ANTICIPATING A REGIME CHANGE IN THE DAILY SHARE OF DELAYED AND CANCELED FLIGHTS AT SÃO PAULO/GUARULHOS INTERNATIONAL AIRPORT

Rosana Batista Teixeira
Instituto Tecnológico de Aeronáutica
Universidade Federal de São Paulo - Campus São José dos Campos
Rodrigo Arnaldo Scarpel
Instituto Tecnológico de Aeronáutica

ABSTRACT
Flight delays and cancellations are frequent occurrences at most airports around the world. In Brazil, the liberalization of air transport has caused flight concentration at some airports, increasing the occurrence of delays and cancellations due to congestion. In Brazil, São Paulo/Guarulhos International Airport is one of the most affected by delays. The objective of this work is to anticipate the occurrence of congested days at São Paulo/Guarulhos International Airport. The accuracy of the prediction model in anticipating the regime change in the daily share of delayed and cancelled flights one period ahead was considered satisfactory.

1. INTRODUCTION

Congested days at airports due to flight delays and cancellations are a universal problem. Delays are recurrent and increasingly common in passenger routines, especially at the hubs airports. The excessive concentration at a hub can result in congestion delays, as the number of aircraft tends to approach the maximum capacity of the hub airport. Delays, which occur as a result of this congestion, increase operating costs of the airlines, the passengers travel time and create additional work for air traffic controllers, which increases their stress levels (Wensveen, 2016).

The analysis of air delays are important due to the understanding of their potential causes enable the development of possible solutions to support the performance of air transport system (Abdel-Aty *et al.*, 2007; Rebollo and Balakrishnan, 2014; Scarpel and Pelicioni, 2018). Concerned about Brazil, the liberalization of air transport, led to a concentration of flights in a few hubs *(Costa et al*., 2010). The São Paulo/Guarulhos International Airport, as of now, is the

largest Brazilian hub (Scarpel and Pecicioni, 2018). Therefore, anticipating the occurrence of congested days at São Paulo/Guarulhos International Airport is an important issue for the development of strategies to reduce delays and cancellations flights, and support planning.

The goal of this work is to develop a predictive model to anticipate the occurrence of congestion days at São Paulo/Guarulhos International Airport. Then, homogeneous groups – regimes – were identified within the time series represented by the daily share of delayed and cancelled flights at São Paulo/Guarulhos airport. Based on the identified regimes, a classification model was created. Its results was compared by two different algorithms to predict in advance the occurrence of days with a high concentration of delayed and cancelled flights, hence being a tool to support planning and decision-making. This work intends to contribute to the literature proposing an approach that, sequentially, detect change points in the time series of the daily share of delayed and canceled flights at an airport and to create models to anticipate the occurrence of congested days.

2. BACKGROUND

Delays and flights cancellations have been a subject of a succession of studies. According to Xiong and Hansen (2013), the air system faces big challenges dealing with high demand when system capacity is reduced. In face of delays, airline schedules are often subject to unexpected changes, as some flights are delayed due to the late arrival of a previous flight. These delays can propagate due to tight airline schedules (Abdel-Aty et al., 2007). Conform Jacquillat and Odoni (2015), airport congestion has increased as a result of the growth of air traffic and limitations in capacity at the busiest airports. According to Xiong and Hansen (2013), when airports are scheduled close or above the maximum capacity, a capacity drop will often result in a demand-capacity imbalance, creating disruptions and delays. Furthermore, this problem is worsened by the competitiveness of airlines, which confronted by the high costs of aircraft, aim high utilization rates.

Based on Santos *et al.* (2018), flights delays and cancellations are the main problems associated with the operation disruptions of an air transport network. As stated by Janic (2015), an air transport network consists of airports and routes operated by the airlines. According to the author, large-scale disruptions can compromise the operation of the network. Among them, failures of the transport network components, industrial actions of the transport staff and natural

disasters. Abdel-Aty *et al.* (2007 ascertained that the main reasons for increase of flight delays are the adverse weather around airports, the lack of runway capacity, the increase in the number of aircraft and poor air traffic control.

In data analysis, machine learning methods derived from complex models and elaborate algorithms can be used to perform predictive analysis. Santos and Robin (2010) studied flight delays at European airports using multiple regression analysis to identify the delay causes. Rebollo and Balakrishnan (2014) have employed clustering and classification approaches to prevent delays in times flights departure on a particular route or at a particular airport at some point in the future. Chandramouleeswaran *et al.* (2018) presented an approach to predicting delays at United States airports by considering two model types: neural network model and logistic regression model. They were based on temporal, network-level, and weather related features, in addition to the congestion of the network. Yu *et al.* (2019) used an unsupervised learning method combined with a supervised learning algorithm of regression and classification to perform flight delay prevention analyses.

In Brazil, Scarpel and Pecicioni (2018) employed a data analytics approach to build an early warning model to anticipate the occurrence of congested days at São Paulo/Guarulhos International Airport. The combination of modeling approaches relies on different assumptions and allowed to generate a more flexible model that made improvement in the prediction accuracy. Por uma concepção diferente, Bendinelli *et al.* (2016) examined if the lack of competition could be a source of increased rates of delay and cancellation of flights, a relationship which was confirmed by the authors.

In literature, models are created to predict the average delay in a day or considering information related to the daily share of delayed flights. In this type of model the observations are independent and don't identify associations between different days. In this study, delays in airport were analyzed through the identification of day patterns, dependently. The proposal was to identify regimes that could be characterized using a probability distribution, that is, mean and standard deviation. To create the model, the concept of change point detection was applied from the regime perspective, and to detect regimes was enforced the hidden Markov model.

3. LITERATURE REVIEW

3.1. Change Point Detection and Hidden Markov Models

Change Point Detection - CPD is the estimation of points in a time series with different statistical properties. The intervals between each detected point are called regimes or states. According to Killick and Eckley (2014), change point detection is the estimation of the point at which the statistical properties of a sequence of observations change. Considering observations in time t and $t+1$, if a point in time t belongs to a different group than an observation in time $t+1$, then a change point occurs between the two observations, as shown in Figure 1.

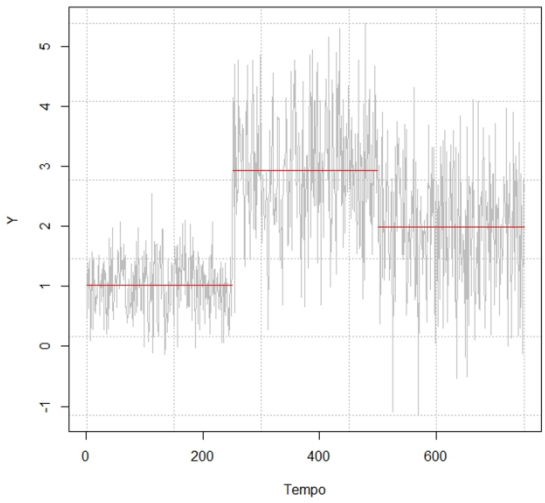

Figure 1: Change point in a time series

For CPD as a clustering problem, the observations placed between two change points, are sets of observations with the same statistical properties, called regimes. To handle the problem of CPD, seen as a clustering problem, the machine learning method, the Hidden Markov Model - HMM was applied.

Hidden Markov Models are models where the probability distribution that generates an observation depends on the state of an unobserved Markov process (Zucchini *et al.*, 2017). The authors describe a hidden Markov model O_t: $t \in \mathbb{N}$ as a particular kind of dependent mixture, in which the history sequence of the observations $O_{1:\,t}$ and the unobserved states $S_{1:\,t}$ are summarized as two process described as,

$$P(S_t|S_{1:(t-1)} = P(S_t|S_{t-1}),\ \ t = 2,3,\dots \quad (1)$$

$$P\left(O_t\middle|O_{1:(t-1)}, S_{1:t}\right) = P(O_t|S_t),\ \ t \in \mathbb{N} \quad (2)$$

where the Equation 1 represents an unobserved parameter process S_t: $t = 1, 2,\dots$ satisfying the Markov property. The Equation 2 represents the state-dependent process, in which the distribution of O_t: $t = 1, 2,\dots$ depends only on the current state and not on previous states or observations. If the MC (S_t) has m states, O_t is an m-state HMM. According to Visser (2011), hidden Markov models are defined as models with discrete states, characterized by their distribution function, and the evolution of the states over time is governed by a Markov process. Regarding the model selection criteria in an HMM with m states, Zucchini *et al.* (2017) state that the model is better adjusted increasing m as judged by the likelihood. However, the number of parameters becomes a quadratic increase, and the model can be very complex. Two known criterions to trade the improvement against this increase are called Akaike Information Criterion (AIC) e Bayesian Information Criterion (BIC), where the best fit model usually is the one with the lowest value of AIC and BIC, considering the increasing number of parameters.

3.2. Classification Models

When observed states are detected by the HMM, they can be seen as groups composed by their own attributes. Therefore, to select crucial variables and generate a predictive model a classification method is necessary. In order to do a comparative analysis, two methods are applied: Classification and Regression Tree – CART and Random Forests – RF. This work presents a short explanation of them.

3.2.1. Classification and Regression Tree

Classification and Regression Tree are conceptually simple and useful for interpretation and visualization, and can be used for regression or classification. The response variable is given by the mean of the training observations that belong to the same node. In contrast, for a classification tree, is predicted that each observation belongs to the most commonly occurring class of training observations in the region to which it belongs (James *et al.*, 2013).

According to Scarpel (2014), CART is attractive when the interpretability is an important issue, since they are designed to detect the important predictor variables and to generate a tree

structure to represent the identified partition. James *et al.* (2013), describes the regression tree algorithm in four steps:

1. Use recursive binary splitting, stopping only when each terminal node has fewer than some minimum number of observations;
2. Apply cost complexity pruning to the large tree, in order to obtain a sequence of best subtrees, as a function of a parameter α;
3. Use K-fold cross-validation to choose α. For each k = 1, …, K: a) Repeat Steps 1 and 2 on all but the kth fold of the training data; b) Evaluate the mean squared prediction error on the data left-out kth fold, as a function of α; Average the results for each value of α, and pick α to minimize the average error;
4. Return the subtree from step 2 that corresponds to the chosen value of α.

The classification tree extending is very similar. However, the residual sum of squares cannot be used as a criterion for making the binary splits. An alternative is the classification error rate by the measures Gini index or entropy.

In order to determine the ideal tree size, a usual method is to consider one-standard deviation rule. By this method one is advised to choose the smallest tree whose cross-validation relative error is close to the minimum cross-validation relative error plus one standard deviation (Scarpel, 2014).

3.2.2. Random Forests

Random Forests (RF) are a classification method composed of several decision trees that combine the bagging concept and the random variables selection of each partition. Breiman (2001) states that random forests consist of using randomly selected inputs or combinations of inputs at each node to grow each tree. Among its advantages are good accuracy, relatively robust to outliers and noise, useful internal estimates and variable importance. The author concludes that RF are an effective tool in prediction and injecting the right kind of randomness makes them accurate classifiers and regressors.

Generally, Kandhasamy e Balamurali (2015) describe the random forest tree growth as follow:

1. If the number of cases in the training set is N, sample N cases at random – but with replacement, from the original data. This sample will be the training set for growing the tree.
2. If there are M input variables, a random number of attributes are selected and the best split used to split the node. The value of M is held constant during the forest growing.
3. Each tree is grown to the largest extent possible. There is no pruning.

4. METHODOLOGY

Initially, the source and selection of the dataset, the treatment and integration to obtain the data used were described. Then, was made an explanation of the method used to group the data into regimes and, finally, the independent variables and the models used in the generation of the classifiers were presented.

The database used for the development of this study is known as Regular Active Flight - VRA, available on the website of the Brazilian National Civil Aviation Agency (ANAC). It consists of flight information that presents changes (delays, anticipations and cancellations), the time they occur and the justifications presented by the companies for such changes (ANAC 2019). The analyses were performed using the R software and the depmixS4, rpart, partykit and randomForest packages.

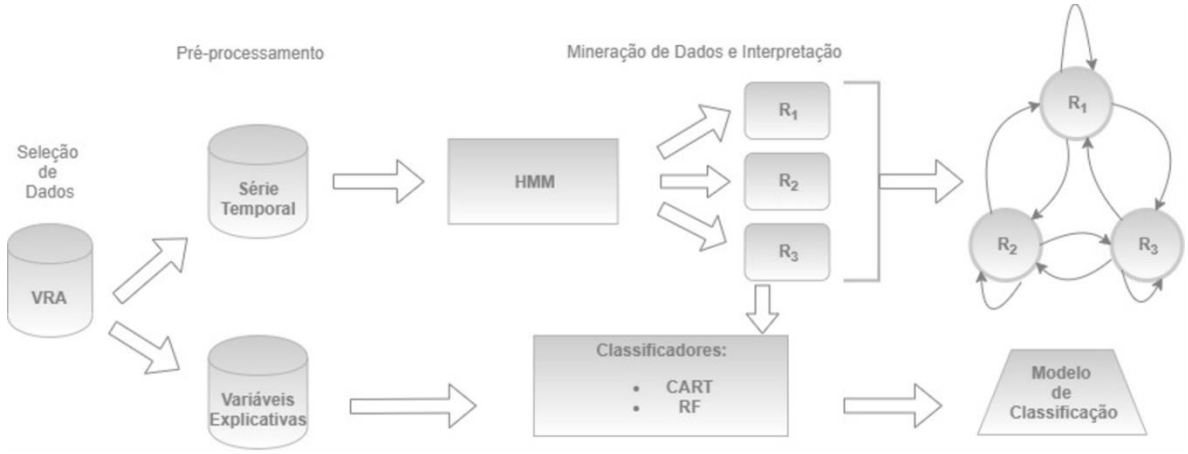

Figura 2: Methodology

The procedure illustrated in Figure 2 presents the methodological process of this study. It is observed that the development of this work took place in two phases. In the first, the data set was considered without the use of independent variables and, applying an unsupervised method. The HMM was used to identify regimes. In the second phase, independent variables were selected to compose the database in which classifiers were used to create the predictive model.

In this context, were considered daily flights canceled and those delayed 15 minutes or more that arrived at or departed from the São Paulo/Guarulhos International Airport. The standard 15-minute delay is used by the U.S. Bureau of Transportation Statistics (2020) when calculating delay rates. The non-scheduled flights were computed as performed with no delay, thus, they were not analyzed in this study. The database is composed of data collected from the years 2011 to 2017, except for the months of June and July 2014. In such months, the data were not provided due to the unavailability of information on scheduled flights and the suspension of the timetable system. These changes in the system occurred due to the World Cup period occurred in Brazil.

Flight delays and cancellations create inconveniences as they cause congestion at airports. Therefore, the variables “canceled flights” and “delayed flights” are considered attributes of interest in the data set. The variable “scheduled arrival/departure” is considered to obtain the daily percentage of delays and cancellations. The integration of the variables occurs through the sum of the total daily flights canceled and delayed, divided by the total of daily arrivals and departures scheduled flights. The time series obtained from the integrated variables represents the daily share (percentage) of delayed and canceled flights.

The HMM was used in the obtained time series, with the purpose of detecting regimes that could represent the intensity of flight delays and cancellations for a certain period of time. In order to select the best adjusted model, the AIC and BIC criteria were used to determine the number of regimes.

After the detection of regimes with the HMM, started the second phase of this work. The database which the classifiers were applied is composed for variables resulting from the HMM model and independent variables. Obtained “Current regime” and “previous regime” are variables resulting from the HMM. The independent variables, “month of the year” and “day of the week”, were extracted from the dataset, in order to investigate whether the occurrence of congested days is associated with a greater or lesser demand during the days of the week and

months of the year. Three independent variables were also considered, which according to Scarpel and Pelicioni (2018), are potential variables to deal with flight delays and cancellations at São Paulo/Guarulhos International Airport: Herfindahl Hirschman Index – HHI of Slots per day, Spacing and ConMov.

"HHI" is the variable that measures the market concentration at an airport and refers to the distribution of the flights daily share operated by companies within the airport (Santos and Robin, 2010). According to Abdel-Aty et al. (2007), Spacing is the time interval between two consecutive movements (arrival or departure) of scheduled flights. This work considered Spacing as two variables: "Average Spacing", represents the daily average time between consecutive arrivals and departures; and "Standard Deviation Spacing", representing the Spacing variability, that is, the difference in the time between actual and scheduled flight, arrivals or departures. ConMov is the daily number of consecutive movements of the same type, arrivals or departures, (Scarpel and Pelicioni, 2018). In this work, was considered the daily average "ConMov". Table 1 shows the variables considered and their definitions.

Table 1: Set of considered variables and their definitions.

Variable	Definition
Canceled flights	Daily canceled flights
Delayed flights	Daily delayed flights
Scheduled arrival/departure	Daily scheduled arrivals/departures flights
Current regime	Regime at time t
Previous regime	Regime at time t-1
HHI	Herfindahl-Hirschman Index
Av. Spacing	Daily average time between consecutive movments - arrival/departure (in minutes)
Std.Spacing	Daily standard deviation from the average time between two consecutive movements - arrival/departure (in minutes)
Av. ConMov	Daily average number of consecutive arrivals and consecutive departures
Month	Month of the year when the flight is scheduled (January, February, March, April, May, June, July, August September October November December)
Day of week	Day of the week when the flight is scheduled (Sunday, Monday, Tuesday, Wednesday, Thursday, Friday and Saturday)

The methods, CART and RF, were chosen because they are conceptually simple, good efficiency for predicting models, and the interpretation allowed by CART. The methods were applied to the dataset dividing it into training (70%) and validation (30%). The procedure for variables selection, variables relevance and the classification of CART and RF were performed simultaneously.

In CART method, it is important to determine the ideal tree size for pruning and avoid overfitting. Thus, a usual method considered was the "one-standard-deviation rule", where

according to Scarpel (2014) one is advised to choose the smallest tree in which cross-validation relative error is close to the minimum cross-validation relative error plus one standard deviation. The plot is composed of the estimated cross-validation errors versus a complexity parameter (cp) associated to the tree size. The complexity parameter measures the additional accuracy that a split adds to the tree. It is estimated by the linear combination of the error rate and the size of the tree, defined by the number of nodes in the terminals.

In RF method, according to Maindonald and Braun (2003), the main tuning parameter is the number mtry (number of variables randomly sampled at each split), which controls the trade-off between the amount of information in each individual tree and the correlation between trees. The mtry standard for classification tree is the square root of the number of variables of the model. To tuning the parameter, different numbers of mtry were tested according to the OOB error (out-of-bag).

After applying the methods, it was necessary to evaluate the performance of the classifiers. The error metric used in this work was the classification accuracy. The Equation 3 calculates the classifier accuracy. Given a classifier l,

$$acc(l) = 1 - err(l) = \frac{1}{n} \sum_{i=1...n} I(y_i = f(x_i)), \tag{3}$$

where n is the number of observations, I the identity function, y_i the known class and $f(x_i)$ the predicted class.

5. RESULTS AND DISCUSSION

To obtain the analyzed time series in this study, HMM was used as an unsupervised way to identify regimes. Figure 3 shows the evolution of the time series share of the delayed and canceled flights between the years 2011 and 2017.

To create the HMM model the number of regimes is defined a priori. To obtain a better fit, models from 2 to 6 regimes were evaluated, observing the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) criteria. They are represented in Figure 4.

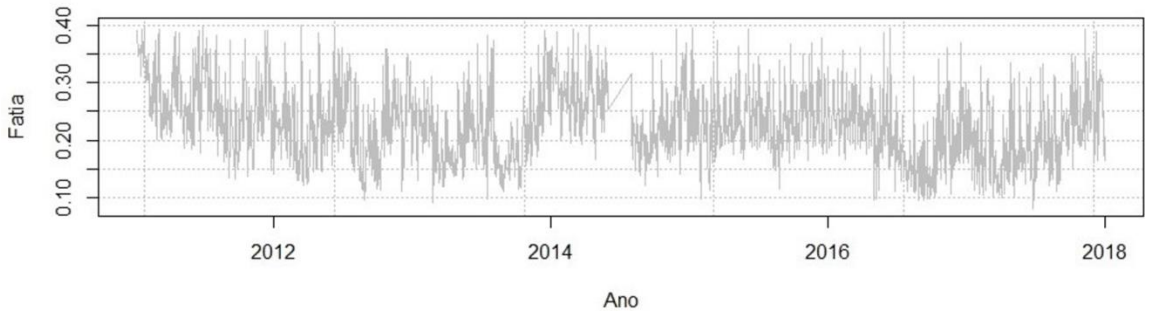

Figure 3: Time series daily share of the delayed and canceled flights

The number of parameters calculated for each regime shows that the models with more than three regimes would be very complex (due to the increase of the number of parameters) and less amount of gain of information (as observed in Figure 4), therefore the model with three regimes was better fitted.

The mean and standard deviation of each detected regime were estimated. The average daily share of delayed and canceled flights under regime 1 was 29.7%, and a standard deviation of 4.6%. The average of regime 2 was 22.1%, and standard deviation 3.0%. Finally, for the regime 3, the average was 15.6% and standard deviation 3.0%. Therefore, it can be inferred that regime 1 is added up to the daily share with the highest rate of flight delays and cancellations and the regime 3 with the lowest rate of flight delays and cancellations.

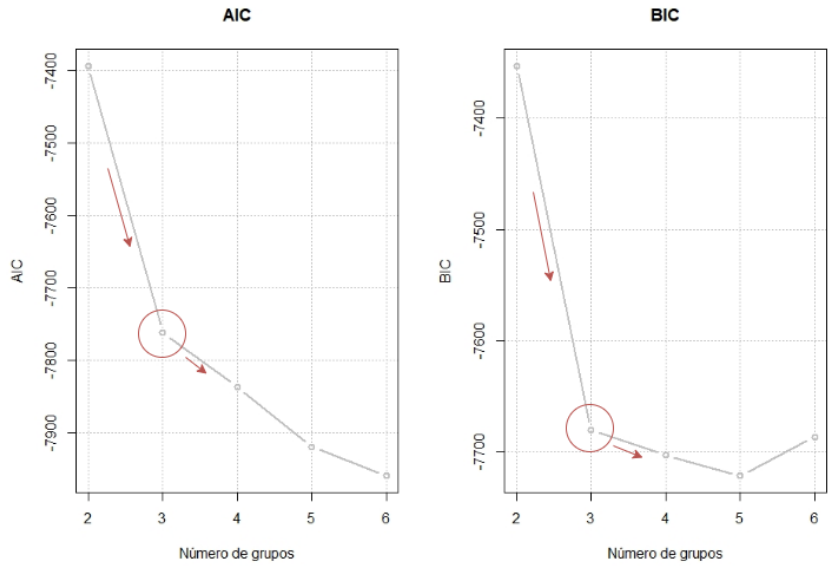

Figure 4: Akaike Information Criterion (AIC) e Bayesian Information Criterion (BIC)

In order to make analysis simpler, the regimes detected by the HMM model were defined, according to the average daily share of each regime, as follows: High congestion (regime 1): 29.7%; Average congestion (regime 2): 22.1%; Low congestion (regime 3): 15.6%. Figure 5 shows the histogram of the delayed and canceled flights share for the identified regimes.

As seen in Figure 5, regime 1 shows higher variation and a higher index of the daily share of delayed and canceled flights. This implies the occurrence of days with high congestion. Regimen 2 shows the majority of occurrences of days with indexes of the daily share of delayed and canceled flights close to 22%, which indicates the occurrence of days with average congestion. And the regime 3 has most occurrences around 15%, which implies days with low congestion.

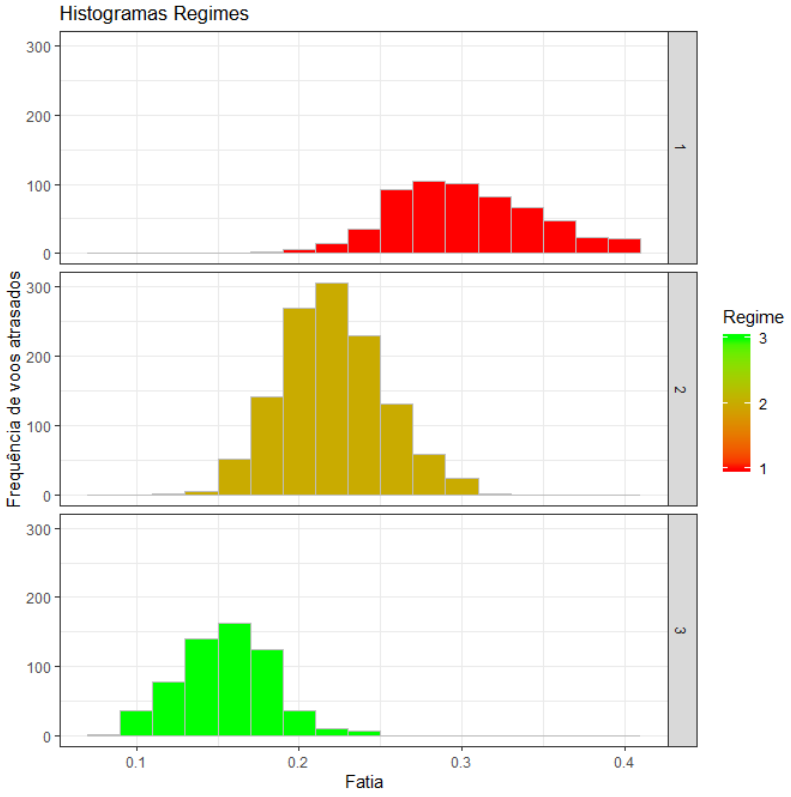

Figure 5: Histogram of HMM regimes

As probabilidades estimadas dos regimes, conhecidas como probabilidades posteriores, são observadas na matriz de transição que é composta por vetores de probabilidade representados pelas linhas, onde cada linha soma um. O vetor indica a probabilidade permanecer no regime

corrente ou de ir para outro no período de tempo *t+1*. A Tabela 2 apresenta a matriz de transição estimada a partir do modelo. A princípio, no vetor regime 1 a probabilidade de permanecer no regime corrente é de 75,0%. O segundo vetor mostra que a probabilidade do regime 2 ocorrer é de 80,0%. O terceiro vetor indica que a probabilidade do regime 3 ocorrer é de 87,0%. De acordo com a matriz de transição, neste cenário, a probabilidade de ir do regime 1 para o regime 3 – ou seja, de um dia pouco congestionado ocorrer após um dia muito congestionado – é insignificante.

The estimated probabilities of the regimes, known as posterior probabilities, are observed in the transition matrix, which is composed of probability vectors represented by the lines, where each line adds up to one. The vector indicates the probability to stay in the current regime or to go to another regime in time *t + 1*. Table 2 shows the transition matrix estimated from the model. In this case, in regime 1, the probability of remaining in the current regime is 75,0%. The second vector shows that the probability of occurs regime 2 is 80,0%. The third vector indicates that the probability of regime 3 to occur is 87,0%. According to the transition matrix, in this scenario, the probability of going from regime 1 to regime 3 – from the high congestion day to the low congestion day – is insignificant.

***Table 2**: Transition matrix*

		For		
	Regime	1	2	3
	1	0,75	0,25	**0,00**
To	2	0,13	0,80	0,07
	3	0,01	0,12	0,87

After the detected regimes, two predictive classification models were generated. The goal is to predict the regime in the period of time t + 1.

5.1. Classification and Regression Trees

The ideal size of the tree of the CART model was determined considering one-standard deviation rule, showed in Figure 6. The plot implies that the ideal size is a tree with six terminal nodes.

Figure 7 shows the tree obtained after pruning, with six terminal nodes and five splits. According to the plot, the variable “previous regime” (period t) is more relevant and, therefore, has a strong influence in predicting regimes at time t + 1. Terminal node 3 contains 698 observations, was classified as regime 1 and depends on the variable “previous regime”. Thus, the day that belongs

to regime 1 (high congestion), has an 88.1% probability that the next day remains congested and stays in the same regime. The error rate is 11.9%.

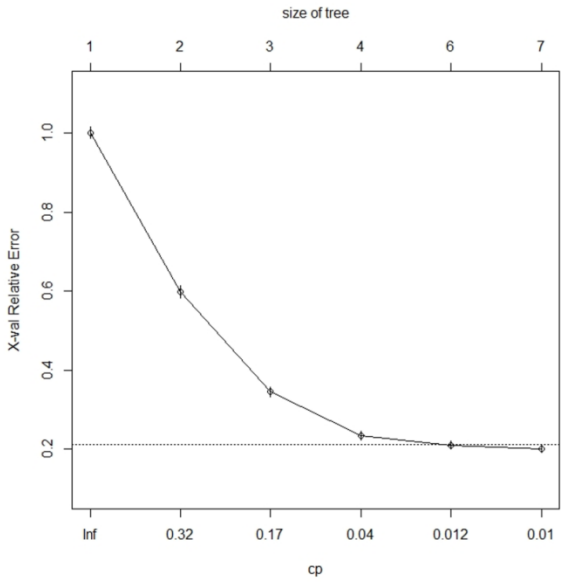

Figure 6: Cross validated error versus a complexity parameter

The terminal node 11 includes 712 observations and is classified as regime 3 (low congestion) and the probability of the following day remains in the current regime is 94.0%. The error rate is 6.0%. Note that the probability of regime 3 (low congestion) go to regime 1 (high congestion) is tiny. Therefore, there is small chance that a day with high congestion rate occurs after a day with low congestion rate of flight delays and cancellations.

The terminal node 10 covers 846 observations, and has high probability (82.4%) of belonging to regime 2 (average congestion). It occurs when the previous regime is 2 and the demand for scheduled flight arrivals and departures is high. When demand is high, the intervals between flight arrivals and departures are shorter, in this case, "Av. Spacing" <0.154. It is important to highlight that "Av. Spacing" and "Std. Spacing" were multiplied by one hundred, due to the rounding pattern of the R package.

The terminal node 6 contains 93 observations and it has been classified as regime 1 (high congestion). It occurs when the previous regime is 2, the demand is low ("Av. Spacing" ≥ 0.154) and the day of the week is Sunday or Friday. Thus, days belonging to regime 2 (average congestion) with scheduled flights where time interval between arrivals and departures is greater

than approximately two minutes, have 98.9% probability of going to regime 1 (high congestion).

In terminal node 8 (139 observations) the probability that regime 1 occurs is 61.2% when the previous regime is 2, the demand is low ("Av. Spacing" ≥ 0.154), the day of the week is Monday to Thursday or Saturday, and the market is less concentrated (HHI < 0.21). These results are in agreement with the results presented by Scarpel and Pelicioni (2018). Thus, on days with higher demand (peak periods) and less concentrated market (more airlines are operating), more congested days are expected.

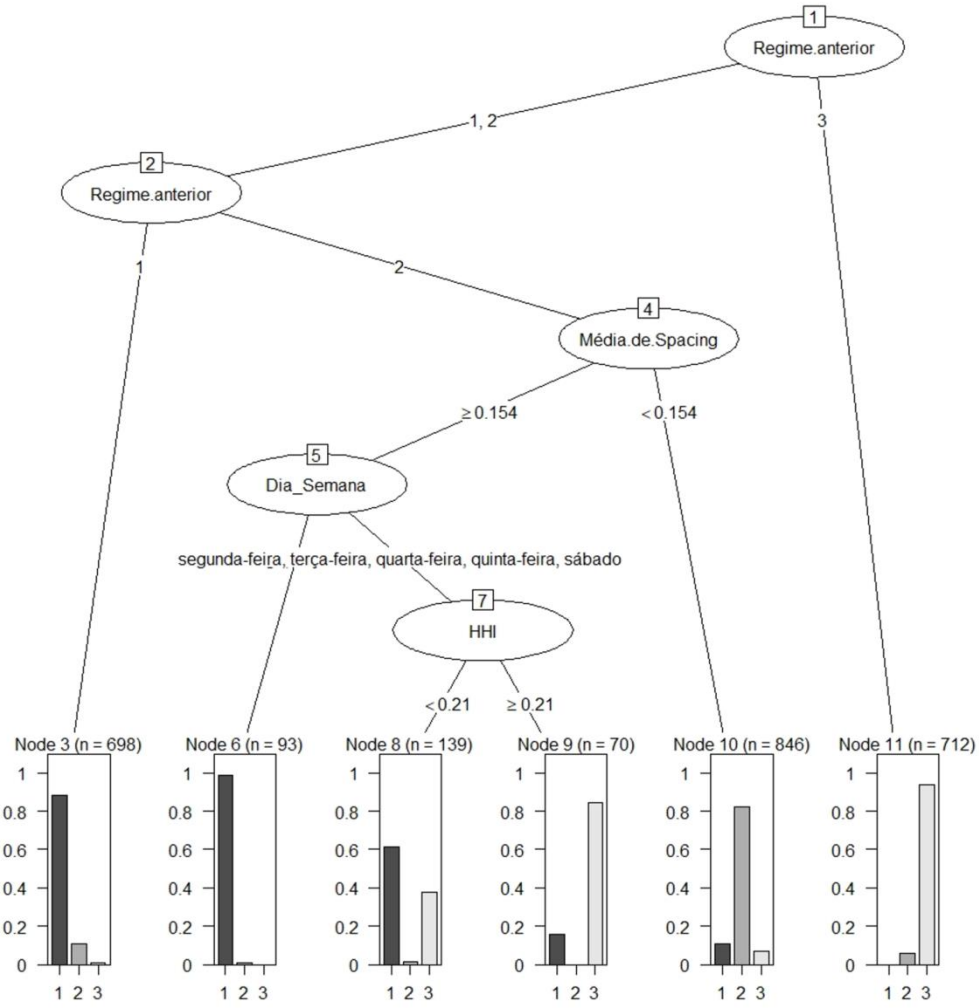

Figure 7: Classification tree with six terminal nodes

The terminal node 9 contains 70 observations and an 84.3% probability of being in regime 3. It is the most concentrated market (HHI ≥ 0.21), and airlines tend to internalize the delays.

According to Scarpel and Pelicioni (2018), lower delay rates are expected when the airport is more concentrated and it has lower demand. Therefore, terminal node 9 has probability above 80.0% of occurring in regime 3 (low congestion). For CART performance evaluation, the training set showed 86.6% accuracy and the test set, 88.3%.

5.2. Random Forests

To generate the RF model, the parameter mtry was defined. It is calculated with the square root of the number of variables of the model. As this model holds eight variables, the parameter is calculated as $\sqrt{8}$. The number resulting is 2, since it is the largest integer extracted from the root. However, the value of mtry was tuned by testing values from 1 to 4. The lowest value of the OOB error was obtained with mtry equal to 2. The generated model RF had an estimated accuracy of 88.9% for the training set and 89.7% for the test set.

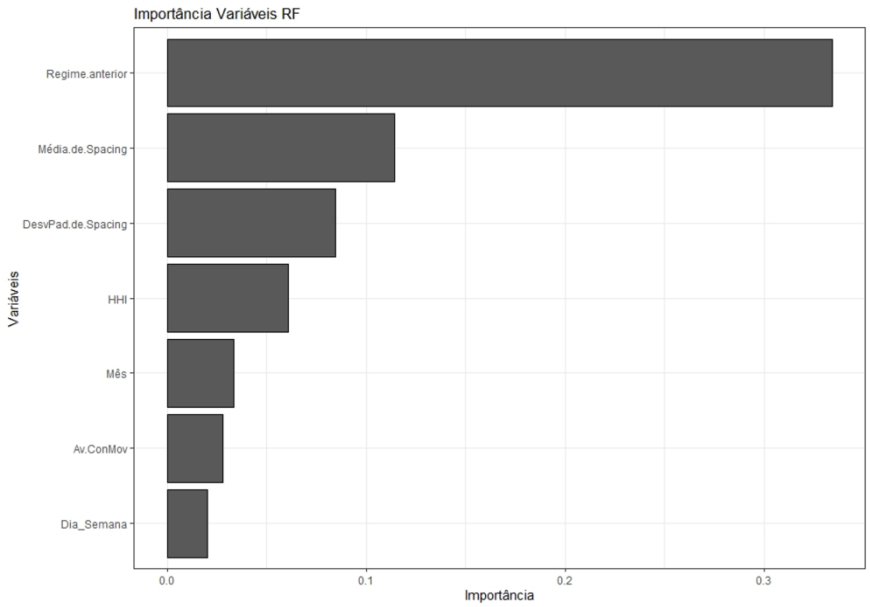

Figure 8: Random Forest model: variables importance

Figure 8 shows the importance variables according to RF model. Although RF does not offer the possibility of interpretation, it is observed that the variable “previous regime” is the most important, which is in agreement with CART.

Both models, CART and RF, presented good classification performance in the predictive model generated to anticipate the occurrence of congested days at the São Paulo/Guarulhos International Airport with a period of one day in advance.

6. CONCLUSION

This study aimed to create a model to anticipate the occurrence of congested days at the São Paulo/Guarulhos International Airport. At first, an HMM model was generated using the time series daily share of delayed and canceled flights. The best model was fitted considering the AIC and BIC criteria. Three regimes were identified and defined according to the average rate of flight delays and cancellations as follows: regime 1 (high congestion, 29.7%); regime 2 (average congestion 22.1%); and regime 3 (low congestion 15.6%).

Subsequently, the predictive model was generated applying the methods CART and RF. Through CART model, the determinant variables for congested days and how they are combined were identified. The produced tree holds six terminal nodes, and the variables “previous regime”, “Av. Spacing”, “day of the week” and “HHI”. The results obtained are consistent with the results found in literature. RF model estimated the importance variables, in which, as in CART model, “previous regime” is the most significant variable.

The classification models employed showed to be acceptable and presented good accuracy to anticipate the occurrence of congested days at São Paulo/Guarulhos International Airport one day in advance. Some limitations were identified in the course of this study: (i) the variables obtained from the result of the HMM model were taken as real data, so the label uncertainty was not considered; (ii) the predictive model generated anticipates the occurrence of congested days only for time t + 1. For future work, it is proposed to investigate how other variables (for example, “spacing” considering peak times and seasons) would influence the predictive model for congested days in order to expand the forecast horizon.

Referências

Abdel-Ary, M.; Lee, C.; Bai, Y.; Li, X. e M. Michalak (2007) Detecting periodic patterns of arrival delay. *Journal of Air Transport Management*, v. 13, n. 6, p. 355–361.

ANAC – Agência Nacional de Aviação Civil, acessado 2020, Metadados do conjunto de dados: Voo Regular Ativo (VRA), <anac.gov.br/acesso-a-informacao/dados-abertos/areas-de-atuacao/voos-e-operacoes-aereas>.

Bendinelli, W. E.; H. F. A. J. Bettini e A. V. M. Oliveira (2016) Airline delays, congestion internalization and non-price spillover effects of low cost carrier entry. *Transportation Research: Part A, Policy and Practice*, v. 85, p. 39-52.

Bureau of Transportation Statistics, acessado 2020. Airline On-Time Performance Data, <transtats.bts.gov>.

Breiman, L. (2001) Random Forests. *Machine Learning*, v. 45, n.1, p. 5-32.

Chandramouleeswaran, K. R.; Krzemien, D.; Burns, K. e H. T. Tran (2018) Machine Learning Prediction of Airport Delays in the US Air Transportation. *2018 Aviation Technology, Integration, and Operations Conference,* AIAA, Atlanta, Georgia, USA, p. 1–10.

Costa, T.F.G.; Lohmann, G.; Oliveira, A.V.M.; (2010) A model to identify airport hubs and their importance to tourism in Brazil. *Research in Transportation Economics,* v. 26, p. 3–11.

Jacquillat, A. e A. R. Odoni (2015) An Integrated Scheduling and Operations Approach to Airport Congestion Mitigation. *Operations Research*, v. 63, n. 6, p. 1390–1410.

James, G.; Witten, D.; Hastie, T. e R. TibshiraniI (2013) *An Introduction to Statistical Learning with Applications in R*. Ed. Springer , NewYork, NY, USA.

Janic, M. (2015) Modelling the resilience, friability and costs of an air transport network affected by a large-scale disruptive event. *Transportation Research Part A: Policy and Practice,* v. 71, p. 1-16.

Kandhasamy, J. P. e S. Balamurali (2015) Performance Analysis of Classifier Models to Predict Diabetes Mellitus. *Procedia Computer Science,* v. 47, p. 45–51.

Killic, R. e A. I. Eckley (2014) changepoint: An r package for changepoint analysis. *Journal of Statistical Software,* v. 58, n. 3, p. 1–19.

Maindonald, J. e W. J. Braun (2003) Data *Analysis and Graphics Using R an Example-Based Approach*. Ed. Cambridge University Press, New York, NY, USA.

Rebollo, J. J. e H. Balakrishnan (2014) Characterization and prediction of air traffic delays. *Transportation Research Part C: Emerging Technologies,* v. 44, p. 231–241.

Santos, G. e M. Robin (2010) Determinants of delays at European airports. *Transportation Research Part B: Methodological,* v. 44, n. 3, p. 392-403.

Santos, T. A. dos; Vendrame, I.; Alves, C. J. P.; Caetano, M. e J. P. S. Silva (2018) Modelo de identificação do impacto futuro de chuvas extremas nos atrasos/cancelamentos de voos. *Transportes*, v. 26, n. 2, p. 44–53.

Scarpel, R. A. (2014) A demand trend change early warning forecast model for the city of São Paulo multi-airport system. *Transportation Research Part A: Policy and Practice,* v. 65, p. 23–32.

Scarpel, R. A. e L. C. Pelicioni (2018) A data analytics approach for anticipating congested days at the São Paulo International Airport. *Journal of Air Transport Management*, v. 72, p. 1–10.

Xiong, J. e M. Hansen (2013) Modelling airline light cancellation decisions. *Transportation Research Part E: Logistics and Transportation Review*, v. 56, p. 64-80.

Visser, I. (2011) Seven things to remember about hidden Markov models: A tutorial on Markovian models for time series. *Journal of Mathematical Psychology*, v. 55, n. 6, p. 403–415.

Wensveen, J. G. (2016) *Air Transportation: A Management Perspective* (8ª. ed.). Routledge, New York, NY, USA and London, UK.

Yu, B.; Guo, Z.; Asian, S.; Wang, H. e G. Chen (2019) Flight delay prediction for commercial air transport: A deep learning approach. *Transportation Research Part E*, v. 125, p. 203–221.

Zucchini, W.; Macdonald, I. L. e R. Langrock (2017) Hidden *Markov Models for Time Series: An Introduction Using R.* Ed. CRC Press, Boca Raton, FL, USA.