TRANSPORTES
www.revistatransportes.org.br
anpet
associação nacional de pesquisa e ensino em transportes
www.anpet.org.br

Antecipação de mudança de regime na fatia diária de voos atrasados e cancelados no aeroporto internacional de São Paulo/Guarulhos

Anticipating a regime change in the daily share of delayed and canceledflights at são paulo/guarulhos international airport

Rosana Batista Teixeira[1], Rodrigo Arnaldo Scarpel[2]

[1]Instituto Tecnológico de Aeronáutica, São Paulo, Brasil, rosana.bteix@gmail.com
[2]Instituto Tecnológico de Aeronáutica, São Paulo, Brasil, rodrigo@ita.br

Recebido:
20 de novembro de 2019
Aceito para publicação:
18 de maio de 2020
Publicado:
30 de abril de 2021
Editor de área:
Alexandre de Barros

Palavras-chaves:
Detecção de Pontos de Mudança.
Modelos Escondidos de Markov.
Modelos de Classificação.
Atrasos de voos.

Keywords:
Change Point Detection.
Hidden Markov Models.
Classification Models.
Flight delays.

DOI:10.14295/transportes.v29.i1.2236

RESUMO
Atrasos e cancelamentos de voos são ocorrências frequentes na maioria dos aeroportos em todo o mundo. No Brasil, a liberalização do transporte aéreo provocou a concentração de voos em alguns aeroportos gerando o aumento da ocorrência de atrasos e cancelamentos de voos em razão de dias congestionados. O Aeroporto Internacional de São pulo/Guarulhos (GRU) é um dos mais afetados por atrasos causados por congestionamento no país. O objetivo deste trabalho é a criação de um modelo de previsão para a antecipação da ocorrência de dias congestionados no Aeroporto Internacional de São Paulo/Guarulhos. A precisão do modelo foi considerada satisfatória e antecipou a mudança de regime na fatia diária de voos atrasados e cancelados para um período à frente.

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. INTRODUCÃO

Dias congestionados nos aeroportos em decorrência de atrasos e cancelamentos de voos são um problema universal. Os atrasos são recorrentes e cada vez mais comuns na rotina dos passageiros, principalmente nos aeroportos considerados *hubs* (onde se concentram as conexões). A alta concentração de voos em um aeroporto para a realização de conexões pode gerar atrasos por congestionamento, pois o número de aeronaves tende a se aproximar da capacidade máxima do aeroporto. Os atrasos que podem ocorrer em consequência deste congestionamento aumentam os custos operacionais das companhias aéreas e o tempo de viagem dos passageiros,

além de gerar uma carga de trabalho adicional para os controladores de voo, o que aumenta o nível de estresse (Wensveen, 2016).

A análise dos atrasos aéreos é importante, pois a compreensão de suas potenciais causas pode possibilitar o desenvolvimento de possíveis soluções para ajudar no desempenho do sistema de transporte aéreo (Abdel-Aty *et al.*, 2007; Rebollo e Balakrishnan, 2014; Scarpel e Pelicioni, 2018). No que diz respeito ao Brasil, com a liberalização do transporte aéreo houve a concentração de voos em alguns aeroportos *hub (Costa et al*., 2010). O Aeroporto Internacional de São Paulo/Guarulhos, atualmente o maior *hub* dos aeroportos brasileiros, é o que mais sofre com atrasos por congestionamento (Scarpel e Pelicioni, 2018). Logo, antecipar a ocorrência de dias congestionados no aeroporto de São Paulo/Guarulhos é de grande importância para o desenvolvimento de estratégias com o propósito de reduzir os atrasos e cancelamentos de voos e apoiar seu planejamento.

O objetivo deste trabalho é criação de um modelo de previsão para a antecipação da ocorrência de dias congestionados no Aeroporto Internacional de São Paulo/Guarulhos (GRU). Para alcançar este objetivo, buscou-se a identificação de grupos homogêneos – regimes – dentro da série temporal representada pela fatia diária de voos atrasados e cancelados do Aeroporto Internacional de São Paulo/Guarulhos. A partir dos regimes identificados, foi criado um modelo de classificação, comparado em dois algoritmos diferentes, para prever com antecedência mudanças de regime, podendo assim ser uma ferramenta de apoio ao planejamento e de suporte à tomada de decisão no aeroporto GRU. Este trabalho pretende contribuir para a literatura propondo uma abordagem que permite, sequencialmente, detectar pontos de mudança na série temporal da fatia diária de voos atrasados e cancelados em um aeroporto e criar modelos para antecipar a ocorrência de dias congestionados.

2. REVISÃO BIBLIOGRÁFICA

Atrasos e cancelamentos de voos tem sido objeto de uma série de estudos. De acordo com Xiong e Hansen (2013), o sistema de aviação enfrenta grandes desafios ao lidar com a alta demanda quando a capacidade do sistema é reduzida. Diante dos atrasos, os horários das companhias aéreas podem sofrer mudanças não previstas, pois alguns voos atrasam em razão da chegada tardia do voo anterior e, devido aos horários apertados, estes atrasos podem se propagar (Abdel-Aty et al., 2007). De acordo com Jacquillat e Odoni (2015), a maioria dos atrasos de voos é resultante do desbalanceamento entre demanda e capacidade. Segundo os autores, este desequilíbrio é causado pelo crescimento do tráfego aéreo e as limitações de capacidade dos aeroportos com grande movimento. De acordo com Xiong e Hansen (2013), os atrasos são problemas significativos resultantes da excessiva demanda de voos e estão fortemente associados com as operações, duração do voo e condições climáticas dos aeroportos de origem e destino. Os autores reiteram que este problema é agravado pela competitividade das companhias aéreas, que confrontadas pelos altos custos das aeronaves, buscam o máximo aproveitamento.

De acordo com Santos *et al.* (2018), atrasos e cancelamentos de voos apresentam-se como alguns dos principais problemas associados à interrupção das operações de uma rede de transporte aéreo. Janic (2015) afirma que uma rede de transporte aéreo consiste em aeroportos e rotas operadas pelas companhias aéreas. Segundo o autor, as perturbações de grande escala podem comprometer o funcionamento da rede. Ente eles estão o mau tempo, falhas de determinados componentes da rede considerados cruciais (sistemas dos computadores centrais),

ações relacionadas aos funcionários de transporte aéreo (por exemplo, greves), desastres naturais, ameaças e ataques terroristas, incidentes ou acidentes aéreos. De acordo com Abdel-Aty *et al.* (2007), o aumento do atraso de voos deve-se principalmente ao clima adverso nas proximidades dos aeroportos, à falta de capacidade das pistas, ao aumento do número de aeronaves e ao controle de tráfego aéreo deficiente.

No campo da análise de dados, métodos de aprendizado de máquina derivados de modelos complexos e algoritmos elaborados podem ser utilizados para a realização de análises preditivas. Santos e Robin (2010) abordaram atrasos de voos nos aeroportos europeus utilizando análise de regressão múltipla para identificar causas de atrasos. Rebollo e Balakrishnan (2014) empregaram abordagens de geração de agrupamento e classificação para a prevenção de atrasos nos horários de partida dos voos em uma ligação específica ou em um aeroporto específico em algum momento no futuro. Chandramouleeswaran *et al.* (2018) apresentaram uma abordagem para prever atrasos nos aeroportos dos Estados Unidos fazendo uma comparação entre redes neurais e análise de regressão. Foram considerados dados temporais (dia e hora), congestionamentos, redes de aeroportos e o clima. Yu *et al.* (2019) utilizaram um método de aprendizado não supervisionado combinado com um algoritmo de aprendizado supervisionado de regressão e classificação para realizar análises de prevenção de atrasos de voos.

No Brasil, Scarpel e Pecicioni (2018) empregaram uma abordagem de análise de dados para construir um modelo de alerta com a finalidade de prever a ocorrência de dias congestionados no GRU. A combinação de abordagens de modelagem que se baseiam em diferentes premissas permitiu gerar um modelo com maior flexibilidade e trouxe melhorias na precisão das previsões. Por uma concepção diferente, Bendinelli *et al.* (2016) analisaram se a ausência de concorrência favorecia o aumento das taxas de atrasos e cancelamento de voos, relação que foi confirmada pelos autores.

Na literatura, há a criação de modelos para prever atraso médio no dia ou a fatia de voos atrasados, considerando apenas informações relativas ao dia. Nesse tipo de modelo as observações são independentes, não se busca identificar associações entre dias diferentes. Este estudo tratou os atrasos em aeroportos por meio da identificação de padrões de dias, de forma dependente em que se buscou identificar regimes que pudessem ser caracterizados utilizando uma distribuição de probabilidade, ou seja, média e desvio padrão. Para a geração do modelo foi aplicado o conceito de detecção de pontos de mudança pela perspectiva de regimes, onde para detectá-los, foi empregado o modelo escondido de Markov.

3. REFERENCIAL TEÓRICO

3.1. Detecção de pontos de mudança e modelos escondidos de Markov

Detecção de Pontos de Mudança (*Change Point Detection* - CPD) é a estimação de pontos em uma série temporal para os quais as propriedades estatísticas são diferenciadas e os intervalos entre os pontos encontrados são chamados de regimes ou estados, representando um conjunto de observações com características comuns entre si. De acordo com Killick e Eckley (2014), detecção de pontos de mudança é a determinação de pontos nos quais as propriedades estatísticas são alteradas em uma sequência de observações. Considerando observações no tempo t e $t+1$, se um ponto no tempo t pertence a um grupo diferente de uma observação no tempo $t+1$, então um ponto de mudança ocorre entre as duas observações, como mostra a Figura 1.

Pela perspectiva de CPD como um problema de geração de agrupamento, as observações que pertencem à série temporal, localizadas entre um ponto de mudança e outro, formam conjuntos

de observações que compartilham das mesmas propriedades estatísticas, denominados regimes. Para tratar o problema de CPD, visto como um problema de formação de agrupamentos, neste trabalho foi empregado o método de aprendizagem de máquinas Modelo Escondido de Markov (*Hidden Markov Models* - HMM).

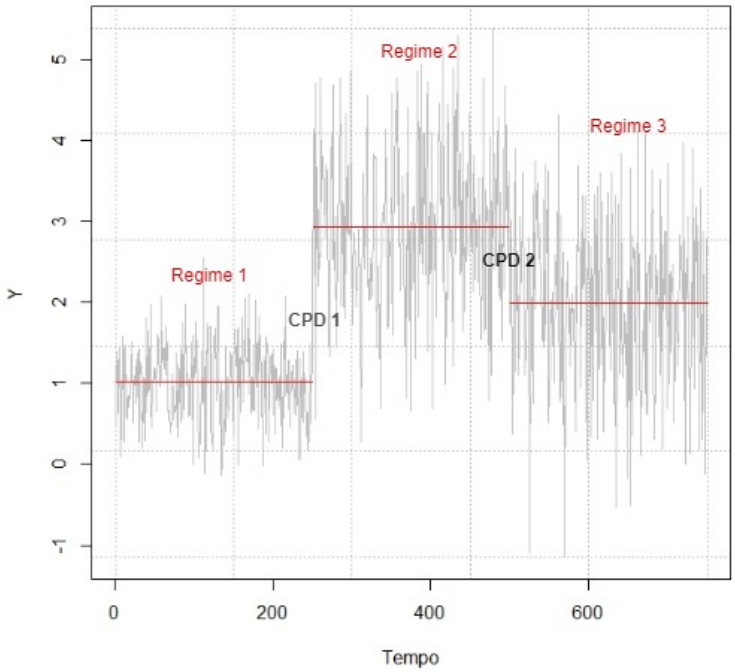

Figura 1. Pontos de mudança de uma série temporal

Os Modelos Escondidos de Markov são modelos nos quais a distribuição de probabilidade que gera uma observação depende de um estado pertencente ao processo oculto de Markov (Zucchini *et al.*, 2017). Os autores descrevem o modelo escondido de Markov O_t: $t \in \mathbb{N}$ como um tipo de mistura dependente, em que as sequências históricas das observações $O_{1:t}$ e dos estados escondidos $S_{1:t}$ são representadas por dois processos descritos como,

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

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

Em que a Equação 1 representa um processo de parâmetros não observado S_t: $t = 1, 2,\ldots$ que satisfaz a propriedade de Markov. A Equação 2 representa um processo dependente de estados de forma que a distribuição de O_t: $t = 1, 2,\ldots$ depende somente do estado atual e não de estados anteriores ou das observações. Se a MC (S_t) tem m estados, O_t será um HMM de m estados. Os modelos escondidos de Markov são definidos como modelos com estados discretos, caracterizados por suas funções de distribuição onde a evolução dos estados no decorrer do tempo é governada por um processo de Markov (Visser, 2011). No que diz respeito aos critérios de seleção dos modelos, em um HMM Zucchini *et al.* (2017) afirmam que o modelo é melhor ajustado à medida que se aumenta o número de estados (critério *likelihood*). Entretanto, o aumento do número de estados aumenta exponencialmente o número de parâmetros, o que torna o problema muito complexo. Dois critérios populares de seleção do melhor ajuste de modelo são *Akaike Information Criterion* (AIC) e *Bayesian Information Criterion* (BIC), onde o modelo melhor ajustado é aquele que apresenta menor valor de AIC e BIC, considerando o aumento do número de parâmetros.

3.2. Modelos de classificação

Diferentes estados observados são identificados pelo HMM e podem ser vistos como grupos imbuídos com atributos próprios. Assim, faz-se necessário o uso de um classificador para que, por meio da seleção de variáveis, seja criado um modelo de previsão. Com o propósito de uma análise comparativa, são considerados dois modelos: Árvores de Classificação e Regressão (*Classification and Regression Tree* – CART) e Florestas Aleatórias (*Random Forests* - RF). Neste trabalho é feita uma breve apresentação destes métodos de classificação.

3.2.1. Árvores de Classificação e Regressão

Árvores de classificação e regressão (CART) são conceitualmente simples, úteis para interpretação e visualização. Como sugerido pelo nome podem ser utilizadas tanto para regressão quanto para classificação. As variáveis respostas das árvores de regressão são dadas pela resposta média das observações de treino que pertencem ao mesmo nó. No entanto, as árvores de classificação preveem que cada observação pertence à classe de observações de treinamento mais comum na região à qual pertence (James *et al.*, 2013). De acordo com Scarpel (2014), CART é atraente quando a interpretação é uma questão importante, uma vez que os dados são projetados para detectar as variáveis de predição importantes e gerar uma estrutura de árvore para representar a partição identificada. O algoritmo de uma árvore de regressão pode ser resumido em quatro etapas:

1. Fazer partições binárias recursivas para expandir a árvore, parando apenas quando cada nó terminal tiver menos que o número mínimo de observações;
2. Aplicar a poda em função da complexidade de custos em árvores grandes, para obter uma sequência das melhores subárvores, em função de um parâmetro α;
3. Utilizar a validação cruzada *K-fold* para escolher α. Para cada *k=1,..., K*; a) Repetir os passos 1 e 2 em todos, exceto na k-enésima dobra dos dados de treinamento; b) Avaliar o erro quadrático médio previsto na k-enésima dobra deixada de fora em função de α; calcular a média dos resultados para cada valor de α e escolher α para minimizar o erro médio;
4. Retornar à subárvore da etapa 2 que corresponde ao valor escolhido de α.

A expansão de uma árvore de classificação é bastante similar. Entretanto, não se pode utilizar soma de quadrados dos resíduos como critério para partições. A taxa de erro de classificação é feita por meio das medidas *Gini* e Entropia. Para que se determine o tamanho ideal de uma árvore, um método considerado é a regra de "um desvio padrão". Por esse método, escolhe-se a menor árvore cujo erro relativo de validação cruzada seja próximo ao erro relativo mínimo de validação cruzada mais um desvio padrão (Scarpel, 2014).

3.2.2. Florestas Aleatórias

Florestas Aleatórias (RF) são um método de classificação composto por várias árvores de decisão que combina o conceito de *bagging* com a seleção aleatória de variáveis a cada partição. De acordo Breiman (2001), RF consiste em se usar um subconjunto de variáveis de entrada, selecionadas aleatoriamente para expandir cada árvore. Dentre os benefícios de RF inclui-se boa acurácia, relativamente robusto a *outliers* e ruídos, estimativas internas úteis de erro e avaliação da importância relativa das variáveis. O autor concluiu que RF é uma ferramenta eficaz para fazer previsões e, ao injetar o tipo certo de aleatoriedade, torna-se um método de classificação e regressão preciso.

De modo geral, Kandhasamy e Balamurali (2015) descrevem a expansão das árvores que compõem o método RF como segue:

1. Se o número de observações no conjunto de treinamento é N, faz-se a amostragem de N observações aleatoriamente, com reposição a partir dos dados originais. As amostras serão o conjunto de treinamento;
2. Se houver M variáveis de entrada, subconjunto delas é selecionado aleatoriamente e a melhor partição é utilizada para dividir o nó. O Valor M é mantido constante durante a expansão da floresta;
3. Cada árvore deve se expandir o máximo possível e não há podas.

4. METODOLOGIA

Inicialmente, foram descritos a fonte e seleção do conjunto de dados, o tratamento e integração para a obtenção dos dados utilizados. Em seguida, foi feita uma explanação do método utilizado para agrupar os dados em regimes e, por fim, foram apresentadas as variáveis explicativas e os modelos utilizados na criação dos classificadores.

A base de dados utilizada para o desenvolvimento deste estudo é conhecida como Voo Regular Ativo – VRA, disponibilizada no site da Agência Nacional de Aviação Civil (ANAC). É constituída de informações de voos que apresentam alterações (atrasos, antecipações e cancelamentos), horários em que ocorreram e as justificativas apresentadas pelas empresas para tais alterações (ANAC 2019). As análises foram realizadas utilizando o software R e os pacotes depmixS4, rpart, partykit e randomForest.

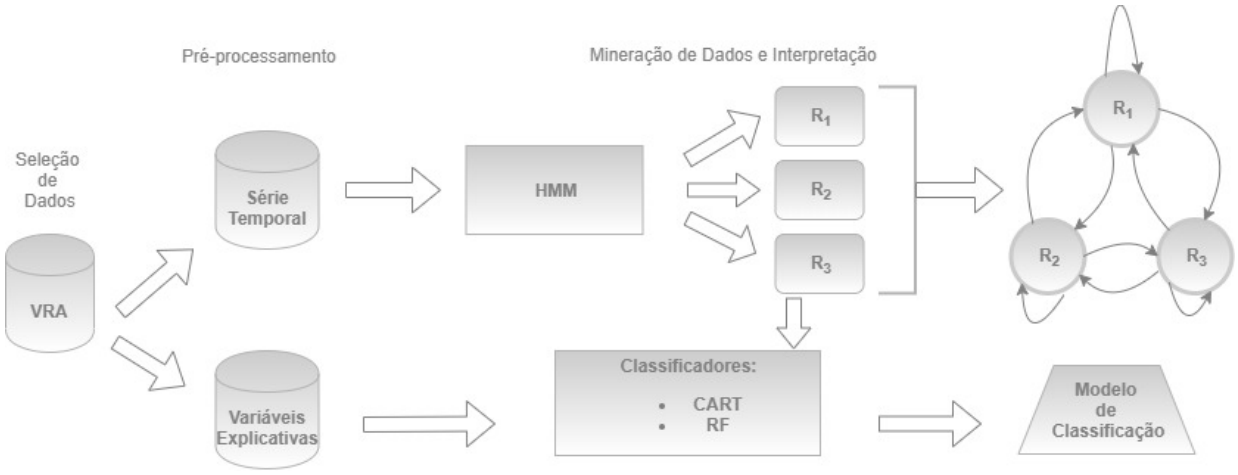

Figura 2. Procedimento Metodológico

O procedimento ilustrado na Figura 2 apresenta o processo metodológico deste estudo. Observa-se que o desenvolvimento deste trabalho deu-se em duas fases. Na primeira, foi considerado o conjunto de dados sem o uso de variáveis explicativas e, de modo não supervisionado, foi empregado o método de HMM para a identificação de regimes. Na segunda fase, foram selecionadas variáveis explicativas para a composição da base de dados na qual foram empregados classificadores para a criação do modelo de previsão.

Neste contexto, foram considerados os voos diários cancelados e os realizados com atraso de 15 minutos ou mais que chegaram ou partiram do Aeroporto Internacional de São Paulo/Guarulhos. O padrão de 15 minutos de atraso é utilizado pelo Bureau of

Transportation Statistics (2020) dos EUA no cálculo de índices de atrasos. Os voos com horário não programado foram tidos como voos realizados sem atraso; logo, não foram analisados neste estudo. A base de dados é composta por dados coletados dos anos de 2011 a 2017, exceto os meses de junho e julho do ano de 2014 cujos dados não foram fornecidos devido à indisponibilidade de informações de voos regulares e da suspensão do sistema de horários. Tais mudanças no sistema ocorreram em razão do período da Copa do Mundo que ocorria no Brasil.

Os atrasos e cancelamentos de voos geram transtornos por causarem congestionamentos nos aeroportos. Logo, são considerados como atributos de interesse do conjunto de dados as variáveis "voos cancelados" e "voos realizados com atraso". A variável "chegada/partida programada" também é considerada para que seja possível observar a porcentagem de atrasos e cancelamentos em relação às chegadas e partidas de voos programados do dia. A integração das variáveis ocorre por meio da soma do total de voos diários cancelados e realizados com atraso, dividida pelo total de chegadas e partidas diárias de voos programados. A série temporal obtida a partir das variáveis integradas representa a fatia (porcentagem) diária dos voos atrasados e cancelados.

O HMM foi empregado na série temporal obtida, com o propósito de detectar regimes que pudessem representar a intensidade de atrasos e cancelamentos de voos de um determinado período de tempo. A fim de selecionar o modelo melhor ajustado foram utilizados os critérios de AIC e BIC para determinar o número de regimes.

Após a detecção de regimes por meio do HMM, foi iniciada a segunda fase deste trabalho. A base de dados em que foram empregados os classificadores é composta por variáveis resultantes do modelo de HMM e por variáveis explicativas. Obtidas a partir do HMM são "regime atual" e "regime anterior". As variáveis explicativas "mês do ano" e "dia da semana" foram extraídas do conjunto de dados, com o fim de investigar se a ocorrência de dias congestionados está associada a uma demanda maior ou menor durante os dias da semana e meses do ano. Ainda foram consideradas três variáveis explicativas que, de acordo com Scarpel e Pelicioni (2018), são variáveis potenciais para tratar atrasos e cancelamentos de voos no Aeroporto Internacional de São Paulo/Guarulhos: Índice *Herfindal Hirschman* (*Herfindal Hirschman* Index – HHI) de Slots por dia, *Spacing* e ConMov.

"HHI" é a variável que mede a concentração de mercado em um aeroporto e se refere à distribuição da fatia diária de voos operada pelas empresas dentro do aeroporto (Santos e Robin, 2010). De acordo com Abdel-Aty *et al.* (2007), *Spacing* é o intervalo de tempo entre dois movimentos (chegada ou partida) de voos programados consecutivos. Neste trabalho, *Spacing* foi tratado com duas variáveis: "Média de *Spacing*", variável que representa a média diária de intervalo entre chegadas e partidas consecutivas; e o "Desvio Padrão de *Spacing*", representando a variabilidade de *Spacing*, ou seja, a diferença no intervalo de tempo entre as chegadas e partidas de voo real e o programado. *ConMov* é o número diário de movimentações consecutivas do mesmo tipo, chegadas e partidas, (Scarpel e Pelicioni, 2018). Neste trabalho foi considerada a "média diária de *ConMov*". Seguem na Tabela 1 as variáveis consideradas neste estudo e suas definições.

Quanto aos métodos, CART e RF foram escolhidos por serem conceitualmente simples, apresentarem boa eficácia para modelos de previsão, e pela interpretação permitida pelo CART. Os métodos foram aplicados ao conjunto de dados divididos em treino (70%) e validação (30%). O procedimento de verificação da relevância e seleção das variáveis e a classificação das observações dos algoritmos CART e RF foram executados simultaneamente.

Tabela 1 – Variáveis do conjunto de dados e suas definições

Variável	Definição
Voos cancelados	Total diário de voos cancelados
Voos realizados com atraso	Total diário de voos realizados com atraso
Chegada/partidas programadas	Total de chegadas/partidas de voos programados para o dia
Regime atual	Regime no tempo t
Regime anterior	Regime no tempo t-1
HHI de slot por dia	Índice Herfindal-Hirschman de concentração (concentração de mercado)
Média de Spacing	Média diária do tempo entre chegada/partida consecutivas (em minutos)
DesvPad de Spacing	Desvio padrão diário do tempo médio entre chegada/partida consecutivas (em minutos)
Média de ConMov	Número médio diário de chegadas consecutivas e partidas consecutivas
Mês	Mês do ano em que o voo está programado (janeiro, fevereiro, março, abril, maio, junho, julho, agosto, setembro, outubro, novembro, dezembro)
Dia da Semana	Dia da semana em que o voo está programado (domingo, segunda, terça, quarta, quinta, sexta e sábado)

No método CART, há a necessidade de se determinar o tamanho ideal da árvore para realizar a poda e evitar superajustes. Desta forma, foram aplicadas a validação cruzada 10-*fold* e a regra "um desvio padrão". O gráfico gerado é composto pelos erros estimados da validação cruzada versus o parâmetro de complexidade (cp) associado ao tamanho da árvore. O parâmetro de complexidade mede a precisão adicional que uma partição adiciona à árvore. A precisão é estimada pela combinação linear da taxa de erro e o tamanho da árvore, definida pelo número de nós nos terminais.

Quanto ao RF, de acordo com Maindonald e Braun (2003), o principal hiperparâmetro a ser otimizado é o número *mtry* (número de variáveis testadas aleatoriamente a cada partição), que controla a quantidade de informações em cada árvore individual e a correlação entre elas. O padrão *mtry*, para árvores de classificação, é a raiz quadrada do número de variáveis que compõe o modelo. Para otimizar o hiperparâmetro foram testados diferentes números de *mtry* em função do erro OOB (*out-of-bag*).

Após a aplicação dos métodos foi necessário avaliar o desempenho dos classificadores para este conjunto de dados em estudo. A métrica de erro utilizada neste trabalho foi a acurácia de classificação. A acurácia de um classificador é calculada pela Equação 3. Dado um classificador l,

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

em que *n* é o número de observações, *I* a função identidade, y_i a classe conhecida e $f(x_i)$ a classe predita.

5. RESULTADOS E DISCUSSÃO

Para a obtenção da série temporal analisada neste estudo, foi empregado de modo não supervisionado o HMM para a identificação de regimes. A Figura 3 apresenta a evolução da série temporal da fatia de voos atrasados e cancelados entre os anos de 2011 e 2017.

A criação do modelo de HMM depende do número de regimes definidos a *priori*. Para obter um melhor ajuste, foram avaliados modelos com 2 a 6 regimes observando os critérios *Akaike Information Criterion* (AIC) e o *Bayesian Information Criterion* (BIC), representados graficamente na Figura 4.

Quanto ao número de parâmetros calculado para cada regime, constatou-se que os modelos com mais de três regimes seriam muito complexos (devido ao aumento do número de parâmetros) e sem muito ganho de informações (como observado no gráfico da Figura 4), logo o modelo com três regimes mostrou-se melhor ajustado.

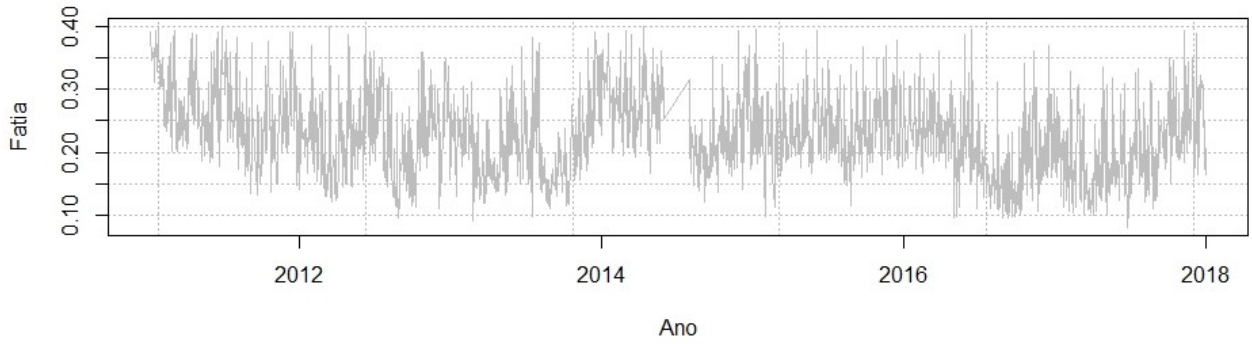

Figura 3. Série temporal da fatia diária de voos atrasados e cancelados

Foram estimadas a média e desvio padrão de cada regime detectado. A média da fatia diária de voos atrasados e cancelados do regime 1 foi de 29,7%, com desvio padrão de 4,6%. A média do regime 2 foi de 22,1%, com desvio padrão de 3,0%. Por fim, para o regime 3, a média foi de 15,6% e desvio padrão de 3,0%. Pode-se inferir assim que o regime 1 é composto pela fatia diária com maior índice de atrasos e cancelamentos de voos e o regime 3 é composto por dias menos congestionados, onde a fatia diária tem o menor índice de atrasos e cancelamentos de voos.

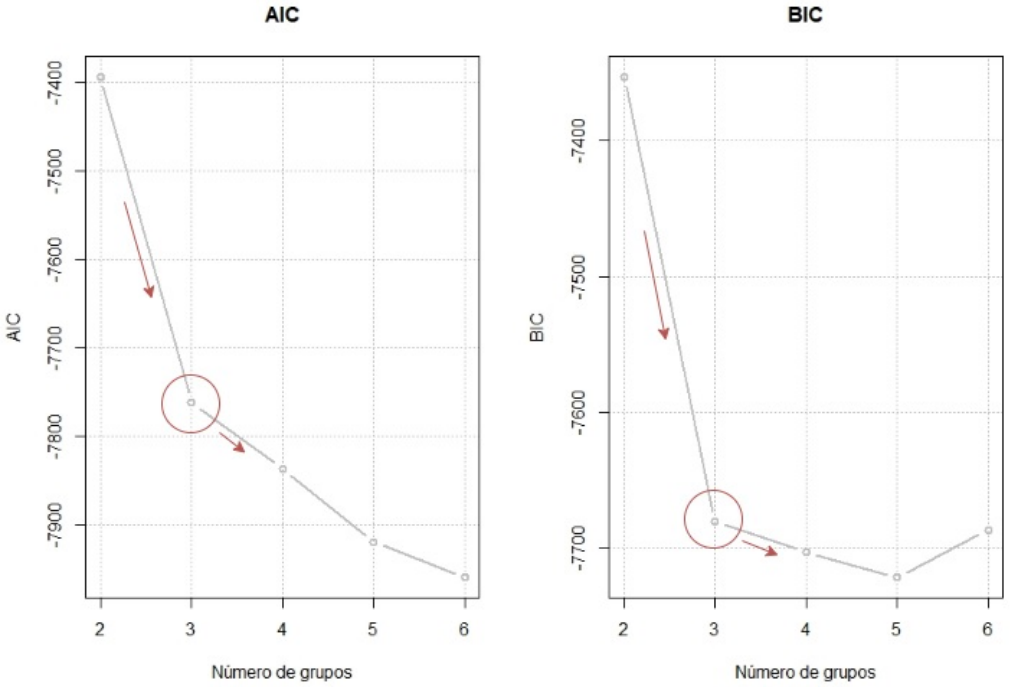

Figura 4. Gráfico Akaike Information Criterion (AIC) e Bayesian Information Criterion (BIC)

Para a simplificação da análise, os regimes detectados pelo modelo HMM foram definidos, de acordo com as medias da fatia diária de cada regime, como segue: Congestionamento alto (regime 1): 29,7%; Congestionamento médio (regime 2): 22,1%; Congestionamento baixo (regime 3): 15,6%. A Figura 5 apresenta o histograma da fatia de voos atrasados e cancelados para os regimes identificados pelo modelo de HMM.

Como observado na Figura 5, o regime 1 apresenta maior variação e maior índice da fatia diária de voos atrasados e cancelados. Isto implica a ocorrência de dias com congestionamento alto neste regime. O regime 2 apresenta a maioria das ocorrências de dias com índices da fatia diária de voos atrasados e cancelados próximos a 22%, o qual indica a ocorrência de dias com congestionamento médio. Já o regime 3 tem a maioria das ocorrências em torno 15%, o que implica em dias com congestionamento baixo.

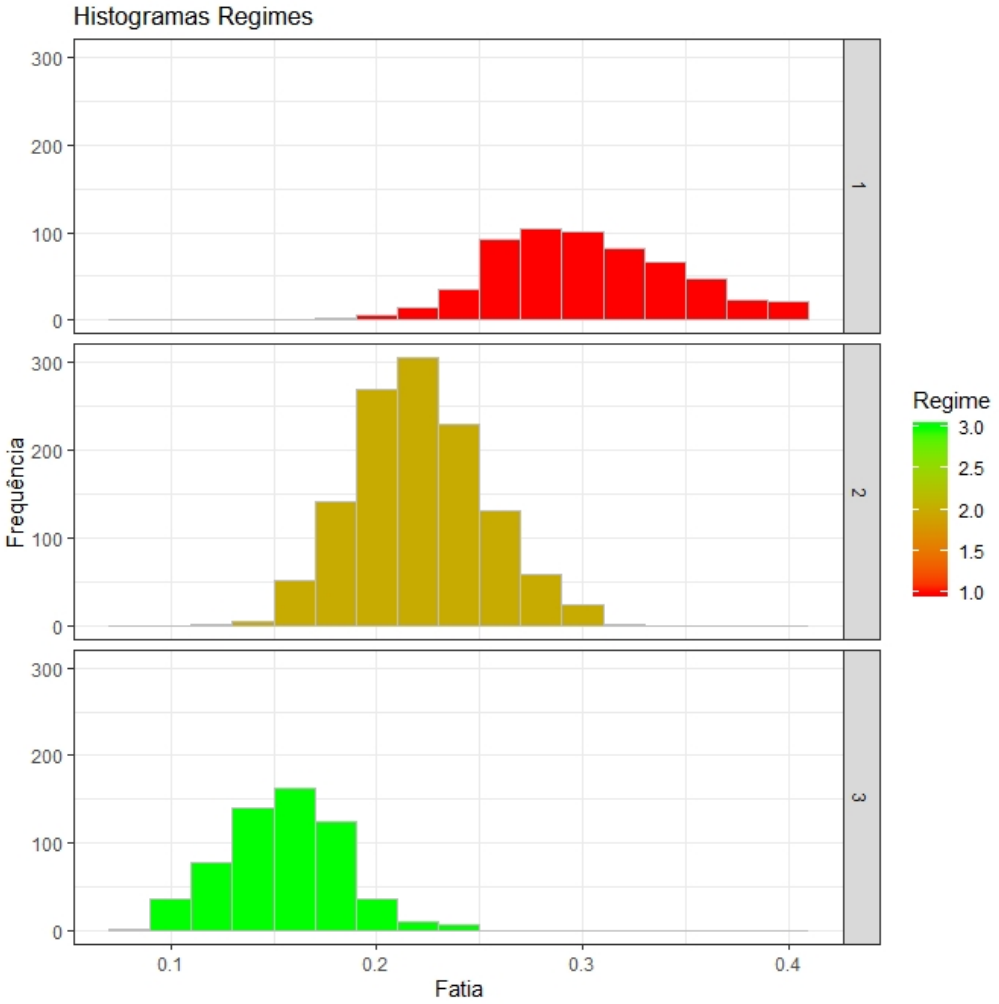

Figura 5. Histograma dos regimes de HMM

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.

Tabela 2 – Matriz de Transição dos regimes de HMM

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

Após os regimes detectados, dois modelos preditivos de classificação foram gerados, tendo em vista a previsão do regime no período de tempo *t+1*.

5.1. Árvores de Classificação e Regressão

O tamanho ideal da árvore, ou seja, o seu número ideal de nós terminais tendo em vista evitar superajuste do modelo CART, foi determinado utilizando a validação cruzada 10-*fold* considerando a regra "um desvio padrão" e pode ser observado na Figura 6.

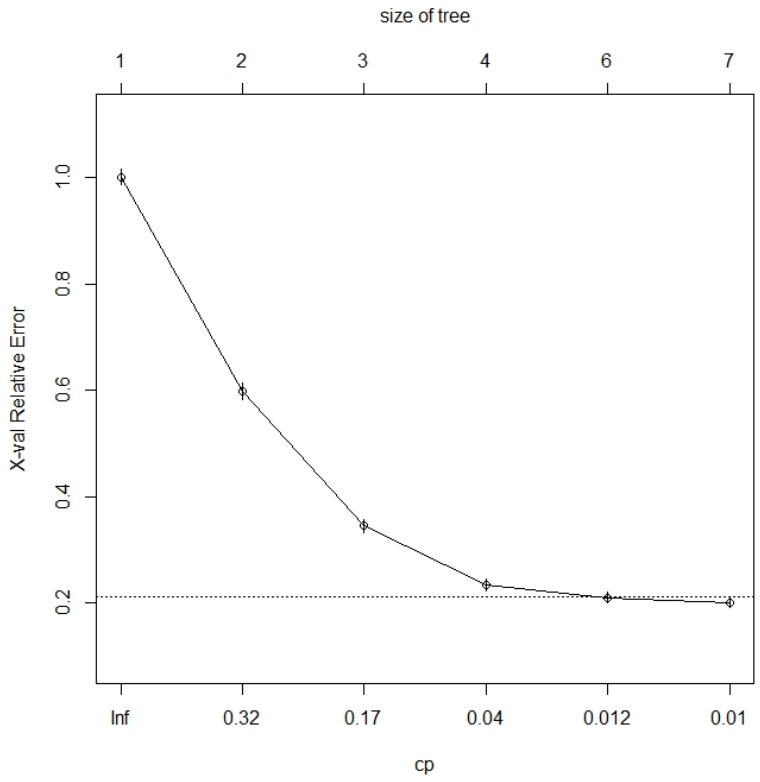

Figura 6. Erro de classificação da validação cruzada versus tamanho da árvore

Verifica-se, portanto, que o tamanho ideal indicado por esse procedimento é uma árvore com seis nós terminais.

A Figura 7 apresenta a árvore obtida após a poda, com seis nós terminais e cinco partições. De acordo com o gráfico, a variável "regime anterior" (período t) tem maior relevância e, portanto, forte influência na previsão de regimes no tempo *t+1*. O nó terminal 3 contém 698 observações, foi classificado como regime 1 e depende da variável "regime anterior". Desta forma, o dia que pertence ao regime 1 (congestionamento alto), tem a probabilidade de 88,1% de que o dia seguinte permaneça congestionado e pertença ao mesmo regime. A taxa de erro é de 11,9%.

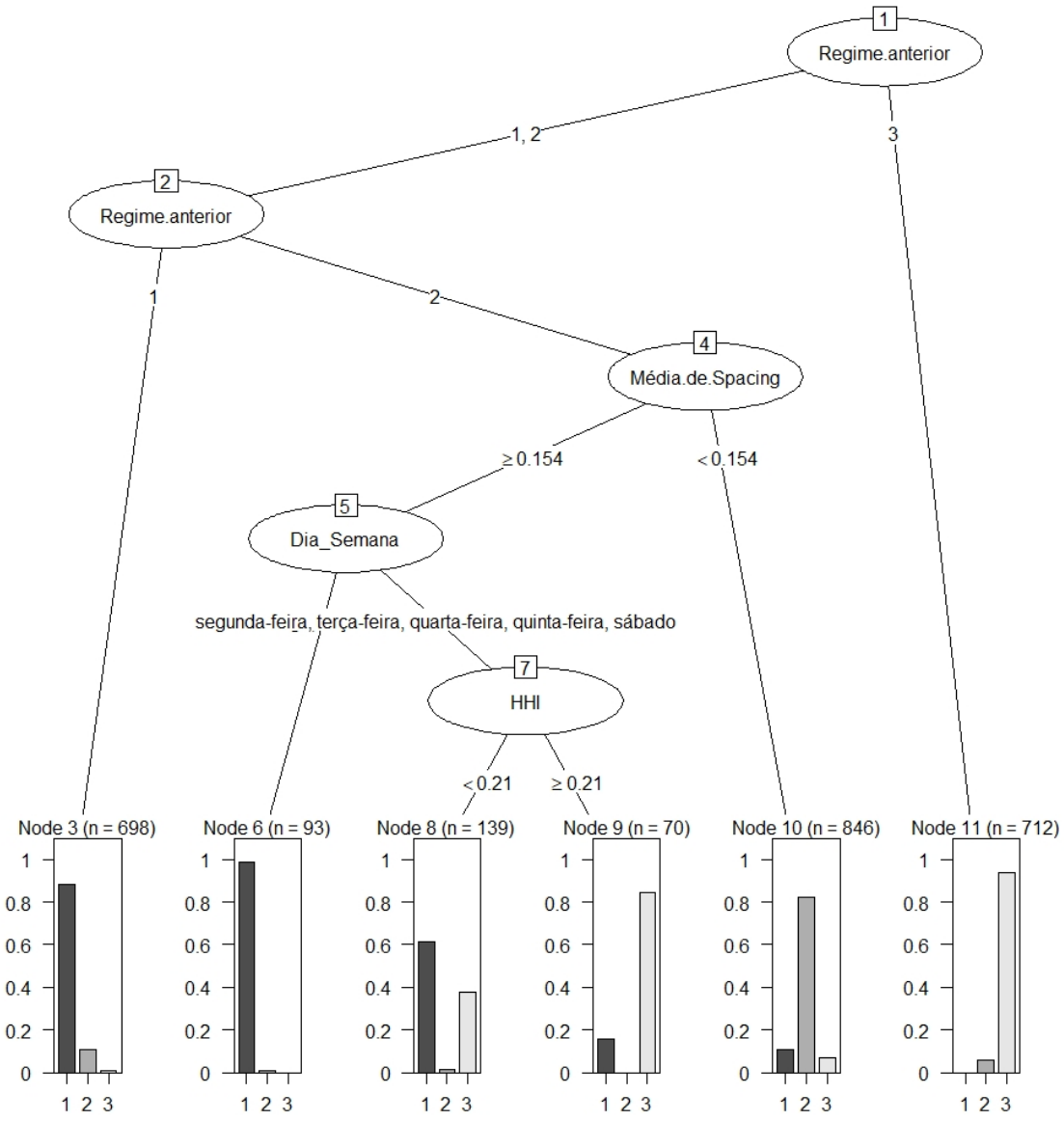

Figura 7. Árvore de classificação com seis nós terminais

O nó terminal 11 inclui 712 observações, foi classificado como regime 3 (congestionamento baixo). A probabilidade de que o dia seguinte permaneça no regime corrente é de 94,0%. A taxa de erro é de 6,0 %. Observe que a probabilidade do regime 3 (congestionamento baixo) ir para o regime 1(congestionamento alto) é muito pequena. Portanto, há pouca possibilidade de um dia com a maior taxa média, de atrasos e cancelamentos de voos, ocorrer após um dia com a menor taxa média de atrasos e cancelamentos de voos.

O nó terminal 10 abrange 846 observações, e apresenta alta probabilidade (82,4%) de pertencer ao regime 2 (congestionamento médio). Ocorre quando o regime prévio é 2 e a demanda de chegadas e partidas de voos programados é alta. Quando a demanda é alta os intervalos de tempo entre as chegadas e partidas de voos são menores, neste caso, "média de *Spacing*" < 0,154. É importante destacar que foram multiplicadas por 100 as variáveis "média de *spacing*" e "desvio padrão de *spacing*", devido ao padrão de arredondamento do pacote do R que gerou o modelo CART.

O nó terminal 6 contém 93 observações e foi classificado como regime 1 (congestionamento alto). Ocorre quando o regime anterior é 2, a demanda é baixa ("média de *Spacing*") ≥ 0,154) e o dia da semana é domingo ou sexta-feira. Deste modo, com o intervalo de tempo entre as chegadas e partidas de voos diários programados maior que aproximadamente dois minutos, o dia pertencente ao regime 2 (congestionamento médio) tem 98,9% de probabilidade de ir para o regime 1 (congestionamento alto). No nó terminal 8 (composto por 139 observações) a probabilidade de que o regime 1 ocorra é de 61,2% quando o regime anterior é 2, a demanda é baixa ("Média *Spacing*" ≥ 0,154), o dia da semana é de segunda-feira a quinta-feira ou sábado, e o mercado é menos concentrado (HHI < 0,21). Estes resultados estão de acordo com os resultados apresentados por Scarpel e Pelicioni (2018). Assim, em dias com uma demanda maior (períodos de pico) e mercado menos concentrado (mais companhias aéreas operando), são esperados dias mais congestionados.

O nó terminal 9 contém 70 observações e probabilidade de 84,3% de pertencer ao regime 3. Diferentemente do nó terminal 8, no nó terminal 9 é o mercado mais concentrado (HHI ≥ 0,21), e as companhias aéreas tendem a internalizar os atrasos. De acordo com Scarpel e Pelicioni (2018), são esperadas menores taxas de atrasos quando o aeroporto é mais concentrado e a demanda é menor. Logo, o nó terminal 9 tem probabilidade acima de 80,0% de ocorrer no regime 3 (congestionamento baixo), com o mercado mais concentrado e a demanda menor. Quanto à avaliação de desempenho do CART, o conjunto de treino apresentou acurácia de 86,6% e o conjunto de teste de 88,3%.

5.2. Florestas Aleatórias

Para gerar o modelo RF é necessário definir o parâmetro *mtry*. O valor padrão para esse parâmetro, como este modelo que possui oito variáveis, é $\sqrt{8}$. Logo, o valor padrão (2) é satisfatório, pois é o maior número inteiro cujo quadrado é menor ou igual a 8, entretanto, o valor de *mtry* foi otimizado testando valores de 1 a 4. O menor valor do erro OOB foi obtido com *mtry* igual a 2. O modelo gerado pelo RF teve acurácia estimada de 88,9% para o conjunto de treino e 89,7% para o conjunto de teste.

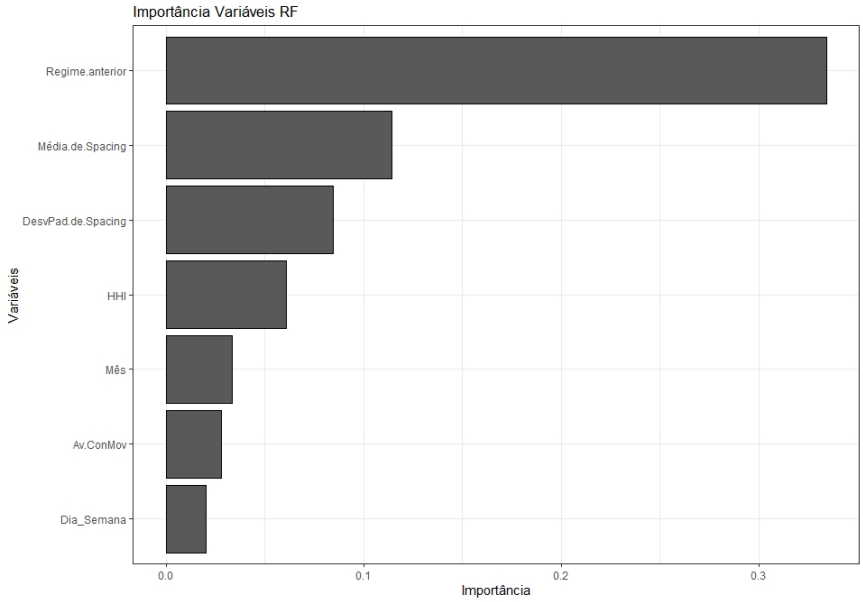

Figura 8. Importância das variáveis no modelo RF

A Figura 8 apresenta a importância das variáveis de acordo com o modelo RF. Apesar de RF não oferecer possibilidade de interpretação, observa-se que a variável "regime anterior" é a de maior importância, o que está em concordância com CART.

Os modelos, tanto o CART quanto o RF, apresentaram bom desempenho de classificação no modelo de previsão gerado para antecipar a ocorrência de dias congestionados no Aeroporto Internacional de São Paulo/Guarulhos com o período de um dia de antecedência.

6. CONCLUSÃO

O objetivo deste estudo foi criar um modelo para antecipar a ocorrência de dias congestionados no Aeroporto Internacional de São Paulo/Guarulhos. Para isto, em um primeiro momento, foi gerado um modelo de HMM utilizando a série temporal fatia diária de voos atrasados e cancelados. O ajuste do melhor modelo ocorreu considerando os critérios AIC e BIC. Foram identificados três regimes e definidos de acordo com a taxa média de atrasos e cancelamentos de voo como segue: regime 1 (congestionamento alto, 29,7%); regime 2 (congestionamento médio 22,1%); e regime 3 (congestionamento baixo 15,6%).

Posteriormente, os classificadores CART e RF geraram o modelo de previsão pretendido. Por meio do modelo do CART foram identificadas as variáveis determinantes para dias muito congestionados e como estão combinadas. A árvore gerada possui seis nós terminais, é composta pelas variáveis "regime anterior", "média de *spacing*", "dia da semana" e "HHI". Os resultados obtidos são consistentes com os resultados presentes na literatura. O modelo de RF estimou a importâncias das variáveis, em que assim como no modelo de CART, "regime anterior" é a variável mais significativa.

Os modelos de classificação empregados se mostraram adequados e com boa acurácia para fazer a antecipação da ocorrência de dias congestionados no Aeroporto Internacional de São Paulo/Guarulhos com um dia de antecedência. Algumas limitações foram identificadas no decorrer deste estudo: (i) as variáveis obtidas as partir do resultado do modelo de HMM foram tidas como dados reais, assim não foi considerada a incerteza dos rótulos; (ii) o modelo de previsão gerado faz a antecipação da ocorrência de dias congestionados somente para o instante de tempo t+1. Para trabalhos futuros, propõe-se investigar como outras variáveis (por exemplo, "spacing" considerando horários de pico e estações do ano) influenciariam o modelo de previsão de dias congestionados visando ampliar o horizonte da previsão.

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.