataset average of 7 minutes.

The best description is centered around the middle of the year, from the end of March to the end of August, for flights departing after noon. The subgroup with the largest delays is the 5^{th}, which has an average delay of 23 minutes and affects 11% of the flights.

Similarly to departure delays it can be noticed that most subgroups are centered around the same dates, departure, and arrival times. This is a clear indicator that this is the part of the data that exhibits the highest delays affecting most airplanes. Also, the subgroups of this experiment are very similar to the ones obtained for departure delays in Table III, which is expected as departure delay is correlated with arrival delay.

E. Discussion of the results

In general all the results agree that the summer months—June to August—and later departures tend to lead to longer delays and are more prevalent than others.

It is interesting to observe the difference between the top-10 non-diverse and the diverse results of Tables II and III,

TABLE II: Top-10 non-diverse subgroups obtained departure delays as target variable. Rank, average delay in minutes, M-score, coverage in percentage, and description of the subgroup.

#	Del. (min.)	$q(X)$	Cov. (%)	Description of the subgroup
1	22	1010	24	$date \in [01/03, 02/09] \wedge dep. \in [13:59, 23:59] \wedge holiday = no$
2	22	1010	24	$date \in [01/03, 02/09] \wedge dep. \in [13:59, 23:59] \wedge holiday = no$
3	22	1010	24	$date \in [01/03, 02/09] \wedge dep. \in [13:59, 23:59] \wedge holiday = no$
4	22	1010	24	$date \in [01/03, 02/09] \wedge dep. \in [13:59, 23:59] \wedge holiday = no$
5	22	1007	27	$date \in [01/03, 02/09] \wedge arr. \in [15:52, 03:42] \wedge holiday = no$
6	22	1007	27	$date \in [01/03, 02/09] \wedge arr. \in [15:52, 03:42] \wedge holiday = no$
7	22	1006	24	$date \in [01/03, 02/09] \wedge dep. \in [13:59, 23:59] \wedge day\_month \in [1, 30]$
8	22	1006	24	$date \in [01/03, 02/09] \wedge dep. \in [13:59, 23:59] \wedge day\_month \in [1, 30]$
9	22	1006	24	$date \in [01/03, 02/09] \wedge dep. \in [13:59, 23:59] \wedge day\_month \in [1, 30]$
10	22	1006	24	$date \in [01/03, 02/09] \wedge dep. \in [13:59, 23:59] \wedge day\_month \in [1, 30]$

TABLE III: Diverse subgroup set obtained with departure delay as target variable. Dataset average delay is 12 minutes. Rank, average delay in minutes, M-score, coverage in percentage, and description of the subgroup.

#	Del. (min.)	$q(X)$	Cov. (%)	Description of the subgroup
1	23	1015	23	$date \in [01/03, 31/08] \wedge dep. \in ([11:30, 11:59] \vee [13:30, 23:59]) \wedge arr. \in [16:35, 11:59]$
2	30	1012	8	$date \in [01/06, 31/08] \wedge day\_month \in [1, 24] \wedge dep. \in [11:30, 23:59] \wedge arr. \in [15:45, 03:49]$
3	22	1004	28	$date \in [01/03, 31/08] \wedge dep. \in [11:30, 23:59] \wedge arr. \in [15:20, 23:59] \wedge holiday = no$
4	21	988	27	$date \in [01/03, 31/08] \wedge dep. \in [11:30, 23:59] \wedge arr. \in [13:30, 03:49] \wedge holiday = no$
5	22	1007	27	$date \in [01/03, 31/08] \wedge dep. \in [11:30, 23:59] \wedge arr. \in [15:30, 03:49] \wedge holiday = no$
6	20	836	24	$date \in [01/01, 29/08] \wedge dep. \in [14:30, 23:59] \wedge arr. \in [16:10, 05:59] \wedge time \in [130, \inf[\wedge orig. \neq CLT$
7	24	884	13	$date \in [01/05, 11/08] \wedge day\_month \in [1, 24] \wedge arr. \in [12:00, 09:25] \wedge time \in [130, \inf[\wedge holiday = no$
8	22	885	19	$date \in ([24/02, 31/08] \vee [01/12, 31/12]) \wedge dep. \in [13:20, 22:00] \wedge arr. \in [18:44, 23:02]$
9	21	865	24	$date \in [01/01, 30/08] \wedge dep. \in [14:55, 21:56] \wedge arr. \in [13:41, 22:56]$
10	19	751	30	$date \in [01/12, 31/08] \wedge dep. \in [13:20, 22:00] \wedge arr. \in [12:00, 05:54] \wedge time \in [130, 204[\wedge orig. \neq CLT$

TABLE IV: Diverse subgroup set obtained with arrival delay as target variable. Dataset average delay is 7 minutes. Rank, average delay in minutes (Del.), M-score, coverage in percentage, and description of the subgroup.

#	Del. (min.)	$q(X)$	Cov. (%)	Description of the subgroup
1	20	1120	21	$date \in [26/03, 31/08] \wedge dep. \in [11:30, 23:59] \wedge arr. \in [15:40, 3:50]$
2	21	1110	16	$date \in [01/03, 02/08] \wedge day\_week \in [1, 6] \wedge dep. \in [13:20, 23:59] \wedge dest. \neq LAX$
3	19	1106	24	$date \in [26/03, 31/08] \wedge dep. \in [11:30, 23:59] \wedge arr. \in [15:40, 3:50] \wedge holiday = no$
4	17	1066	29	$date \in [23/02, 31/08] \wedge dep \in [11:30, 23:59] \wedge arr \in [15:20, 3:50]$
5	23	1021	11	$date \in [01/06, 02/08] \wedge day\_month \in [1, 24] \wedge arr. \in [12:00, 3:50]$
6	18	1036	22	$date \in ([23/02, 31/08] \vee [01/12, 31/08]) \wedge day\_week \in [1, 6] \wedge arr. \in [15:51, 23:59] \wedge dest. \neq LAX$
7	18	1035	24	$date \in [03/03, 31/08] \wedge arr. \in [13:41, 22:55] \wedge dest. \neq LAX$
8	16	987	33	$date \in ([23/02, 31/08] \vee [01/12, 31/08]) \wedge dep. \in [15:20, 03:49] \wedge arr. \in [15:20, 23:59] \wedge dest. \neq LAX$
9	20	967	15	$date \in [01/06, 20/08] \wedge day\_month \in [1, 24] \wedge arr. \in [12:00, 9:25] \wedge holiday = no$
10	18	955	21	$date \in [01/12, 31/08]) \wedge dep. \in [12:00, 21:15] \wedge arr. \in [13:41, 22:56] \wedge dest. \neq LAX$

respectively. These results reveal the importance of enforcing diversity when searching for subgroups that show distinct perspectives on variables associated with delay. In the non-diverse case most subgroups were redundant and in practice only one subgroup was found in the top-10.

Arrival and departure delay are clearly correlated, as can be seen from the figures of single variable analysis and the similar subgroups obtained for departure (Table III) and arrival (Table IV). However, there are some differences, as arrival delays subgroups tend to have earlier arrival times, and more often include the day of the month and day of the week as variables. For departure delay we also found that flights have longer delays if they do not depart from CLT, and arrival delays tend to be longer subgroups if the destination airport is not LAX. Finally, it should be noted that both descriptions are centered around June to August, which coincides with the dates of school vacations in the US.

From the point of view of scheduling, the information provided by the subgroups could contribute to making schedules more robust with regard to delays. This could be done by, e.g.,

providing more resources to those groups of flights that tend to have longer delays. From this perspective our subgroup discovery approach seems useful, as the descriptions can be very specific. For example, the flights of the subgroup with the longest arrival delay in Table IV have an average delay that is 16 minutes longer than the overall average, and can be attributed to around 11% of the flights.

VI. Conclusion

We used diverse subgroup set discovery (DSSD) to find groups of flights with relatively long delays from historic flight data. In doing so, were the first to apply DSSD to a real world application. The results show that potentially useful associations can be discovered between flight delay and explanatory variables that are available six months before the flight takes place. Specifically, the results suggest that variables seemingly unrelated to delay, such as day of the week and day of the month, can help to identify large subgroups of flights that suffer from relatively long delays. To achieve this, the diverse selection of subgroups turned out be crucial, as it helped to strongly reduce the redundancy in the result sets. The proposed approach may provide valuable insights for two important scheduling problems in the airline industry, i.e., for block time and crew scheduling.

Future work. As future work we plan to validate the obtained results on an external dataset, such as 2018 flight data once it is available, to assess how well the obtained subgroups generalize. Further, restricting the analysis to a certain type of delays, such as air traffic control delay, could return more specific subgroups associated with the given type of delay.

Acknowledgment

This work is part of the research programme Indo-Dutch Joint Research Programme for ICT 2014 with project number 629.002.201, SAPPAO, which is financed by the Netherlands Organisation for Scientific Research (NWO).

References

[1] (2017) IATA annual review. [Online]. Available: http://www.iata.org/publications/

[2] (2018) The Bureau of Transportation Statistics (BTS). [Online]. Available: https://www.bts.gov/

[3] A. Sternberg, J. Soares, D. Carvalho, and E. Ogasawara, "A review on flight delay prediction," *arXiv preprint arXiv:1703.06118*, 2017.

[4] P. Balakrishna, R. Ganesan, and L. Sherry, "Accuracy of reinforcement learning algorithms for predicting aircraft taxi-out times: A case-study of tampa bay departures," *Transportation Research Part C: Emerging Technologies*, vol. 18, no. 6, pp. 950–962, 2010.

[5] J. J. Rebollo and H. Balakrishnan, "Characterization and prediction of air traffic delays," *Transportation research part C: Emerging technologies*, vol. 44, pp. 231–241, 2014.

[6] S. Khanmohammadi, C.-A. Chou, H. W. Lewis, and D. Elias, "A systems approach for scheduling aircraft landings in jfk airport," in *Fuzzy Systems (FUZZ-IEEE), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1578–1585.

[7] J.-T. Wong and S.-C. Tsai, "A survival model for flight delay propagation," *Journal of Air Transport Management*, vol. 23, pp. 5–11, 2012.

[8] N. Pyrgiotis, K. M. Malone, and A. Odoni, "Modelling delay propagation within an airport network," *Transportation Research Part C: Emerging Technologies*, vol. 27, pp. 60–75, 2013.

[9] N. Xu, G. Donohue, K. B. Laskey, and C.-H. Chen, "Estimation of delay propagation in the national aviation system using bayesian networks," in *6th USA/Europe Air Traffic Management Research and Development Seminar*. FAA and Eurocontrol Baltimore, MD, 2005.

[10] J. Xiong and M. Hansen, "Modelling airline flight cancellation decisions," *Transportation Research Part E: Logistics and Transportation Review*, vol. 56, pp. 64–80, 2013.

[11] M. van Leeuwen and A. Knobbe, "Diverse subgroup set discovery," *Data Mining and Knowledge Discovery*, vol. 25, no. 2, pp. 208–242, 2012.

[12] D. Klabjan, "Large-scale models in the airline industry," in *Column generation*. Springer, 2005, pp. 163–195.

[13] 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.

[14] A. Siebes, "Data surveying: Foundations of an inductive query language." in *KDD*, 1995, pp. 269–274.

[15] G. I. Webb, "Opus: An efficient admissible algorithm for unordered search," *Journal of Artificial Intelligence Research*, vol. 3, pp. 431–465, 1995.

[16] W. Klösgen, "Explora: A multipattern and multistrategy discovery assistant," in *Advances in knowledge discovery and data mining*. American Association for Artificial Intelligence, 1996, pp. 249–271.

[17] M. Atzmueller, "Subgroup discovery," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 5, no. 1, pp. 35–49, 2015.

[18] B. R. Goldsmith, M. Boley, J. Vreeken, M. Scheffler, and L. M. Ghiringhelli, "Uncovering structure-property relationships of materials by subgroup discovery," *New Journal of Physics*, vol. 19, no. 1, p. 013031, 2017.

[19] N. Jin, P. Flach, T. Wilcox, R. Sellman, J. Thumim, and A. Knobbe, "Subgroup discovery in smart electricity meter data," *IEEE Transactions on Industrial Informatics*, vol. 10, no. 2, pp. 1327–1336, 2014.

[20] F. Lemmerich, M. Atzmueller, and F. Puppe, "Fast exhaustive subgroup discovery with numerical target concepts," *Data Mining and Knowledge Discovery*, vol. 30, no. 3, pp. 711–762, 2016.

[21] M. Boley, B. R. Goldsmith, L. M. Ghiringhelli, and J. Vreeken, "Identifying consistent statements about numerical data with dispersion-corrected subgroup discovery," *Data Mining and Knowledge Discovery*, vol. 31, no. 5, pp. 1391–1418, 2017.

[22] B. Bringmann and A. Zimmermann, "The chosen few: On identifying valuable patterns," in *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*. IEEE, 2007, pp. 63–72.

[23] H. Großkreutz, D. Paurat, and S. Rüping, "An enhanced relevance criterion for more concise supervised pattern discovery," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012, pp. 1442–1450.

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

[25] N. Lavrač, B. Kavšek, P. Flach, and L. Todorovski, "Subgroup discovery with cn2-sd," *Journal of Machine Learning Research*, vol. 5, no. Feb, pp. 153–188, 2004.