Les Cahiers du GERAD ISSN: 0711–2440

Flight deck crew absenteeism: From data to forecasting

A.-H. Homaie-Shandizi, V. Partovi Nia, M. Gamache, B. Agard

G–2015–108

October 2015

GERAD HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada H3T 2A7

Tél. : 514 340-6053
Téléc. : 514 340-5665
info@gerad.ca
www.gerad.ca

Flight deck crew absenteeism: From data to forecasting

Amir-Hosein Homaie-Shandizi [a]
Vahid Partovi Nia [b]
Michel Gamache [b]
Bruno Agard [a]

[a] *Department of Mathematics and Industrial Engineering, Polytechnique Montréal, Montréal (Québec) Canada, H3C 3A7*

[b] *GERAD & Department of Mathematics and Industrial Engineering, Polytechnique Montréal, Montréal (Québec) Canada, H3C 3A7*

vahid.partovinia@polymtl.ca
michel.gamache@polymtl.ca
bruno.agard@polymtl.ca

October 2015

Les Cahiers du GERAD
G–2015–108

Abstract: Airline companies are subject to a considerable amount of disruptions during their operations. It is vital for many industries including the airline industry to predict the sources of disruptions in different levels of management in order to reduce schedule recovery costs. One of the most important and costly sources of disruption in the airline industry is absenteeism of pilots at the time of a flight operation.

We propose a supervised learning method to predict total monthly absence hours of pilots. The proposed method uses characteristics of the monthly schedule as factors and, using an iterative algorithm, makes a prediction. The model was tested with real data and a substantial improvement was observed in the results.

Key Words: Forecasting, decision trees, absenteeism.

Résumé : Toutes les compagnies aériennes sont sujettes à un nombre considérable d'interruptions dans leurs opérations. Il est vital pour plusieurs industries, y compris les compagnies de transport aérien, de prédire les causes de ces interruptions à différents niveaux de gestion afin de réduire les coûts associés à la mise à jour des horaires. Une des plus importantes et coûteuses sources d'interruptions dans le domaine du transport aérien provient de l'absence des pilotes au moment d'opérer les vols.

On propose une méthode d'apprentissage supervisée pour prédire le nombre d'heures d'absence des pilotes. La méthode proposée utilise les caractéristiques des horaires mensuels comme facteurs et, en utilisant un algorithme itératif, fait des prévisions. Le modèle a été testé sur des données réelles et des améliorations substantielles ont été observées.

Acknowledgments: We would like to thank MITACS and our industrial partner for their financial support in this project.

1 Introduction

Airline companies face many kinds of disruptions during operations and they have to spend millions of dollars each year on disruption management issues, ranging from weather conditions to security issues. The airline industry is involved with many uncontrolled situations that may cause changes to their schedules. Flight delays, flight cancellations, passenger dissatisfaction, etc., are only few of these numerous challenges.

The loss of revenue as a result of such disruptions has prompted airlines to have strategic cost-saving plans in order to reduce this loss. Airline disruptions are inevitable, but by using different mathematical and engineering tools, it is feasible to have continuous improvement in minimizing these losses.

One of the ways to improve the decision making tools that are used in disruption management is with scientific prediction for future random events. Although in any situation and for any random event it is impossible to make an exact prediction, having a systematic prediction will cover a portion of the uncertainty. Another reason to use prediction in industries is related to the nature of decision-making systems in business. Having good predictions leads to a reduction in the number of decisions that must be taken at the operation level. At this level, the decisions must be taken faster but are perhaps less optimal compared to the tactical level decisions.

In the airline industry, the cost of the crew is the second most significant cost after fuel, and obviously pilots are the most important member of an airline crew; without them a flight is not even possible. It is difficult to replace a pilot with another one because every pilot is qualified for just one position. However, during the operations, there are different situations in which a reserve pilot must replace a planned pilot: for instance, when a pre-selected pilot does not show up.

It is important to predict the number of required reserves well. The wrong number of reserves can cause two different extra operational costs. First, if the number of reserves is greater than the number of absent pilots, the company pays the non-operational pilots. Second, if the number of reserves is less than the number of absent pilots, then the airline calls an out-of-duty pilot for a much higher cost or, in the worst case, this can lead to the cancelation of a flight. Therefore, costs of under-prediction are higher than costs of over-prediction.

The number of pilots in reserve for the block of a month must be determined in advance in order to cover the needs of replacements. Obviously, the number of reserve pilots depends on the monthly schedule. In high seasons, when there is more demand for travel, airlines encounter more flight hours so the number of reserve pilots should be higher than in low seasons.

We attempt to develop a methodology for predicting the absence hours of pilots based on the flight schedule and history of absenteeism by using decision trees as a statistical tool. Prediction and forecasting with data mining has already proved to be successful in numerous industrial applications (Verma et al., 2013), (Suryadevara et al., 2013) in particular with decision trees (Li and Chan, 2010), (Pham et al., 2014) even in the present application domain of airlines (Burdun and Parfentyev, 1999). This methodology was developed for use in a major airline and the results show considerable improvement with prediction.

In Section 2, the problem is described in detail. The proposed methodology for solving this problem is presented in Section 3. The results of implementing this methodology in an airline are reported in Section 4. The paper will end with a conclusion.

1.1 Disruption management

Unlike strategic and tactical problems, during flight operations the majority of problems need to be solved in a short period of time. Therefore, managing irregular operations, called disruptions, is a subject of considerable interest.

In the airline industry, disruptions can occur for several reasons: mechanical problems, weather conditions, crew absenteeism, security, etc. These kinds of problems may cause flight delays or even flight cancelations. However, in many cases, crew reassignment is still feasible.

One of the first works on disruption management in the airline discipline is the two minimum-cost flow network models presented by Jarrah et al. (1993) for reducing shortages. The first model chooses the set of delayed flights and the second model chooses the set of cancelations. Based on these models, a decision support system (DSS) was implemented at United Airlines and the result of valuable cost savings for using this DSS was published by (Rakshit et al., 1996).

From a different point of view, Bratu and Barnhart (2006) deal with the airline schedule recovery problem and develop an optimal trade-off between airline operation costs and passenger delay costs. They consider either passenger disruption or delay cost.

Kohl et al. (2007) discuss developing a system that uses multiple resource methods and integrating these resources to improve the quality of the decision-making process. They indicate that the development of flexible tools must be considered in order to have an added-value contribution to the business. They also conclude that emphasizing finding an optimum solution in the strict academic sense without weighing on the operational restrictions cannot be applicable in real situations. Cauvin et al. (2009) propose a multi-agent approach to the problem in a disrupted and distributed environment. They propose a method based on existing methods for managing disruptions. They suggest identifying the actors, their interactions and their consequent activities in a disruptive environment.

Disruption management in the airline industry has been an increasingly active domain throughout the past decades, but in most of the cases, the proposed solutions consider just one aspect of the problem, e.g. aircraft type, crew, passenger, etc. This is an important field of research since there is a fundamental gap between the proposed prototype tools by software companies and the ideal integrated recovery tool; see Clausen et al. (2010) for more information.

A model for estimating the number of required reserve crews for covering aircraft delays callout is considered by Gaballa (1979). Gaballa minimized costs of both reserve crews and overnight delays. The implementation of this method provided considerable cost savings at Qantas Airlines.

Another example of a disruption management system is an automated system that has been implemented at US Airways. This system constructs optimal scheduling for a reserve crew through an advanced reserve bid line (Dillon and Kontogiorgis, 1999).

Wei et al. (1997) develop a modeling framework for the crew reassignment by using a heuristic branch-and-bound search algorithm. Their proposed algorithm is more flexible compared to the traditional operational research approaches. They also consider business rules to bind the optimal solution.

Lettovski et al. (2000) claim that it is necessary to reduce the complexity of the problem for crew reassignment during the operations. They suppose that the published schedule is optimum, and by using a tree-based data structure, they generate integer optimal solutions in a short time.

The prediction of passenger show rate can be accomplished by using logical analysis of data (LAD), an approach presented by Dupuis et al. (2012). The objective of their study is to find conditions in which the airline can allow over-booking on flights that some passengers may not show up to at the gate.

RESOPT (reserve optimization) is another model developed by Sohoni et al. (2006) that effectively increases reserve availability. Their model needs a good estimator for open-time reserve demand to be used as a reserve manpower controller.

Another automated decision support tool has been developed by Abdelghany et al. (2004). The tool can be used in large-scale commercial airlines that use the hub-spoke network structure for the crew recoveriy problem. A hub-spoke network is a network in which all the points are connected through spokes to the hubs instead of a point-to-point connection. This tool is flexible in different scenarios and can proactively manage the future disruptions in a chain.

1.2 Classification and regression tree

Classification and regression trees were presented by Morgan and Sonquist (1963) as an automatic interaction detection technique. Two decades later, Breiman et al. (1984) developed the first modern and comprehensive

algorithm for growing trees. Their famous method, CART, is a fundamental basis for classification and regression trees and the book of Breiman (1993) on the classification and regression trees is now a classic reference.

For a long time, classification and regression trees (CART) have been popular for modeling and predicting among statisticians, machine learning experts, and data mining practitioners. Like many other methods, these tree-based models are used when there is a response variable, Y, and some predictors $X_1, X_2, \ldots, X_p$. Regression tree is used when the response variable is a continuous variable while the usage of a classification tree is for cases with a categorical response. We consider here just a binary decision tree i.e. the decision trees with a two level response variable. Although decision trees with n-level response variables, called n-ary decision trees, exist in theory and practice, we do not consider this part of literature because the nature of our problem is binary; the pilots are either absent or present. For further reading and an in-depth discussion of this subject, see Chapter 9 of Hastie et al. (2009), Chapter 6 of Witten et al. (2011), and the classic textbook of Breiman et al. (1984).

The main idea of tree-based models is to partition the variable space into non-overlapping rectangles and fit a constant in each partition. This is a simple but powerful idea, since a large class of functions can be approximated by a piece-wise constant (step) function. Furthermore, a constant model is computationally fast to fit. The fitting process requires only averaging over the response variable of observations that belong to that partition. Over-fitting has been recognise has a delicate problem in different industrial applications (Karabadji et al., 2014).

A decision tree has a hierarchical top-down order and at each node just one variable splits the space of inputs. In decision trees, the algorithm chooses a variable and a splitting point on that variable by using an impurity measure. The splitting criteria with the stopping rule make the growing phase of the decision tree. There are two types of splitting criteria, univariate and multivariate.

Univariate splitting criteria create each node using just one variable; therefore the splitting function is univariate. These are the well-known criteria that are used in almost all of the most well-known algorithms. *Information Gain* (Quilan, 1993) uses the entropy measure as the impurity measure. *Gini Index* (Breiman et al., 1984) is the divergence measure defined over probability distribution of the response variable. *Likelihood ratio chi-square statistic* (Ciampi et al., 1987) in addition provides statistical inference about the information gain. Normalizing information gain by the entropy leads to the *Gain Ratio* (Quinlan, 1993). *Twoing Criteria* (Breiman et al., 1984) is the same as the Gini Index for the binary response model and a more accurate generation of it for multi-level response variables.

Multivariate splitting criteria create each node by a linear combination of variables (Breiman et al., 1984) and (Sethi and Yoo, 1994). It is obvious that finding the best solution is more complicated and less interpretable, so multivariate splitting criteria are less popular in practice.

Another important criterion for making a tree is the stopping criteria being applied at the end of the growing phase. Common stopping criteria are one of the following (Rokach and Maimon, 2005):

- In each terminal node, there exists just a single value of the response.
- The maximum tree depth reaches a pre-specified limit.
- The number of cases in the node is lower than a pre-specified value.
- The best splitting criterion is not greater than a pre-specified threshold.

After the tree is grown, and until a stopping rule is met, the tree is pruned to keep the balance between bias and variance of the model for an accurate prediction.

One of the problems that occurs after growing the tree is the over-fitting, i.e. the training accuracy is high, while prediction accuracy is low. In order to increase prediction accuracy, pruning is necessary (Bohanec and Bratko, 1994).

There are two types of pruning. First, where pruning is a part of the tree construction, it is called *pre-pruning*. Second, where pruning is a separate procedure after the growing phase, it is called *post-pruning* (Esposito et al., 1997).

There are many pruning methods (Rokach and Maimon, 2005). Among them, *cost-complexity pruning* (*cp*) is the most popular one. The *cp* is a post-pruning method proposed by Breiman et al. (1984).

2 Problem description

2.1 Crew scheduling

Crew scheduling for the airlines consists of different tasks and interested readers are directed to the textbooks on airline operations such as Barzagan (2010) or Grosche (2009). For a detailed analysis of preferential bidding systems at airlines see Gamache et al. (1998) and Barnhart et al. (2003).

The first step in the crew scheduling process, as can be seen in Figure 1, is publishing the list of monthly flights. Based on tactical and strategic decisions, a list of flights for a business month is published by the commercial department of airlines. Each flight has its own planned characteristics such as flight departure, arrival date and time, assigned aircraft type, departure and arrival airports, flight duration, flight credits, etc.

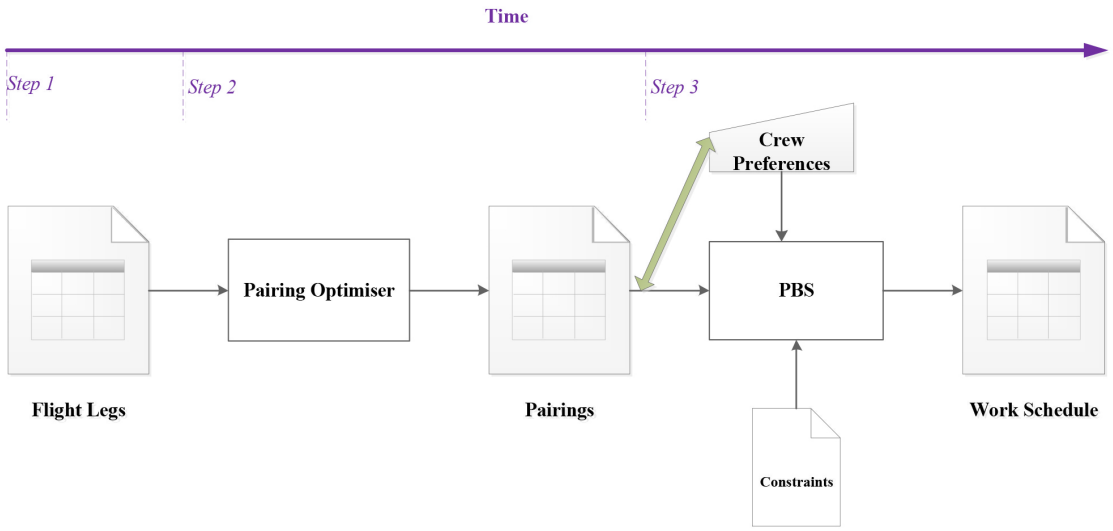

Figure 1: Crew scheduling process in airlines.

Despite other industries in which the working duty consists of shifts or days, in airlines there is an additional duty period called *pairing*. A pairing is a combination of consecutive flight legs which starts and ends at the same domicile.

After publishing the pairing table, in the second step a computer program called the *preferential bidding system* optimises the airline workforce schedule. The preferential bidding system is usually executed for a monthly schedule and its inputs could be *airline operations requirements* (list of the pairings), *crew preferences* (bidding) and some other *constraints* such as government regulations, collective bargaining agreements and airline policies.

After pairings have been made, the bidding process begins. In the bidding process, each crew requests a certain schedule or determines their preferences. Then, the preferential bidding system is assigned to the list of the pairings of the pilots, considering all of the constraints. This process minimizes the costs of the operations and matches aircraft type, flying routes, and pairings in a way that each pairing is assigned to only one qualified pilot.

2.2 Problem overview

Being absent depends on many different environmental and personal conditions. However, in this study we do not consider these conditions as the factors for modeling because the predictions are performed exactly after publishing a pairing schedule; at that time, pilots have not been assigned to the pairings (see Figure 2). Therefore, we can only include pairing characteristics as the covariates of the model. As a result of using these covariates, if there exists undesirable characteristics of the pairing that some pilots prefer to use their yearly pay absent days, this model is able to distinguish those characteristics.

Another reason for considering pairing characteristics as the covariates of the model is the fact that in the airline industry, pilots bid on the pairings and not on flights. Therefore, we may expect that pairing characteristics have an effect on their choices. Some pairings are for 4 or 5 consecutive days and some pairings are performed on only one day. Some of them contain a lot of deadhead credit, in which pilots transfer to the duty airport as a passenger, and some happen during the weekend and on holidays. As there is no available information about the pilots' preferences in this study, we consider pairing characteristics as the only covariates of the model and will extract their hidden information.

Figure 2 shows a simplified flowchart of airline operations from the scheduling phase to the end of operations. The main objective is predicting the monthly total absence hours of pilots after publishing the schedule of pairings based on the monthly schedule and also by using past records of absenteeism from the previous months. Prediction must be implemented without information about bidding results (work schedule in Figure 2).

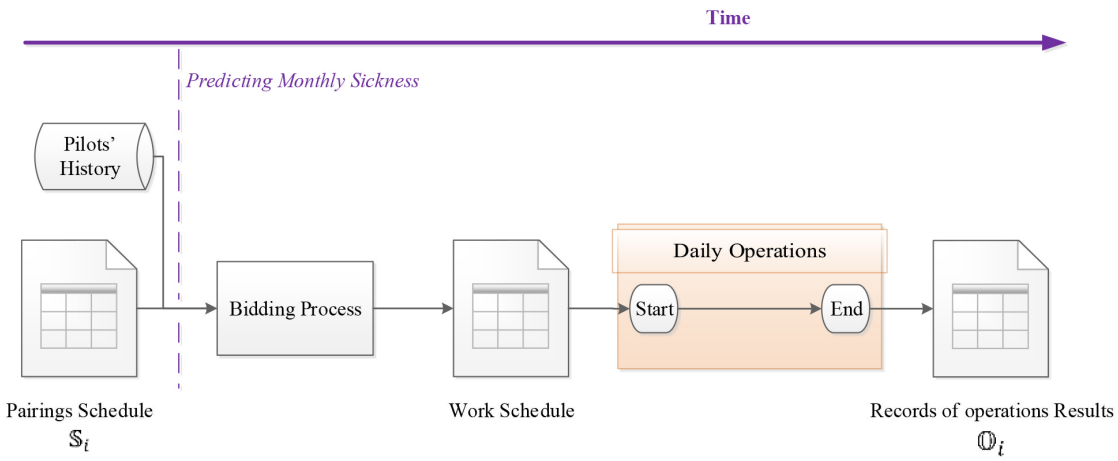

Figure 2: Simplified airline process from scheduling to the end of monthly operations. We predict monthly absenteeism after publishing a pairing schedule and before the bidding process.

Let $\mathbb{S}_i$ denote the table of pairings schedule for month i. Each record in this table is a pairing with its characteristics or attributes. In the pairing schedule table, as shown in Figure 2, there is no information about the pilots that may operate a pairing.

At the end of each month, the results of the operations are published in another table that is shown with $\mathbb{D}_i$ in Figure 2. This table describes how each scheduled pairing has been operated. Suppose that $\mathbb{D}_i$ has K_i pairings $1, 2, \ldots, K_i$ and s_{ij} indicates the scheduled flying minutes in which the assigned pilot has been absent for the month i and the pairing j. Therefore, total absenteeism in month i is the sum of pairing absenteeism over all of its pairings,

$$s_i = \sum_{j=1}^{K_i} s_{ij}. \tag{1}$$

Our objective is to present a systematic method for predicting the total value shown in (1) for a published pairing schedule and absence history.

3 Methodology

Two main databases are available for predicting absenteeism in a new month. Let n merged tables be available, $\mathbb{D}_1, \mathbb{D}_2, \ldots, \mathbb{D}_n$. After publishing a new month pairing schedule ($\mathbb{S}_{n+1}$), a suitable method for prediction should predict the total absence hours for this new month, say s_{n+1}.

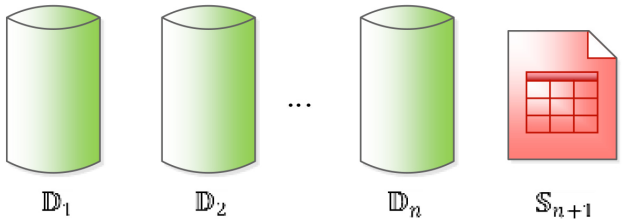

Figure 3: Available datasets for predicting new month absenteeism. All the previous databases ($\mathbb{D}_1, \mathbb{D}_2, \ldots, \mathbb{D}_n$) and the schedule of new month, ($\mathbb{S}_{n+1}$), can be used in a prediction.

In each of the datasets $\mathbb{D}_1, \mathbb{D}_2, \ldots, \mathbb{D}_n$, the response variable is defined as the *absent indicator*. This variable is denoted by y_{ij} which indicates the absenteeism in pairing j of the month i and is a binary variable

$$y_{ij} = \begin{cases} 1, & \text{if } s_{ij} > 0 \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

Let $\mathfrak{T}$ be the complete decision tree obtained by using all the datasets $\mathbb{D}_1, \mathbb{D}_2, \ldots, \mathbb{D}_n$, where the absent indicator is the response variable and the pairing characteristics are the explanatory variables. If $\mathfrak{T}$ has m terminal nodes which represent m regions on the space of pairing characteristics as $r_1, r_2, \ldots, r_m$ and the assigned probability of absenteeism in each region is $p_1, p_2, \ldots, p_m$; then the estimation of absenteeism for the new month can be calculated as

$$\hat{s}\left(\mathfrak{T}, \mathbb{S}_{n+1}\right) = \sum_{k=1}^{m} p_k c_k, \tag{3}$$

where c_k is the sum of total credits in kth region of the pairing schedule of the new month.

Equation (3) is our proposition for calculating the estimation of monthly absenteeism based on a decision tree and a monthly-published pairing schedule. The next step is then to prune the tree for an accurate prediction.

Figure 4 shows an example of a tree and the absenteeism estimations in each level of the tree for a fixed monthly pairing schedule. Figure 4 confirms that the differences between prediction values are large, with a range of more than 500 hours.

To solve this problem, we propose a learning process in which the algorithm of tree growing and tree pruning will be explained.

3.1 Tree growing method

Our objective is to provide a decision support system for predicting a new month's absence hours ($\hat{s}_{n+1}$) based on the new month pairing schedule ($\mathbb{S}_{n+1}$), pairing characteristics, and absenteeism history ($\mathbb{D}_1, \mathbb{D}_2, \ldots, \mathbb{D}_n$). The proposed learning process suggests using a loop for choosing the best decision trees.

Our first attempt in learning is to provide a stable decision tree. Unlike random forests, we do not create a random tree by resampling the data set. Suppose $\mathbb{D}_1, \mathbb{D}_2, \ldots, \mathbb{D}_a$ is the data of month 1 until month a, we merge these datasets into a unique set, denoted Γ_a. In the first step of the loop, a decision tree is made for Γ_a and the predictions of absence hours are calculated for each level of this tree by using Equation (3) and $\mathbb{S}_{a+1}$ as the input. Then the tree is pruned at the level that gives the minimum *level error* for the month $a+1$. The pruned tree is called $\widetilde{\mathfrak{T}}_a$.

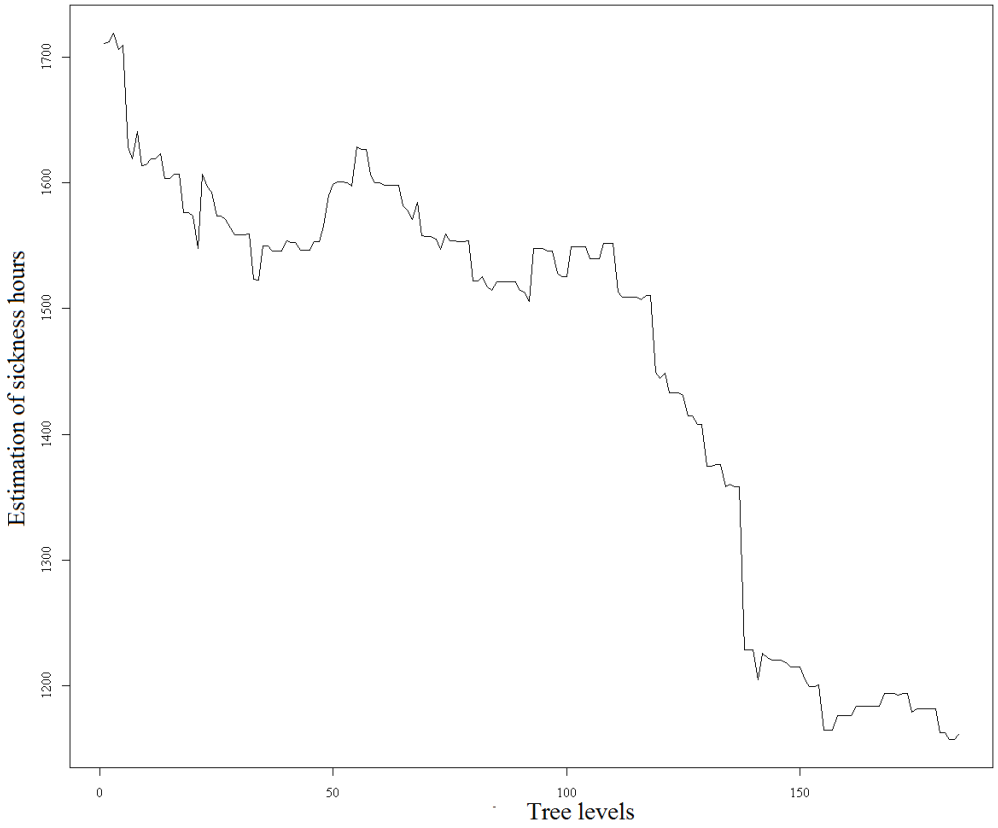

Figure 4: Estimation of absence hours in different levels of the tree, see Equation (3). Non-monotone values in different levels of the tree with a wide range of variability make the pruning a non-standard problem.

This loop will continue monthly to obtain $(n-a)$ pruned trees $\widetilde{\mathfrak{T}}_a, \widetilde{\mathfrak{T}}_{a+1}, \ldots, \widetilde{\mathfrak{T}}_{a+n-1}$.

The experts of a company consider a tree stable when its associated rules do not change considerably and they are acceptable. We fix a which is the number of months that must be used for making such a decision tree. Let Γ_a be the dataset obtained by merging $\mathbb{D}_1, \mathbb{D}_2, \ldots, \mathbb{D}_a$.

In the airline industry, we propose setting a equal to 12 to have a complete year's history when starting the construction of the tree. The reason for this choice is that the explanatory variables, pairing characteristics that are extracted from pairing schedule, are different in high and low seasons and in the airlines' flight schedules are planned at the tactics level for one year. This is why we suggest a 12-month dataset sequence as the preliminary training set.

By using Γ_a as the training dataset, we grow a decision tree. The Gini index is used as the splitting measure and the stopping rule is applied when either the *cp* is less than a pre-defined threshold or a node is split into child nodes with a population that is less than a pre-defined percent of the number of pairings in Γ_a. The *cp* criteria helps to keep the decision tree at a level where the variance of a terminal node is acceptable and the minimum population criteria avoids the creation of terminal nodes with rare cases.

Another parameter used for growing the tree is the *loss matrix*. In binary cases, a loss matrix is a 2×2 matrix with zero on the diagonal elements and the misclassification cost rates on the off-diagonal elements. In our problem, the cost of misclassifying an absent pairing as non-absent is equal to an odds ratio of not being absent. This means that if the cost of misclassifying a non-absent pairing as absent is 1, we consider the cost of misclassifying an absent pairing as non-absent equal to $\frac{1-p}{p}$, where p is the proportion of absent

pairings in Γ_a. We can write the loss matrix as the following

$$\begin{bmatrix} 0 & 1 \\ \frac{1-p}{p} & 0 \end{bmatrix}. \tag{4}$$

We use a loss matrix because absenteeism is a rare event and a decision tree requires adjustment for the prediction of unfair events.

Figure 5 represents the whole process in a flowchart. We repeat this process month by month to obtain $n-a-1$ trees, say $\mathfrak{T}_a, \mathfrak{T}_{a+1}, \ldots, \mathfrak{T}_{n-1}$. Each of these trees is based on the merged datasets from the first available month up to its related month. For example $\mathfrak{T}_{a+1}$ is the decision tree that is obtained from the explained growing method by using Γ_{a+1}, the merge of datasets $\mathbb{D}_1, \mathbb{D}_2, \ldots, \mathbb{D}_{a+1}$, as the train dataset.

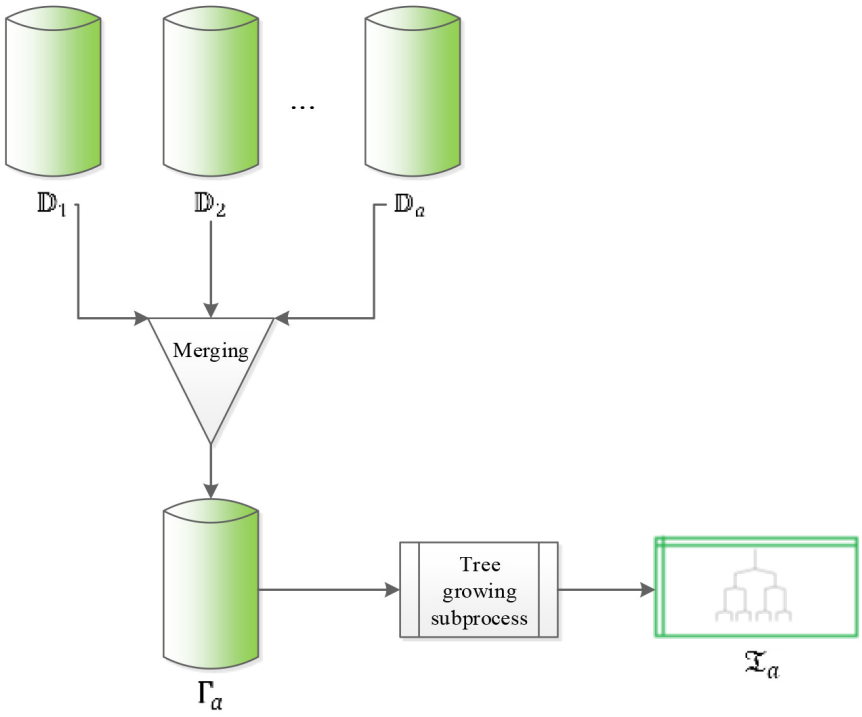

Figure 5: Tree growing process. The tree growing sub-process is applied to the result of merging all of the databases $(\mathbb{D}_{i1}, \mathbb{D}_{i2}, \ldots, \mathbb{D}_{i\ a})$ to create the original decision tree $\mathfrak{T}_{i\ a}$.

3.2 Tree pruning method

In the building step, we obtain $n-a-1$ trees, $\mathfrak{T}_a, \mathfrak{T}_{a+1}, \ldots, \mathfrak{T}_{n-1}$, each of them describes the association rules that increase or decrease the absenteeism in proportion to its terminal nodes. We want to use these rules as the predictor of future absenteeism. We prune each of these trees at the level that gives it the best prediction for the following month. In this way, we obtain the best possible scenarios based on the pairing characteristics and on the absenteeism history and we apply these scenarios to the future.

Let us start with $\mathfrak{T}_a$ and suppose that it has m levels. We can obtain m different trees by pruning the original tree with respect to each of the *cp* values. We denote the pruned tree at level k as $\langle \mathfrak{T}_a \rangle_k$.

By using Equation (3) and the pairing schedule of the following month, $\mathbb{S}_{a+1}$, it is possible to calculate the prediction of absenteeism for the following month at each level of the tree. The actual value of absenteeism in month $a+1$, s_{a+1}, can be calculated from Equation (1). Moreover, the *level error* representing the error

of prediction at level k of the original decision tree $\mathfrak{T}_a$ is

$$e_{ak} = \begin{cases} \hat{s}\left(\langle \mathfrak{T}_a \rangle_k, \mathbb{S}_{a+1}\right) - s_{a+1}, & \text{in the case of over-prediction} \\ \eta\left\{s_{a+1} - \hat{s}\left(\langle \mathfrak{T}_a \rangle_k, \mathbb{S}_{a+1}\right)\right\}, & \text{in the case of under-prediction} \end{cases} \tag{5}$$

where η is the proportion of under-prediction cost on over-prediction cost. We consider this ratio because in our case the cost of under-prediction is higher than the cost of over-prediction.

The *error* of the tree is defined as the minimum of level errors,

$$e_a = \min_{1 \le k \le m} e_{ak} \ , \tag{6}$$

where m is the number of levels of $\mathfrak{T}_a$. We prune the decision tree $\mathfrak{T}_a$ at the level where its level error is equal to e_a and denote it $\widetilde{\mathfrak{T}}_a$. This procedure is applied for all $n-a-1$ trees and the obtained trees, $\widetilde{\mathfrak{T}}_a, \widetilde{\mathfrak{T}}_{a+1}, \ldots, \widetilde{\mathfrak{T}}_{n-1}$, are used for predicting a new month of absenteeism s_{n+1}.

The following pseudo code provides the details of the prediction procedure.

1. Choose a,
2. Consider an empty set as the tree set,
3. For z from a to $n-1$ do:
   - 3.1. Merge $\mathbb{D}_1, \mathbb{D}_2, \ldots, \mathbb{D}_z$,
   - 3.2. Grow a decision tree for the merged dataset,
   - 3.3. Calculate the prediction for the next month by using $\mathbb{S}_{z+1}$ for each level of the obtained tree,
   - 3.4. Calculate the level error,
   - 3.5. Prune the tree at the level having the minimum level error,
   - 3.6. Call the pruned tree $\widetilde{\mathfrak{T}}_z$,
   - 3.7. Add $\widetilde{\mathfrak{T}}_z$ to the tree set.

3.3 Prediction of absenteeism

For predicting absenteeism in a new month, we use the pairing schedule table for the new month, $\mathbb{S}_{n+1}$, and all the pruned trees, $\widetilde{\mathfrak{T}}_a, \widetilde{\mathfrak{T}}_{a+1}, \ldots, \widetilde{\mathfrak{T}}_{n-1}$. This collection of trees explains the best possible prediction for the following month's absenteeism. If we use the $\mathbb{S}_{n+1}$ table as the input of these trees, they give $n-a-1$ different values for the new month's prediction, $\hat{s}\left(\widetilde{\mathfrak{T}}_a, \mathbb{S}_{n+1}\right)$, $\hat{s}\left(\widetilde{\mathfrak{T}}_{a+1}, \mathbb{S}_{n+1}\right)$, $\ldots$, $\hat{s}\left(\widetilde{\mathfrak{T}}_{n-1}, \mathbb{S}_{n+1}\right)$. Each of these values is based on the rules that best predict the previous month's absenteeism. Therefore, we replicate the possibility of the occurrence of previous scenarios in the future to make a virtual evaluation set. Figure 6 represents the procedure for calculating individual estimations.

In the next step, we weigh the individual predictions to provide a prediction for a new, unobserved month.

For the new month's prediction of absenteeism, the only available information is the pairing schedule. Thus, we propose the *similarity* of the new month's schedule with the previous month's schedules. We use intuition that it is more likely to have the same scenario in the months with a more similar pairing schedule.

To calculate the similarity, first we determine the important variables that characterize a pairing schedule. These variables are the pairing characteristics such as monthly total time, total credits, day credits, night credits, deadhead credits, weekend credits, etc. If there are z variables, we use these variables to properly weigh each prediction tree. If $\nu_{i\ j}$ denotes the value of variable j of the pairing characteristics in month i, the similarity vector, $\mathbf{w}_{n+1} = (w_{n+1\ a+1}, w_{n+1\ a+2}, \ldots, w_{n+1\ n})$, can be written as

$$w_{n+1\ l} = \sqrt{\sum_{j=1}^{z} \left(\nu_{n+1\ j} - \nu_{l\ j}\right)^2}. \tag{7}$$

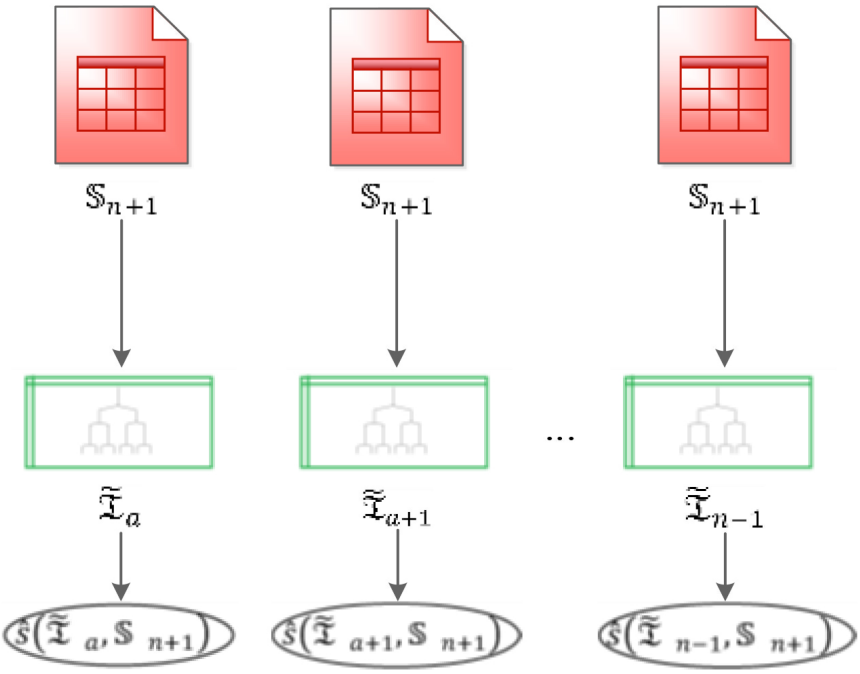

Figure 6: Prediction for the new month is based on the pruned trees. The pairing schedule of the new month is the input of each available pruned tree for calculating possible values of prediction.

For predicting absence hours in the new month, we use the prediction of trees obtained from $\mathbb{S}_{n+1}$, $\hat{s}\left(\widetilde{\mathfrak{T}}_a, \mathbb{S}_{n+1}\right)$, $\hat{s}\left(\widetilde{\mathfrak{T}}_{a+1}, \mathbb{S}_{n+1}\right)$, $\ldots$, $\hat{s}\left(\widetilde{\mathfrak{T}}_{n-1}, \mathbb{S}_{n+1}\right)$, the similarity vector, $\mathbf{w}_{n+1}$, and the errors of the trees, e_a, e_{a+1}, $\ldots$, e_{n-1}. Motivated by Kernel regression method (Watson, 1964), the suggested prediction is the *weighted mean of tree based predictions*, i.e.

$$\hat{s}_{i\ n+1} = \frac{1}{2}\hat{s}_{w_{i\ n+1}} + \frac{1}{2}\hat{s}_{e_{i\ n+1}}, \tag{8}$$

where

$$\hat{s}_{w_{i\ n+1}} = \frac{1}{\sum_{k=a}^{n-1} w_{n+1\ k+1}} \sum_{k=a}^{n-1} w_{n+1\ k+1} \times \hat{s}\left(\widetilde{\mathfrak{T}}_k, \mathbb{S}_{n+1}\right), \tag{9}$$

and

$$\hat{s}_{e_{i\ n+1}} = \frac{1}{\sum_{k=a}^{n-1} \left(match1/\sqrt{e_k}\right)} \sum_{k=a}^{n-1} \frac{1}{\sqrt{e_k}} \times \hat{s}\left(\widetilde{\mathfrak{T}}_k, \mathbb{S}_{n+1}\right). \tag{10}$$

Equation (8) consists of two parts and each part is a weighted mean of $\left\{\hat{s}\left(\widetilde{\mathfrak{T}}_k, \mathbb{S}_{n+1}\right),\ a \leq k \leq n-1\right\}$. The weights in the first part of this equation are a similarity vector. We mathematically formulize the intuition that more similarity between two pairing schedules must have more prediction weight. The weights on the second part are a decreasing function of the errors of the trees. We use these errors as the weight for decreasing the effect of outliers on the prediction. When the prediction of a decision tree is better in the learning process, it gets more weight in the calculation of the new month's prediction.

In the next section, we report the results of implementing this methodology in practice.

4 Implementation

The methodology described has been developed to be applied in an airline company as a decision support system for predicting the absenteeism of pilots. To test the decision support system and verify the quality of the prediction, we used 3 years of monthly pairing schedules. The final table contains 13 different pilot positions and 382,202 pairings for a total of 36 months. Over 3 years, 379,129 of pilots' flight hours were replaced because of absenteeism. This means 7 percent of the total flight hours had absent pilots.

The following two plots (Figure 7 and Figure 8) demonstrate the ideas from the previous section. We search in the space of all variables for relatively frequent regions in terms of the amount of pairings in which absenteeism occurs more frequently than in other, similar regions.

Figure 7 compares 3 pairing characteristics (day credits, night credits and deadhead credits) between absent and non-absent pairings in one position.

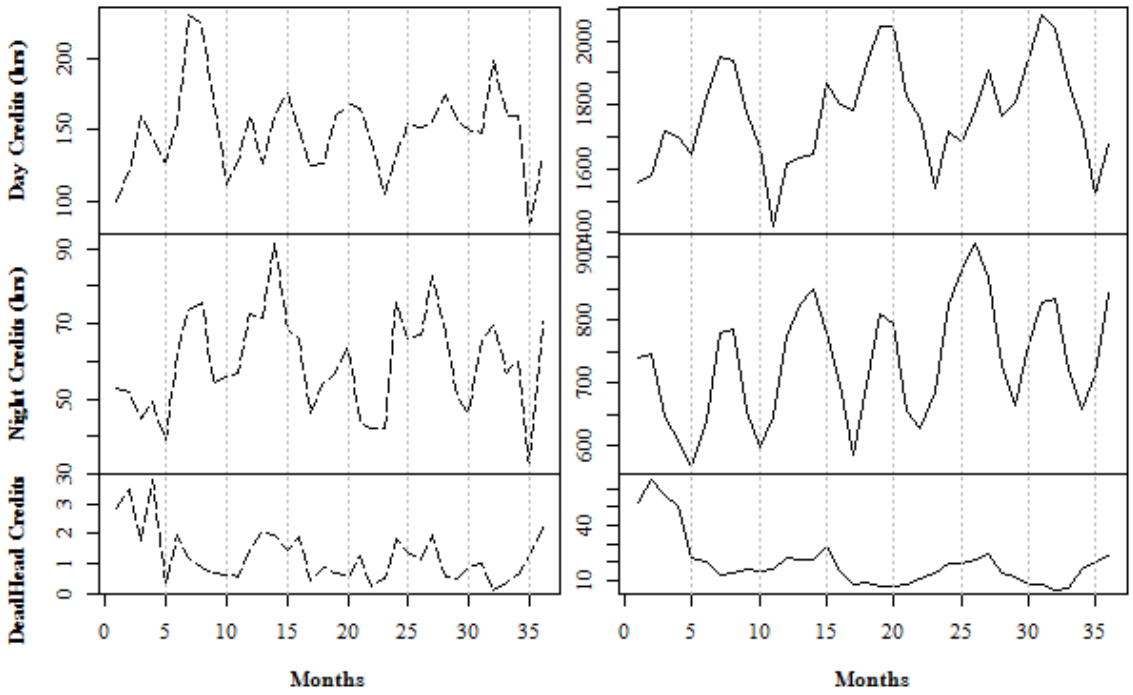

Figure 7: Comparison between absent and non-absent pairings.

The mass plot (Figure 8) presents a four-dimensional representation of absence events, the number of pairings and two variables influenced for prediction (total time and total credit). In Figure 8, the space of total time and total credits is painted in colors ranging from yellow to red. Yellow pixels show the less frequent regions in terms of the number of pairings, while red pixels are the most frequent regions. The dark points show the absence events.

To test the quality of the fit of the proposed methodology and its ability to provide predictions, we applied the procedure to all of the positions and all of the months of 2012. This was the first pre-test of the model before applying it in practice.

In Figure 7, the left and right panel plots show the monthly variation of different variables among the absent and non-absent pairings, respectively. The variables from top to down are day credits, night credits and deadhead credits. The different patterns in the absent plots (left) in comparison with the non-absent plots (right) makes these variables good candidates for the model covariates.

The colors in Figure 8 represent the density of pairings in the space of total time and total credits. About 19 percent of the pairings are scheduled in the yellow region, 22 percent in the gold region, and 59 percent

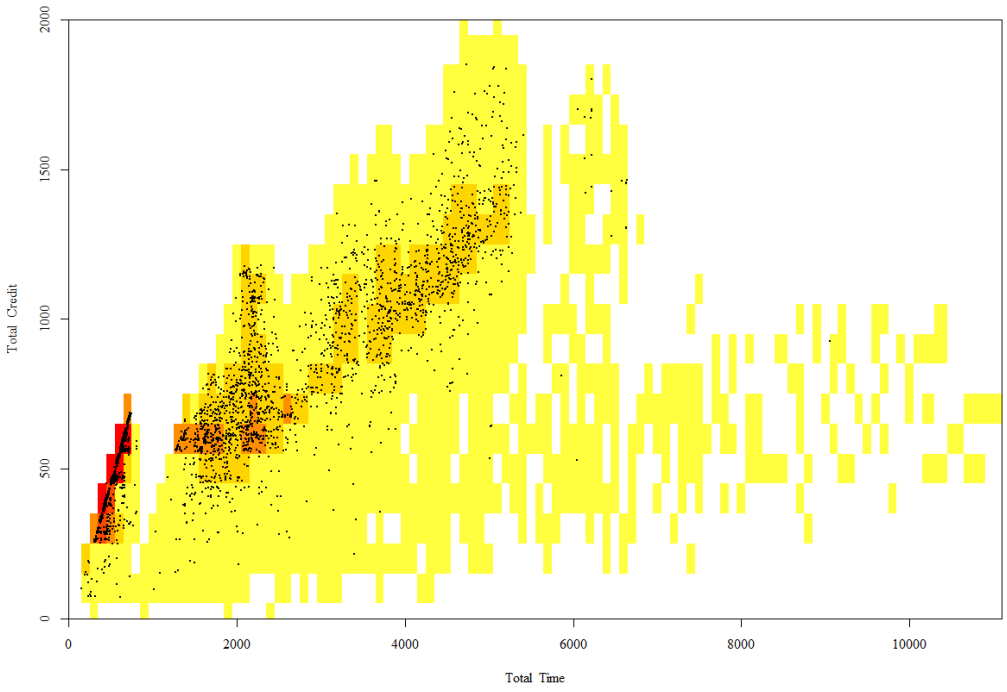

Figure 8: Mass plot for absenteeism, comparing total time against total credits.

of the pairings are scheduled in the small region with dark orange, orange red or red colors, respectively. The dark points show absent pairings. The distribution of absent pairings suggests different patterns of absenteeism for these two variables.

Figure 9 plots the prediction versus observation for absence hours in Position 1. It can be observed that the model is able to determine the general trend of absenteeism. Absenteeism in November 2012 in Position 1 is much lower than in the other months. This outlier is an exception among the three years of data used in this study.

The last examination of our proposed methodology was a comparison of predictions with the current method applied in the airline company. This comparison has been made for 11 positions. The results of the comparison show that in 6 positions, the proposed method performs better than the model that is already being used in this industry. These 6 positions cover 84 percent of annual flight hours of the airline company. The other 5 positions are the small positions with less than 100 pilots in each position.

Table 1 indicates the error of the prediction for both models in hours for a different cost rate of under prediction (1, 1.25, 1.5, 1.75, and 2). A cost rate of under prediction equals to 2 means that we multiply the prediction error by 2 in the case of under prediction error.

If we consider that the cost rate of under prediction equals 1, i.e. the same cost for under and over prediction error, the proposed method improved the predictions by 13 percent and it has 1749 error hours in comparison with the current method. However, the cost of under prediction in airline companies is higher than the cost of over prediction because of the expensive costs in the case of flight cancelations and expenses of calling a non-scheduled pilot to do a pairing. Therefore, one can deduce that the proposed methodology improves the prediction by at least 13 percent.

As an applied data mining study, this methodology and its procedures must be executed in an automatic way and be able to help managers make better decisions. Monthly absenteeism predictions are used at the operations level in the airline industry and therefore the proposed method must be automated like an operations level application.

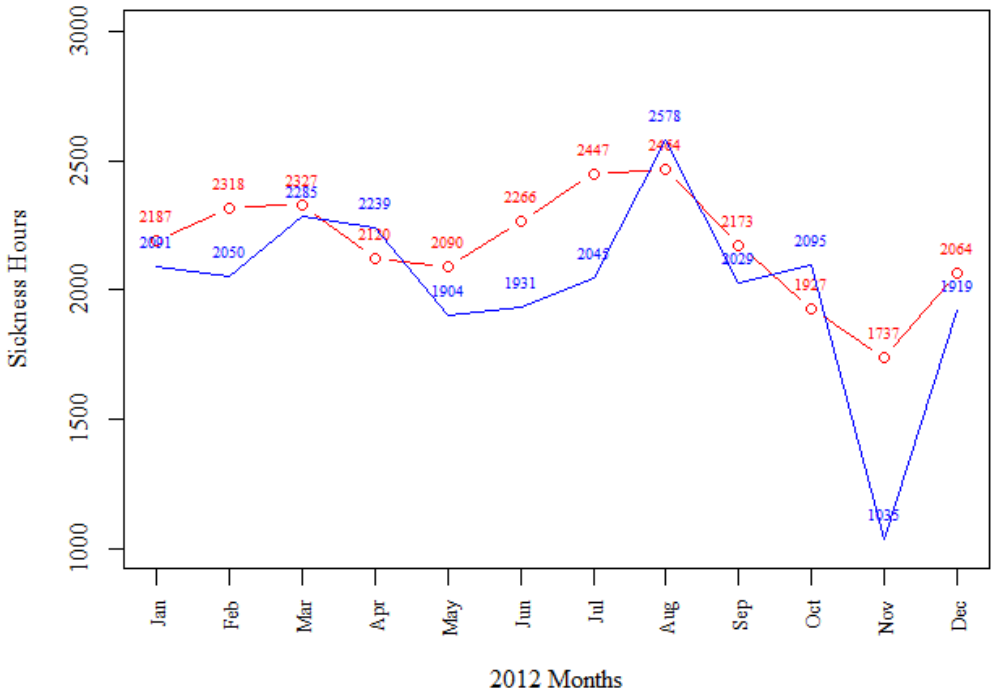

Figure 9: Predictions versus Observation for Position 1 in 2012. The blue line is the actual absenteeism and the red line is the absence prediction.

Table 1: Comparison of annual prediction error (in hours) between the old model and new proposed model, for some positions in 2012.

	Current Model					**New Model**				
Positions	1	1.25	1.5	1.75	2	1	1.25	1.5	1.75	2
1	3277	3695	4113	4530	4948	2652	2741	2829	2917	3006
2	2890	4732	5574	6416	7258	3461	4081	4702	5323	5944
7	1237	1487	1737	1987	2237	1039	1110	1181	1252	1323
8	1519	1822	2125	2429	2732	1332	1501	1669	1827	2006
9	1605	1841	2078	2314	2550	1356	1499	1642	1786	1929
10	1576	1913	2249	2586	2923	1515	1814	2113	2412	2711
Total	13104	15490	17876	20263	22649	11355	12746	14137	15528	16918
				Improvement in Prediction		**13%**	**18%**	**21%**	**23%**	**25%**

This goal leads us to make a decision support system. We created a user-friendly web application as the decision support system for monthly absenteeism predictions. The codes have been written in R (R Development Core Team, 2005) programming language by using *shiny* (RStudio and Inc., 2013), *ggplot2* (Wickham, 2009) and *rpart* (Therneau et al. 2010) packages.

Conclusion

Unwanted events are always a big challenge for airline companies and having good predictions is a helpful tool for disruption management during operations. Absenteeism of the pilots is one of those unwanted events that may cause a flight delay, flight cancelation, customer dissatisfaction, etc., besides the costs of substituting a reserve pilot. For these reasons, determining the number of reserve pilots with a minimum error is an important issue. Absenteeism of the pilots is one of the most compromising examples of absenteeism among the cabin crew, hence we focused on the prediction of monthly absence hours of the pilots.

Our method for dealing with this prediction problem was considering monthly absence hours as the sum of the pairing hours in which the pilot is absent. In this approach, we could relate the response variable to the pairing characteristics as the explanatory variables. This means the prediction is based on the monthly schedule and will change as the schedule changes. The proposed iterative algorithm determines the best possible scenarios of previous months and predicts future monthly absenteeism as the weighted mean of the results of applying these scenarios to the new schedule.

Results of applying this method to real data show that in most cases the predictions have an acceptable error and the proposed methodology improved the monthly predictions of the absenteeism by at least 13 percent in 2012 compared with the current method of prediction by the airline. This means that for this period, and the entire year of 2012, the error of prediction of our model was 1749 hours less than the error of prediction in the current model used by the airline.

The decision support system that is created for this methodology makes its use considerably easy and completely automatic. It provides extra information for helping the managers make the best possible decisions in a user-friendly manner. The algorithm is able to determine outliers automatically and decrease the effect of the outliers on future predictions. The proposed decision support system is the main contribution of this research in the applied area and it is the first attempt in the airline industry to use it, to our knowledge.

Although the proposed model has been explained and its advantages and improvements have been proven, like any other methodology it has limits and restrictions. The predictions may not be accurate in the presence of a lot of outliers. Also, our experience indicates that this technique is inefficient in small positions with a large variation in the monthly absence hours as a result of the lack of useful information.

This field of research can be developed in various regards. One of the current problems is the daily prediction of absence hours. This could be accomplished by considering an awarded pairing table and the time that we know which pilot is assigned to which pairing.

Among the different trees that exist for each month, some of them provide better predictions. Model selection methods could be applied in order to select the best set of trees for each month. This can be done by developing the similarity vector discussed in Section 3.

Focusing on the outliers and developing a model for predicting outliers is another possible improvement for the current technique.

References

Abdelghany, A., Ekollu, G., Narasimhan, R., Abdelghany, K. (2004). A proactive crew recovery decision support tool for commercial airlines during irregular operations. Annals of Operations Research, 127(1–4), 309–331.

Barnhart, C., Belobaba, P., Odoni, A.R., Barnhart, C. (2003). Applications of Operations Research in the Air Transport Industry. Transportation Science, 37(4), 368–391.

Bohanec, M., Bratko, I. (1994). Trading accuracy for simplicity in decision trees. Machine Learning, 15(3), 223–250.

Bazargan, M. (2010). Airline operations and scheduling (2nd ed.). Farnham, Surrey, England: Ashgate.

Bratu, S., Barnhart, C. (2006). Flight operations recovery: New approaches considering passenger recovery. Journal of Scheduling, 9(3), 279–298.

Breiman, L. (1993). Classification and regression trees. Belmont, California: CRC Press.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. (1984). Classification and regression trees. Wadsworth & Brooks. Monterey, CA: Wadsworth International Group.

Burdun, I.Y., Parfentyev, O.M. (1999). Fuzzy situational tree-networks for intelligent flight support. Engineering Applications of Artificial Intelligence, 12(4), 523–541.

Cauvin, A., Ferrarini, A., Tranvouez, E. (2009). Disruption management in distributed enterprises: A multi-agent modelling and simulation of cooperative recovery behaviours. International Journal of Production Economics, 122(1), 429–439.

Ciampi, A., Chang, C.-H., Hogg, S., McKinney, S. (1987). Recursive partition: A versatile method for exploratory data analysis in biostatistics. Biostatistics, 38, 23–50.

Clausen, J., Larsen, A., Larsen, J., and Rezanova, N.J. (2010). Disruption management in the airline industry?concepts, models and methods. Computers & Operations Research, 37(5), 809?-821.

Dillon, J.E., Kontogiorgis, S. (1999). US Airways optimizes the scheduling of reserve flight crews. Interfaces, 29(5), 123–131.

Dupuis, C., Gamache, M., Pagé, J.-F. (2012). Logical analysis of data for estimating passenger show rates at Air Canada. Journal of Air Transport Management, 18(1), 78–81.

Esposito, F., Malerba, D., Semeraro, G., Kay, J. (1997). A comparative analysis of methods for pruning decision trees. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 19(5), 476–491.

Gaballa, A. (1979). Planning callout reserves for aircraft delays. Interfaces, 9(2-Part-2), 78–86.

Gamache, M., Soumis, F., Villeneuve, D., Desrosiers, J., Gélinas, E. (1998). The Preferential Bidding System at Air Canada. Transportation Science, 32(3), 246–255.

Grosche, T. (2009). Computational intelligence in integrated airline scheduling. Berlin, Germany: Springer-Verlag.

Hastie, T., Tibshirani, R., Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York, NY, US: Springer.

Jarrah, A.I., Yu, G., Krishnamurthy, N., Rakshit, A. (1993). A decision support framework for airline flight cancellations and delays. Transportation Science, 27(3), 266–280.

Karabadji, N.E.I., Seridi, H., Khelf, I., Azizi, N., Boulkroune, R. (2014). Improved decision tree construction based on attribute selection and data sampling for fault diagnosis in rotating machines. Engineering Applications of Artificial Intelligence, 35, 71–83.

Kohl, N., Larsen, A., Larsen, J., Ross, A., Tiourine, S. (2007). Airline disruption management—perspectives, experiences and outlook. Journal of Air Transport Management, 13(3), 149–162.

Lettovský, L., Johnson, E.L., and Nemhauser, G.L. (2000). Airline crew recovery. Transportation Science, 34(4), 337-?348.

Li, X., Chan, C.W. (2010). Application of an enhanced decision tree learning approach for prediction of petroleum production. Engineering Applications of Artificial Intelligence, 23(1), 102–109.

Morgan, J.N., Sonquist, J.A. (1963). Problems in the analysis of survey data, and a proposal. Journal of the American Statistical Association, 58(302), 415–434.

Quinlan, J.R. (1993). C4.5: programs for machine learning. San Mateo, CA: Morgan Kaufmann.

R Development Core Team. (2005). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/.

Rakshit, A., Krishnamurthy, N., Yu, G. (1996). System operations advisor: A real-time decision support system for managing airline operations at united airlines. Interfaces, 26(2), 50–58.

Rokach, L., Maimon, O. (2005). Top-down induction of decision trees classifiers-a survey. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 35(4), 476–487.

RStudio and Inc. (2013). shiny: Web Application Framework for R. R package version 0.6.0 Retrieved from http://CRAN.R-project.org/package=shiny.

Sethi, I.K., Yoo, J.H. (1994). Design of multicategory multifeature split decision trees using perceptron learning. Pattern Recognition, 27(7), 939–947.

Sohoni, M.G., Johnson, E.L., Bailey, T.G. (2006). Operational airline reserve crew planning. Journal of Scheduling, 9(3), 203–221.

Suryadevara, N.K., Mukhopadhyay, S.C., Wang, R., Rayudu, R.K. (2013). Forecasting the behavior of an elderly using wireless sensors data in a smart home. Engineering Applications of Artificial Intelligence, 26(10), 2641–2652.

Therneau, T.M., Atkinson, B., Ripley, B. (2010). rpart: Recursive partitioning. R package version 4.1-1. Retrieved from http://CRAN.R-project.org/package=rpart.

Pham, T.-T., Luo, J., Hong, T.-P., Vo, B. (2014). An efficient method for mining non-redundant sequential rules using attributed prefix-trees. Engineering Applications of Artificial Intelligence, 32, 88–99.

Verma, A., Wei, X., Kusiak, A. (2013). Predicting the total suspended solids in wastewater: A data-mining approach. Engineering Applications of Artificial Intelligence, 26(4), 1366–1372.

Watson, G.S. (1964). Smooth regression analysis. Sankhyâ: The Indian Journal of Statistics, Series A, 26(4), 359–372.

Wei, G., Yu, G., Song, M. (1997). Optimization model and algorithm for crew management during airline irregular operations. Journal of Combinatorial Optimization, 1(3), 305–321.

Wickham, H. (2009). ggplot2: elegant graphics for data analysis. New York, NY, US: Springer.

Witten, I., Frank, E., Hall, M. (2011). Data Mining: Practical Machine Learning Tools and Techniques. US: Morgan Kaufmann.