JAFF GUSTAVO CUNHA

MODELOS PARA ANÁLISE DE ATRASOS DE VIAGENS AÉREAS NACIONAIS

SÃO PAULO
2020

JAFF GUSTAVO CUNHA

MODELOS PARA ANÁLISE DE ATRASOS DE VIAGENS AÉREAS NACIONAIS

Trabalho de Formatura apresentado à Escola Politécnica da Universidade de São Paulo para obtenção do Diploma de Engenheiro de Produção

Orientadora: Profa. Dra. Celma de Oliveira Ribeiro

SÃO PAULO
2020

Catalogação-na-publicação

Cunha, Jaff

MODELOS PARA ANÁLISE DE ATRASOS DE VIAGENS AÉREAS NACIONAIS / J. Cunha -- São Paulo, 2020.

86 p.

Trabalho de Formatura - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Produção.

1.Regressão Logística 2.Clusters 3.Aviação 4.Seguros I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Produção II.t.

À minha avó Marcionília de Oliveira (in memoriam), base de nossa família e que sonhava com este momento. Sei que do alto sorri para mim.

AGRADECIMENTOS

À minha mãe, Sueli Cunha, por todo o carinho e sacrifícios dispensados a mim ao longo destes anos de vida. E às minhas irmãs, pelo cuidado e incentivo à minha formação.

À minha parceira e namorada, Keiko Matsuba, por todo o suporte e motivação recebidos nos últimos anos, essenciais para a conclusão deste trabalho.

Às minhas amigas e amigos, pelas conversas sinceras e por acreditarem mais em mim do que eu mesmo.

Ao projeto "Xorume", por me acompanharem nos dias mais difíceis e me fazerem rir quando mais precisei.

À Poli Júnior, por toda a experiência profissional e interpessoal que me possibilitaram e por me despertarem o desejo pela programação.

Aos professores e funcionários da Escola Politécnica da USP, que possibilitaram minha excelente formação.

Aos meus colegas de trabalho, Daniel e Pedro, pela sugestão de tema e ajuda essencial para levá-lo ao fim.

À Professora Celma, pela orientação e apoio recebidos durante o desenvolvimento do meu tema.

"A mind that is stretched by a new experience can never go back to its old dimensions."

Oliver Wendell Holmes Jr

RESUMO

O setor de transporte aéreo é fortemente marcado por atrasos e cancelamentos de voos, por fatores meteorológicos, logísticos e técnicos, impactando negativamente os planos de milhões de viajantes todos os anos. O mercado de seguros, presente em diversas esferas da vida comum, ainda não atua com todo seu potencial no Brasil. Partindo de inspirações externas e a motivação de se vender um projeto para empresas seguradoras, este trabalho visou desenvolver um modelo de previsão de atrasos em viagens aéreas a partir de variáveis simples e disponíveis publicamente pela Agência Nacional de Aviação (ANAC). Para atingir seu objetivo, os dados foram tratados e analisados em diferentes linguagens e plataformas. Foram realizados agrupamentos em *clusters* com o método *K-Medoids* para lidar com a natureza categórica das variáveis e, por fim, regressões logísticas binária com aquelas de maior significância. O resultado obtido foram fórmulas preditivas de atraso e tabelas comparativas entre as previsões e ocorrências.

Palavras-chave: Regressão logística, Voos nacionais, Aviação, Modelo de Risco.

ABSTRACT

The air transport sector is strongly marked by flight delays and cancellations, due to meteorological, logistical and technical factors, negatively impacting the plans of millions of travelers every year. The insurance market, present in several spheres of common life, still does not operate to its full potential in Brazil. Based on external inspirations and the motivation to sell a project to insurance companies, this work aimed to develop a model for predict flight delays based on simple and publicly available variables provided by the National Aviation Agency (ANAC). In order to achieve the objective, the data were processed and analyzed in different languages and platforms. Clustered groupings were performed with the K-Medoids method to deal with the categorical nature of the variables and, finally, binary logistic regressions with those of greater significance. The result obtained was formulas predictive of delay and comparative tables between forecasts and occurrences.

Key-words : Logistic regression, National flights, Aviation, Risk Model.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

1.1 Motivações do estudo

Uma das constantes preocupações dos diversos atores da Economia é o gerenciamento de riscos. Ao longo dos anos, diversos métodos de gerenciamento foram desenvolvidos e um deles se consolidou como opção viável e comum em praticamente todos os setores e casos: as apólices de seguro. Trata-se de um contrato entre duas partes, em que uma delas, a seguradora, protege a outra parte, o segurado, de perdas financeiras em situações pré-acordadas. Em resumo, é estabelecido que os segurados pagam um valor conhecido como "prêmio" à seguradora, que por sua vez garante compensações financeiras na ocasião de um sinistro – evento que causa prejuízo material ao segurado -, a materialização de um risco coberto pela apólice (ROMANO, 2019).

Todas as diferentes indústrias possuem certo grau de risco comercial, incluso o setor de aviação. De acordo com Guzhva *et al*. (2019) em *Aircraft Leasing and Financing*, os eventos associados a riscos enfrentados por empresas e proprietários de aeronaves incluem: acidentes de aeronaves, volatilidade dos preços de combustíveis, ataques terroristas, desastres naturais e instabilidade política. Com isso, é visível que os contratos de seguro são uma importante face da aviação.

Além desses riscos, há um muito recorrente: os atrasos de voos. Na aviação, atraso é comumente definido como a diferença entre os horários agendado e real de partida ou chegada da aeronave (WIELAND, 1997). Todos os voos comerciais possuem uma previsão de horários de partida e chegada, que devem ser estritamente seguidos, pois todo aeroporto possui limite para acomodação de aeronaves e quantidade de pistas. Os atrasos geram problemas para as companhias aéreas e aeroportos, por meio de aumento de custos e dos tempos de viagem. Para o passageiro, parte mais vulnerável aos atrasos, adiciona-se que muitas viagens dependem de escalas - viajantes realizam voos encadeados, um após o outro, para atingir seu destino -, sendo de grande impacto a demora de um dos voos. Os transtornos se estendem a uma menor estadia no destino, maiores gastos com acomodação, adiamento ou cancelamento de reuniões de negócios, entre outros. Estes riscos para o usuário físico do sistema também são merecedores de gerenciamento, sendo possível o uso de apólices de seguros para os eventos de atraso ou cancelamento dos voos.

Os diversos riscos na aviação justificam o estudo de métodos como o de seguro para amenizar os danos para empresas e clientes. Em especial, o risco de atraso, que afeta mais fortemente o consumidor, vem sendo estudado nos últimos anos com resultados relevantes. Um modelo preditivo de atrasos de chegada ao aeroporto JFK dos EUA utilizou redes neurais e obteve bons resultados com o uso de variáveis como: previsão e real partida e chegada, aeroporto de origem e justificativas do atraso (KHANMOHAMMADI et al., 2016).

No Brasil, um estudo que utilizou o método f*requent pattern mining* em busca de padrões recorrentes, observou que os voos no país possuem dificuldade em se recuperar de atrasos anteriores, um dos fatores para o efeito cascata, em que os atrasos propagam-se pela rede de voos que conecta os aeroportos, além de condições meteorológicas levarem ao aumento de até 216% nos atrasos (STERNBERG et al., 2016).

Inexistem seguros nacionais para indenização automática em caso de pequenos atrasos, na escala de frações de hora. Seguros de indenização automática, vinculados aos horários de partida dos aviões, já são ou foram fornecidos por empresas de outras países, como o serviço Fizzy da empresa AXA na França. O seguro era atrelado à uma blockchain, que guarda todos os horários previstos e executados de saída e chegada dos aviões, utilizando os dados oficiais registrados pela agência de aviação nacional. Identificada a discrepância, uma transferência de dinheiro para o segurado é feita automaticamente dentro de prazo estipulado (AXA, 2017).

Além disso, o Brasil conta com leis indenizatórias para atrasos e cancelamentos de voos insuficientes para muitos clientes. Apenas atrasos superiores a 4 horas obrigam a companhia aérea a oferecer o reembolso integral ou reacomodação do passageiro em outro voo (ANAC, 2010). Tal restrição gera transtornos àqueles que têm compromissos inadiáveis no local de chegada, como empresários em viagem à trabalho. Clientes também têm dificuldade para obter indenização por atrasos e cancelamentos de voos, uma vez que os procedimentos são burocráticos e geralmente requerem a contratação de advogados para vencer a causa (ANAC, 2018).

1.2 Objetivos

O trabalho almeja desenvolver um modelo preditivo de atrasos de voos nacionais. Esse modelo tem o propósito de servir de base para a criação de uma ferramenta de

precificação de seguro a partir de probabilidades de atraso dos voos segurados para companhias de seguro, sendo um diferencial de outros estudos da área.

Busca-se um modelo que não necessite de dados de difícil previsão com antecedência, como os meteorológicos, pois o uso esperado pela seguradora é a venda do bilhete de seguro junto à passagem para o comprador. Tal venda pode ocorrer com dias, semanas ou meses de antecedência à viagem, sendo inviável o uso de previsões atreladas às condições meteorológicas.

A revisão da bibliografia indica não somente um ferramental matemático para construção do modelo, mas diversos caminhos alternativos para atingir o objetivo.

2. REVISÃO BIBLIOGRÁFICA

2.1 A evolução da atividade seguradora no Brasil e o seguro na aviação

A Sociedade de Análise de Riscos (SRA) define 'risco' como a consequência (efeitos, implicações) de uma atividade futura em relação a algo que os humanos valorizam, onde há pelo menos um resultado considerado negativo ou indesejável (AVEN *et al.*, 2015), já a ISO 31000 designa 'risco' como efeito da incerteza nos objetivos. Os métodos e técnicas de análise de risco são usados, atualmente, na maioria dos setores da sociedade, a SRA possui diversos grupos de especialidades para representar uma área de relevância relacionada à análise de riscos como: Materiais e tecnologias avançados, Resposta à dose, Avaliação de risco ecológico, Engenharia e infraestrutura, Avaliação de exposição , Análise de risco microbiano, Saúde e segurança ocupacional, Política e lei de risco e Segurança e defesa (SRA, [2019]).

Para as organizações, a importância em evitar os danos dos riscos e saber aproveitar as oportunidades está na probabilidade de sobreviver e crescer de forma sustentável. O mundo ficou mais dinâmico, as tomadas de decisões devem ser feitas de maneira rápida e o número de agentes e fatores críticos aumentou, desse modo, a previsibilidade das ações ficou mais difícil (OLIVA, 2016). Os riscos podem ser gerenciados com o propósito de auxiliar a tomada de decisões, de forma a maximizar os ganhos e reduzir os prejuízos (GRACE et al., 2015). O gerenciamento de riscos se torna, então, uma ferramenta essencial para o planejamento e o processo de tomada de decisões.

Quadro1. Principais riscos de empresas brasileiras

Natureza dos riscos	**Riscos**
Riscos Estratégicos	1) redução da demanda; (2) problemas com clientes e insolvência; (3) problemas com fusões e aquisições; (4) pressões por preço; (5) regulatório;
Riscos Operacionais	6) descontrole dos custos operacionais; (7) problemas contábeis; (8) problemas de logística; (9) não-conformidade com as regulações; (10) elevação dos custos;
Riscos Financeiros	(11) altas taxas de inadimplência e altas taxas de juros; (12) estratégias financeiras fracas; e
Riscos Externos	(13) crise no setor de atuação; (14) problemas econômicos no país sede; (15) vulnerabilidades legais.

Fonte: DOI (2017) baseado no estudo da Deloitte (2005).

A execução do gerenciamento dos riscos é baseada em dois pilares compostas pelo controle dos riscos, que se baseia na organização de um programa de prevenção de perdas, reduzindo a severidade e frequencia dos acidentes, e do financiamento dos riscos remanescentes. Essa última parte consiste na tranferência total ou parcial para as seguradoras que assumem a responsabilidade pelos riscos econômicos dos seus segurados (MOREIRA, 2020). Dessa forma, obtêm-se uma proteção eficaz, reduzindo ou eliminando de maneira efetiva a maioria dos riscos acidentais, desenvolvendo o mercado de seguros.

Essa estrutura financeira recebe um valor que o segurado paga à seguradora, denominado prêmio, para assumir os riscos que o indivíduo pode sofrer; caso o segurado passe por algum infortúnio, este é indenizado pelas seguradoras de acordo com o contrato estabelecido. O mercado de seguros visa minimizar ou transferir o risco decorrente de eventos aleatórios geradores de danos. (SILVA e CHAN, 2015).

O início da atividade seguradora no Brasil ocorreu em 1808 com a abertura dos portos ao comércio internacional, sendo a "Companhia de Seguros BOA-FÉ" a primeira

sociedade de seguros a funcionar no país com o objetivo de operar no seguro marítimo; a atividade seguradora era regulada pelas leis portuguesas (SUSEP, 1997).

Em 1850 o "Código Comercial Brasileiro" da Lei n° 556, de 25 de junho de 1850 foi promulgado, incentivando o aparecimento de inúmeras seguradoras que passaram a atuar não só com o seguro marítimo, bem como com seguro terrestre. O setor se expandiu e, por volta de 1862, as primeiras empresas de seguros estrangeiras fixaram as suas filiais em terras brasileiras (SUSEP, 1997).

Em 1966, foram reguladas todas as operações de seguros e resseguros e instituído o Sistema Nacional de Seguros Privados, constituído pelo Conselho Nacional de Seguros Privados (CNSP); Superintendência de Seguros Privados (SUSEP) e Instituto de Resseguros do Brasil (IRB); sociedades autorizadas a operar em seguros privados e corretores habilitados, por meio do Decreto-lei n° 73, de 21 de novembro de 1966 (SUSEP, 1997).

Conforme consta no “Relatório de análise e acompanhamento dos mercados supervisionados” da SUSEP (2018), o mercado segurador apresentou significativa expansão nos últimos anos, refletido em uma crescente participação no Produto Interno Bruto (PIB) do país, saindo de um nível de 2,59% em 2003 e alcançando o patamar de 3,77% em 2017, mostrando-se um importante ramo da economia.

De acordo com a classificação adotada pela SUSEP (BRASIL, 2016), o segmento de Seguros Gerais (seguros de danos) é composto pelos seguintes grupos: patrimonial, riscos especiais, responsabilidades, cascos, automóvel, transporte, riscos financeiros, crédito, pessoal, coletivo, habitacional, pessoas individual, rural, marítimos, aeronáuticos, micro-seguros e outros.

O seguro-viagem é um dos seguros recomendado para as pessoas que necessitam viajar a lazer ou a negócios, os planos podem incluir extravio de bagagem, assistência médica, repatriação, assistência jurídica, assistência odontológica, medicamentos, entre outros, sendo válidos apenas dentro do período estipulado e dentro do destino informado. Entretanto, somente 15% a 20% dos brasileiros contratam o seguro viagem em viagens nacionais, aumentando para 35% em viagens internacionais (APENAS... 2017). Outro seguro interessante para aqueles que viajam frequentemente é o seguro para atrasos de voos, seguradoras estrangeiras como Visitors Coverage Inc, Swiss Re e Alpha Travel Insurance oferecem esse tipo de serviço para atrasos acima de 12 horas.

O Brasil possui um importante sistema de aviação comercial que contém mais de 100 aeroportos, somente no primeiro trimestre de 2018 foram transportados 29,3 milhões de passageiros pagos em voos nacionais e internacionais de acordo com a ANAC (2018). Este grande volume de passageiros em milhares de voos gera uma preocupação com os diversos riscos associados ao setor. Atrasos, necessidade de manutenção, acidentes, condições meteorológicas e desastres naturais são eventos que causam danos financeiros e/ou físicos. Embora muitos desses sejam raros, os enormes custos materiais e imateriais envolvidos obrigam que empresas e governos tomem precauções.

Uma delas, voltada para a proteção das companhias aéreas, é o Código Brasileiro de Aeronáutica, que obriga, no Art. 281º da Lei nº 7.565, de 19 de dezembro de 1986, a contratação do Seguro RETA (Seguro de Responsabilidade Civil do Transportador Aéreo) pelos operadores de aeronaves civis, com o objetivo de cobrir riscos aos tripulantes e viajantes, pessoal técnico a bordo, às pessoas e bens na superfície e ao valor da aeronave (BRASIL, 1986).

Do lado do consumidor, o transporte aéreo se tornou cada vez mais acessível, com sites que facilitam a busca de passagens a preços módicos, assim como os programas de fidelidade e milhas. As preocupações dos usuários dos serviços de aviação se diferem daquelas apresentadas pelas companhias aéreas. Além dos riscos de acidentes, os clientes preocupam-se com fatores relacionados à execução do voo e o serviço prestado como um todo. A plataforma do governo, registrou 5.801 reclamações em relação às empresas aéreas; dessas reclamações, 98,3% correspondem às empresas brasileiras e 1,7% são de empresas estrangeiras (ANAC, 2018).

Um boletim de monitoramento do consumidor elaborado pela ANAC (2018) classifica os problemas relatados pelos usuários em onze temas: oferta e compra; alteração pelo passageiro; alteração pela empresa aérea; check-in e embarque; execução do voo; transporte de bagagem; reembolso; reclamações contra valores e regras do contrato; assistência ao PNAE; programas de fidelidade; e outros. O tema ‘execução de voos’, correspondente a 13,10% das queixas totais, inclui as reclamações por atrasos, cancelamentos, interrupção do serviço, perda de conexão e preterição. Os atrasos nos vôos são monitorados pela ANAC, oficialmente, eles são medidos por voo no aeroporto de partida,

considerando a hora em que os motores começam a funcionar, todos os voos com diferenças entre os horários programados e reais devem relatar uma causa (ANAC, 2015.).

De acordo com a resolução nº 400, de 13 de dezembro de 2016 (ANAC, 2016) sobre as Condições Gerais de Transporte Aéreo; no caso de atrasos e cancelamentos, as empresas são obrigadas a:

- Manter o passageiro informado a cada 30 minutos quanto à previsão de partida dos voos atrasados;
- Informar imediatamente a ocorrência do atraso, do cancelamento e da interrupção do serviço;
- Oferecer gratuitamente, de acordo com o tempo de espera, assistência material. Se for superior a uma hora: facilidades de comunicação; superior a duas horas: alimentação, de acordo com o horário, por meio do fornecimento de refeição ou de voucher individual; e quando houver atraso de voo superior a 4 horas, cancelamento ou preterição de embarque: oferecer reacomodação, reembolso integral e execução do serviço por outra modalidade de transporte, cabendo a escolha ao passageiro.

A resolução também esclarece que, no caso de preterição de embarque, que ocorre quando a empresa aérea precisa negar embarque a passageiros que compareceram para viajar, mesmo que cumprindo todos os seus requisitos de embarque, a empresa deverá em um primeiro momento, procurar por voluntários que aceitem embarcar em outro voo, mediante a oferta de vantagens (dinheiro, passagens extras, milhas, diárias em hotéis etc), negociadas livremente com o passageiro. No entanto, caso haja um número insuficiente de passageiros que aceite as vantagens oferecidas e um passageiro tenha seu embarque negado, a empresa deverá pagar a ele, imediatamente, uma compensação financeira, no valor correspondente a 250 DES, no caso de voos domésticos, e 500 DES, para voos internacionais. O Direito Especial de Saque (DES) é uma moeda do Fundo Monetário Internacional, cujo preço varia diariamente.

Além da compensação financeira, o passageiro que foi impedido de embarcar deverá receber alternativas de reacomodação em outro voo da própria empresa ou de outra, reembolso do valor total pago e assistência material. Ressalta-se que nos casos de atraso, cancelamento ou preterição a empresa poderá suspender a prestação da assistência material para proceder ao embarque imediato (ANAC, 2016). Caso uma companhia não cumpra as

exigências estabelecidas por lei, o passageiro pode registrar queixa. Para reivindicar indenizações por danos morais e/ou materiais, o cliente deve entrar em contato com os órgãos de Defesa do Consumidor ou o Poder Judiciário (Ministério da Infraestrutura, 2020).

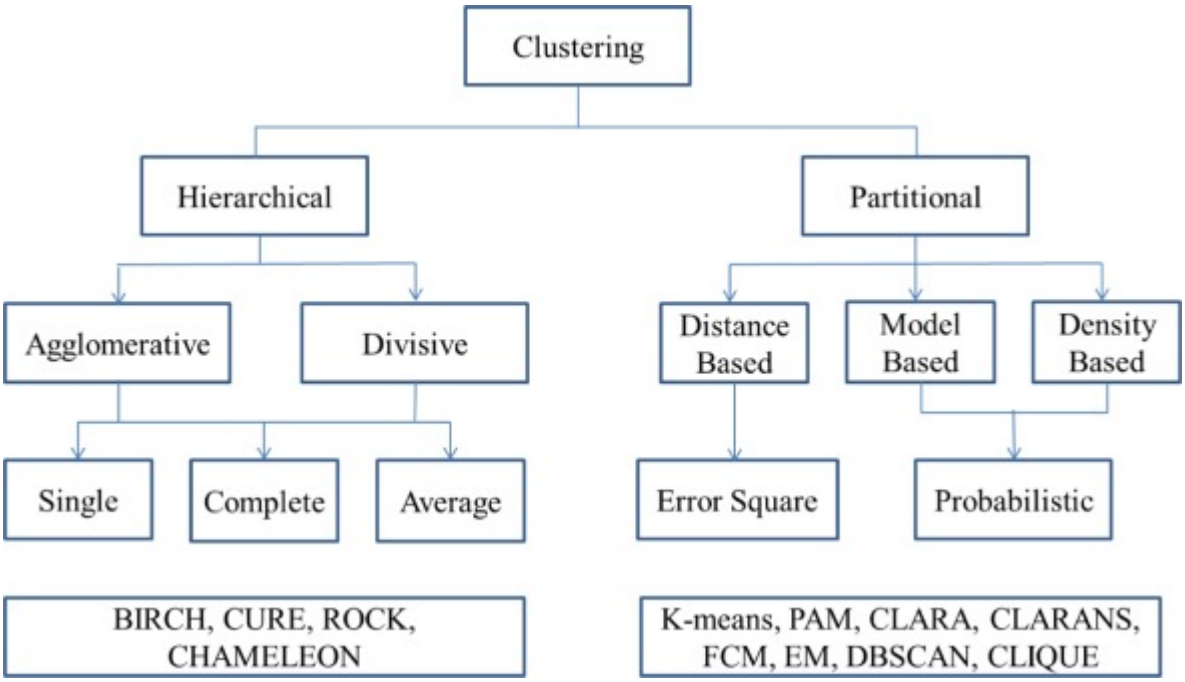
2.2 Agrupamentos por *clusters*

É extremamente importante que os dados sejam analisados com muita cautela para a criação de um modelo preditivo, sendo essencial verificar a existência de padrões nos atrasos dos voos.

O agrupamento por *clusters* divide os padrões de dados em subconjuntos, de maneira que padrões semelhantes sejam agrupados. Dessa forma, um *cluster* é um conjunto de dados cuja as entidades são comparativamente mais similar às entidades desse grupo do que as dos outros grupos (SAXENA *et al.*, 2017). A clusterização é extensamente utilizada para segmentação de imagens, reconhecimento de padrões, quantização de vetores (VQ), aproximação de funções e mineração de dados. Como uma técnica de classificação não supervisionada, o agrupamento identifica estruturas inerentes em um conjunto de objetos com uso de uma métrica de similaridade. Os métodos de clusterização podem ser baseados em modelo estatístico de identificação ou aprendizagem competitiva (DU, 2010).

Na maior parte dos casos, o número de clusters a serem formados é especificado pelo usuário. Existindo apenas dados de tipo numérico para representar características dos padrões em um grupo, a única maneira de extrair qualquer informação da relação entre padrões é fazer uso da aritmética numérica. As características dos objetos são representadas por valores numéricos e uma abordagem comum para definir similaridade é tomar como medida a distância entre os padrões: quanto menor a distância entre dois objetos, maior a semelhança; quanto menor, maior a não similaridade (SAXENA *et al.*, 2017). Os principais métodos de clusterização estão representados na figura abaixo:

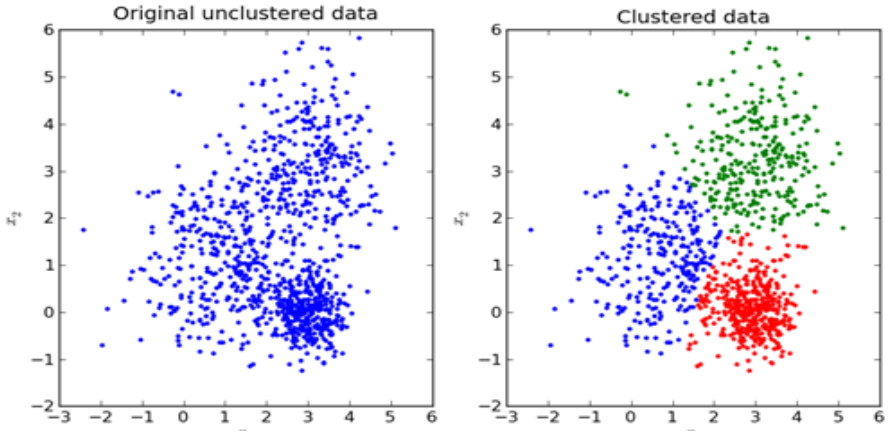
Figura 1. Principais métodos de clusterização.



Fonte: SAXENA et al. (2017).

Um dos algoritmos mais comuns para a formação de clusters é o K-means, um tipo de método por particionamento com uso de centroides. Para que o K-means retorne resultados válidos, é preciso que os dados de *input* sejam variáveis numéricas. Centroide, neste algoritmo, trata-se de um ponto com os valores médios do cluster. Esse ponto pode ser ou não pertencente ao conjunto de dados e graficamente representa o centro gravitacional do cluster (LIKAS *et al.*, 2003).

Figura 2. Exemplo de clusters gerados por K-Means por SAKHALKAR *et al*. (2015)



É possível a realização de agrupamentos por *clusters* também para variáveis não numéricas – as categóricas e ordinais. Variáveis categóricas possuem mais de uma categoria, mas sem ordenamento intrínseco entre elas. Também são conhecidas como variáveis nominais. Um clássico exemplo é o gênero, em que há duas categorias (masculino e feminino), mas sem uma ordem intrínseca deles. Já as variáveis ordinais são semelhantes às categóricas, mas com uma ordenação clara das variáveis (INSTITUTE .... 2020).

Há autores como GUPTA *et al.* (1999) que apontam formas de se transformar variáveis categóricas em vetores, que permitem a aplicação do algoritmo, mas com um passo de preparação bastante complexo. O algoritmo *K-means* gera *clusters* com centroides que possuem valores médios, não necessariamente pontos do conjunto de dados. Isso significa que transformar categorias em números inteiros (empresa "A" em 1, empresa "B" em 2, etc) levará a resultados sem significado real (por exemplo, *cluster* com centroide com a variável empresa de valor 1,3).

Uma alternativa para as variáveis categóricas é o uso de K-medoids, um algoritmo com o mesmo objetivo de agrupamento, mas que trabalha clusterizando com *medoids*, que são pontos como os centroides, mas que também fazem parte do conjunto de dados. Os *medoids* são os elementos mais próximos ao centro de gravidade do *cluster*, portanto sempre têm significado real (PARK; JUN, 2009).

Há diversas medidas de distância para observações de variáveis categóricas binárias. As mais utilizadas são a distância de *Hamming* e o coeficente de *Jaccard* (DUDA; HART, 1973). Considere-se que duas observações possuam variáveis que assumem o valor 1 se certa característica está presente e 0 no caso contrário. O quadro (2) a seguir resume o observado nas variáveis da matriz X:

$$X = \begin{pmatrix} 1 & 1 & \cdots & 0 \\ 0 & 1 & & 0 \end{pmatrix} \qquad (1)$$

Quadro 2. Observações resumidas

Valor assumido	0	1	Total
0	W	x	w + x
1	Y	z	y + z
Total	w + y	x + z	m =w + x + y + z

Fonte: Adaptado de Matos (2007)

Os valores w e z representam o número de variáveis que concordam quanto à existência ou inexistência da característica analisada nas observações. Os valores x e y indicam o número de variáveis em discordância. A distância de *Hamming* pode ser calculada como $\frac{w+y}{m}$.

Variáveis não binárias exigem o uso de variáveis indicadoras (*dummies*). Cada valor assumível pelas variáveis, chamados de categorias, será transformado em uma nova variável, indicando presença ou ausência daquele valor. Assim, o coeficiente de similaridade simples pode ser diretamente usado. Essa abordagem leva ao aumento do número de variáveis e pode inviabilizar o tratamento para um grande número de variáveis e categorias iniciais (EVERITT, 1993).

Para exemplificar no contexto do trabalho, se entre duas linhas de voos a única diferença em variáveis categóricas é a rota, sua distância de *Hamming* é de 1 sobre o número de variáveis totais (HARIKUMAR; PV, 2015).

2.3 Regressão Logística

A técnica de regressão logística, apesar de antiga, passou a ganhar maior visibilidade, melhorias e aplicações após 1950, com os trabalhos de Cox e Snell (1989) e Hosmer e Lemeshow (2000), que originaram testes para os resultados modelados amplamente utilizados desde então. A técnica caracteriza-se por descrever a relação entre uma variável dependente qualitativa binária e um conjunto de variáveis independentes qualitativas. Apesar da regressão logística incialmente ter sido mais utilizada para a área médica, sua eficiência viabilizou a implementação em outras diversas áreas de estudo, com grande popularização entre os que faziam uso de modelos tradicionais de regressão. Assim, a regressão logística tornou-se uma das mais poderosas ferramentas para a análise de variáveis dicotômicas (CRAMER, 2003).

Os modelos de regressão são técnicas que possibilitam explicar a relação entre uma variável dependente e um grupo de variáveis independentes. De acordo com Hair *et al* (1998), a transformação realizada com a variável dependente permite o cálculo direto da probabilidade de ocorrência de um evento em análise. A regressão logística possui especificidades que diferenciam a técnica de demais modelos de regressão tradicionais: a

variável dependente é qualitativa binária (atraso ou não atraso, realizado ou cancelado, nacional ou internacional, etc) e segue uma distribuição de Bernoulli (ANÁLISE... 2020).

Assim, tome-se uma variável aleatória definida como x_i={1,0}, com distribuição de Bernoulli, cuja função de probabilidade é:

$$f(x,p) = p^x(1-p)^{1-x} \text{ para } x \in \{0,1\} \qquad (2)$$

Em que, *p* identifica a probabilidade de ocorrência do evento e *x* o evento ocorrido.

Por se tratar de eventos com distribuição de Bernoulli, a soma dos sucessos ou fracassos terá distribuição Binomial de *n* parâmetros (o número de observações) e *p* (probabilidade da ocorrência de sucesso). A função de distribuição de probabilidade da Binomial é a seguinte:

$f(x;n,p) = \binom{n}{x} p^x(1-p)^{n-x}$, para x = 0, 1, 2, ..., n e $\binom{n}{x}$ uma combinação. (3)

As variáveis independentes devem ser categóricas ou numéricas e, mesmo se a variável de interesse não for binária, é possível torná-la, com o objetivo de aplicar a regressão logística. A distribuição informa a probabilidade de obtenção de uma das categorias da variável resposta e é preciso uma grande amostra, com mais de 30 casos por variável independente (CORRAR et al, 2011).

Um modelo de regressão linear indica como uma variável dependente (Y) se relaciona a uma ou mais variáveis independentes (X). Assim, a variável dependente pode assumir qualquer valor conforme altera-se a variável independente. Ou seja, quando se variam as independentes de menos infinito para mais infinito, a variável dependente também se altera de menos infinito até mais infinito. Há, no entanto, muitos casos em que a variável dependente (Y) é categórica e binária, ou seja, só pode assumir valores de duas categorias. Nestas situações a regressão logística mostra-se como um dos modelos mais adequados para se modelar o comportamento da variável dependente. No caso da variável dependente assumir dois estados (0 ou 1) e haver um conjunto *p* de variáveis independentes (X1, X2, ..., Xp), o modelo pode ser descrito como (MESQUITA, 2014):

$$P(Y = 1) = \frac{1}{1+e^{-g(x)}} \qquad (4)$$

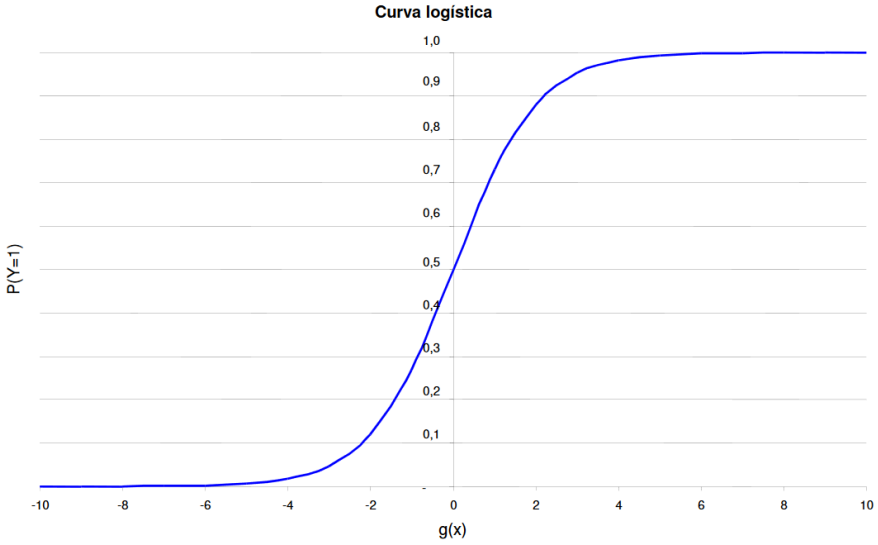
Em que,

$$g(x) = \beta_0 + \beta_1 \beta X_1 + \cdots + \beta_p \beta_p \qquad (5)$$

A partir do conjunto dados, estimam-se os coeficientes βi utilizados, com uso da máxima verossimilhança, buscando-se uma combinação que maximize a probabilidade da amostra ter sido observada (REGRESSÃO... 2020). Com dada combinação de coeficientes βi e variando os valores de X, nota-se que a curva da regressão logística possui comportamento probabilístico com forma de sigmoide, variando e menos infinito a mais infinito (HOSMER; LEMESHOW, 1989).

- Quando g(x) $\longrightarrow + \infty$, então P (Y = 1)$\longrightarrow$ 1
- Quando g(x) $\longrightarrow - \infty$, então P (Y = 1)$\longrightarrow$ 0

Figura 3. Curva da função logística



Fonte: Regressão... ([201-])

Da mesma forma que se pode estimar diretamente a probabilidade de ocorrência de um evento, pode-se estimar a probabilidade de não ocorrência pela diferença (MINUSSI et al., 2002):

$$P(Y = 0) = 1 - P(y = 1) \qquad (6)$$

Os coeficientes de um modelo de regressão logística são avaliados de maneira semelhante aos coeficientes dos modelos de regressão linear, mas com uma interpretação diferente.

Os coeficientes estimados na regressão logística apontam a variação da probabilidade de ocorrência de um evento, conforme se altera uma unidade na variável independente. Para coeficientes positivos, quanto mais alto seu valor, maior o poder de predição da variável independente sobre a probabilidade da ocorrência do evento estudado. Elevando-se a constante ao coeficiente da variável independente, têm-se o impacto que ele exerce na razão de chance (MESQUITA, 2014).

Já para o cálculo dos parâmetros β_i da regressão, deve-se utilizar o método da máxima verossimilhança, com um procedimento iterativo. Atribuem-se valores arbitrários aos coeficientes da regressão e assim obtém-se um modelo inicial para previsão dos valores observados. O passo seguinte é avaliar os erros da predição e mudar os coeficientes da regressão, buscando aumentar a probabilidade dos dados observados no novo modelo. Isso é repetido até que as diferenças entre o modelo mais recente e o modelo anterior sejam desprezíveis (MESQUITA, 2014).

2.3.1 Medidas de avaliação

Após estimativa dos coeficientes da regressão, o próximo aspecto a se observar antes da progressão na análise é a significância da variável, com testes de hipóteses para descobrir se a variável é, ou não, significativamente correlacionada com o retorno do modelo. Algumas das principais medidas de avaliação são:

a) Teste da razão de verossimilhança

O *likelihood value* possibilita testar a significância do coeficiente de uma variável do modelo pela comparação dos valores observados da variável dependente com os valores

previstos por cada um dos dois modelos: um com a variável presente e o outro sem tal variável. Para se comparar os valores preditos e observados, usa-se a função de verossimilhança (NELDER; WEDDERBURN, 1972):

$$D = -2ln\frac{verossimilhança\ do\ modelo}{verossimilhança\ do\ modelo\ saturado} \qquad (7)$$

Apelidada de *deviance* (desvio), ela avalia o valor ajustado do modelo, com a mesma função que a soma de quadrados residuais em outros modelos não logísticos. A *deviance* é sempre positiva e, quanto menor seu valor, melhor torna-se o ajuste do modelo (MESQUITA, 2014). Para a estimativa da significância de certa variável independente, comparam-se o valor de *D* com e sem a presença da variável independente na equação. Espera-se, pela inclusão da variável independente no modelo, a seguinte alteração no valor de *D*:

$$G = D(modelo\ sem\ variável) - D(modelo\ com\ variável) \qquad (8)$$

Pode-se expressar a estatística G por:

$$G = -2ln\frac{verossimilhança\ sem\ variável}{verossimilhança\ com\ variável} \qquad (9)$$

Com a hipótese de que ao menos um β é igual a zero, a estatística *G* terá distribuição assintótica qui-quadrado, com grau de liberdade igual à diferença da quantidade de parâmetros dos modelos em comparação. O valor *G* é comparado com o qui-quadrado no nível de significância pré-determinado, descobrindo-se se é possível remover do modelo as variáveis em estudo (MESQUITA, 2014).

b) Teste de Wald

O teste de Wald baseia-se na distribuição assintótica de β e é uma generalização do teste t de Student (TESTES... 2020). Ele avalia o modelo por inteiro e seu objetivo é determinar o grau de significância de cada coeficiente, incluindo a constante. O intuito é rejeitar a hipótese nula, o que indica que ao menos um dos coeficientes possui impacto no

valor da variável independente. Sua distribuição, como de outros testes do modelo de regressão logística, segue o Qui-quadrado (REGRESSÃO... 2020). As hipóteses do teste Wald são:

$$\begin{cases} H_0: \beta_i = 0, & \forall\, i \in \{1, \dots, n\} \\ H_1: ao\ menos\ um\ \beta \neq 0 \end{cases} \qquad (10)$$

Porém, há casos em que o teste de Wald não rejeita a hipótese nula, quando isso deveria ocorrer (JENNINGS, 1986). Assim, deve-se utilizar forma alternativas para a avaliação do modelo, como o teste anterior de verossimilhança, e os a seguir apresentados. O teste de Wald também é falho quando os coeficientes são muito grandes.

Os coeficientes (β) são divididos por seus respectivos erros padrão (SE).

$$Wald = \frac{\beta_j}{SE_{\beta_j}} \qquad (11)$$

c) Pseudo R^2

Nos modelos de regressão logística, as medidas de qualidade do ajuste são funções dos resíduos definidos como a diferença entre o valor observado e o valor ajustado (MESQUITA, 2014).

Na regressão logística não há uma estatística equivalente ao R^2 usado nos modelos de regressão linear. Assim, a denominação de pseudo R^2 é utilizado pelo fato de se parecem com o R^2do modelo de regressão linear, por estarem em escala similar, indo de 0 a 1. E quanto mais próximo o valor estiver de 1, melhor o ajuste do modelo (ZANINI, 2007).

Entretanto, apesar da semelhança, o pseudo R^2 não pode ser interpretado como se interpretaria o R^2 da regressão linear. Adiciona-se o fato de que diferentes definições do pseudo R^2 levam a valores muito distintos. No contexto da regressão logística, são dois os principais pseudo R^2 utilizados: R^2 Cox & Snell e R^2 Nagelkerke.

d) Teste Hosmer & Lemeshow

O teste Hosmer e Lemeshow tem o intuito de avaliar a validade preditiva do modelo de Regressão Logística medindo seu grau de acurácia. Não utiliza a verossimilhança, usando a visão real da variável dependente (HOSMER; LEMESHOW, 1989).

Os criadores apontam esta estatística como correspondente ao teste qui-quadrado, consistindo em dividir a quantidade de observações em cerca de dez classes e logo após comparar as frequências previstas com as observadas. O intuito do teste é verificar se há diferenças significativas entre as classificações feitas pelo método e a realidade das observações. Antes de se calcular a estatística teste, é preciso estimar a probabilidade de sucesso para cada observação e, de forma crescente, ordenar as probabilidades previstas. A seguir, os dados devem ser agrupados de acordo com os decis de probabilidades previstas. Em cada decil, dividem-se os valores previstos e os observados para o sucesso e o fracasso. A um nível de significância pré-determinado, visa-se não rejeitar a hipótese de não existirem diferenças entre os valores previstos e observados (ZANINI, 2007).

O critério de avaliação difere-se do convencional, pois usualmente deseja-se rejeitar a hipótese nula. Neste teste, se existirem diferenças significativas entre as classificações previstas pelo modelo e as observadas, o modelo não representa a realidade adequadamente. Nesta situação, o modelo não é capaz de realizar estimativas e classificações com confiabilidade.

e) Teste MAPE

Por fim, um dos gráficos mais utilizados no desenvolvimento deste trabalho é o de dispersão, com o intuito de representar visualmente como as previsões se correlacionam com os atrasos ocorridos. Assim, um dos testes auxiliares ao modelo de regressão logística será o MAPE. Ele não valida o modelo em si, mas informa o quão acurado foi o resultado retornado frente às médias ocorridas. O *Mean Absolute Percentage Error* ou erro absoluto percentual médio (MAPE) é bastante utilizado por seu fácil entendimento. A métrica expressa a porcentagem média dos erros (em valor absoluto, apenas positivos) ocorridos na previsão da série temporal. A previsão é tão melhor quanto mais baixo for o valor da medida (MIRANDA, 2014). Sua equação é dada por:

$$MAPE = \frac{\sum_{i=1}^{n} \frac{|P_i - O_i|}{O_i}}{n} \quad (12)$$

3 LEVANTAMENTO DE DADOS

3.1 Identificação de variáveis

A identificação das variáveis mais relevantes foi baseada em alguns estudos de referência a este trabalho. Sternberg *et al.* (2016) avaliaram e quantificaram todos os atributos que podem levar a atrasos utilizando dados de voos brasileiros e se guiaram por seis questões relacionadas a causas, momentos, diferenças e relações entre aeroportos e companhias aéreas. Observou-se através do cruzamento de dados fornecidos pela ANAC e por centros meteorológicos que condições meteorológicas adversas, como neblina, trovoadas e chuva, são as principais causas para a ocorrência de atrasos, podendo aumentar em até 216%.

Entretanto, a dificuldade em se prever as condições meteorológicas a longo prazo fez com que essa variável não fosse utilizada no presente trabalho. Conforme o Centro de Previsão de Tempo e Estudos Climáticos (CPTEC), atualmente, as previsões são geradas para até 15 dias, com um acerto de 98% para as 48 h e chance de acerto de 70% para cinco dias (TEIXEIRA *et al.*, [2018]). Para obter uma ferramenta preditiva de atrasos que futuramente possa ser utilizada por seguradoras para determinação dos prêmios para cada viagem, é preciso cuidado com a adoção de variáveis pouco previsíveis a longo prazo, como as meteorológicas. Um modelo que dependa de fatores como chance de chuva, temperatura, pressão e velocidade de vento pode dificultar seu uso para a estimativa de atrasos de voos para datas muito distantes da presente.

Outro fator identificado por Sternberg *et al.* (2016) que influencia na probabilidade de ocorrência de atrasos são os aumentos de demanda nos meses de férias, como dezembro, janeiro, junho, julho e alguns dias da semana, como sexta-feira que aumentam as chances de novos atrasos para 30% e 13%, respectivamente. Como os voos programados são aproximadamente os mesmos de outros meses ou dias da semana, sugere-se que esses atrasos podem ocorrer devido aos procedimentos de embarque ou antes da decolagem do avião.

Além do aumento de demanda em períodos de férias e sextas-feiras, os aeroportos e as companhias aéreas brasileiras parecem sofrer uma reação em cadeia de atraso, onde um atraso anterior atrasa os voos subsequentes do sistema. Os voos domésticos geralmente seguem ciclos de um dia, na qual as operações começam no início da manhã e terminam a noite, desse modo, quanto mais tarde for o voo, maior a chance de ocorrer um atraso. A maioria dos aeroportos demonstraram seguir a tendência de reação em cadeia de atrasos, com um aumento dos atrasos no final da noite e à noite. Contudo, alguns aeroportos tiveram um comportamento diferente como é o caso de Manaus e Belém que apresentaram uma taxa de atraso mais altos no final da manhã e à tarde. Essas cidades estão situadas próxima à floresta amazônica, com condições climáticas específicas e longe dos centros de negócios (STERNBERG *et al.*, 2016).

A localização dos aeroportos demonstra, portanto, ser uma outra variável relevante no atraso dos voos. No estudo realizado por Sternberg *et al.* (2016), observou-se que maioria dos aeroportos brasileiros tendem a produzir novos atrasos quando já possuem altos níveis de atrasos antecedentes e se associados a alguns atributos meteorológicos ou temporais, as chances de novos atrasos podem aumentar. Os aeroportos localizados nas regiões sul ou sudeste do Brasil tendem a apresentar maiores atrasos por condições meteorológicas, especialmente os aeroportos de Santos Dumont (Rio de Janeiro) e Congonhas (São Paulo), como pode ser observado na tabela 1. Sternberg *et al.* (2016) constataram que a relação aeroportos-atrasos são influenciados principalmente por condições meteorológicas adversas (especialmente nas regiões Sul e Sudeste), seguidas pela propagação de atrasos antecedentes e aumento de demanda por períodos.

Tabela 1. Principais motivos de atrasos em alguns dos principais aeroportos brasileiros

Aeroporto	**Principal motivo de atraso**	**Aumento das chances de atraso (%)**
Santos Dumont (Rio de Janeiro)	Voo por instrumento (IFR)	177
Belo Horizonte	Voo por instrumento (IFR)	139

Curitiba	Voo por instrumento (IFR)	137
Porto Alegre	Voo por instrumento (IFR)	136
Guarulhos (São Paulo)	Propagação de atraso (noite)	122
Galeão (Rio de Janeiro)	Chuva leve	118
Belém	Predominantemente nublado	107
Brasília	Aumento de demanda (período de férias)	105
Congonhas (São Paulo)	Chuva leve	102
Recife	Propagação de atraso (noite)	87

Fonte : Adaptação de Sternberg *et al.* (2016)

Por fim, a escolha da companhia aérea também influencia na ocorrência de atraso em voos. As principais companhias aéreas brasileiras apresentam diferenças importantes ao considerar atrasos nos voos. Dois deles parecem ser mais afetados por condições meteorológicas adversas e um por aumento de demandas (Tabela 2).

As performances das companhias aéreas diferem de aeroporto para aeroporto, a Gol é a única que tem mais chances de produzir novos atrasos na maioria dos aeroportos. No entanto, em alguns aeroportos, outras companhias aéreas apresentam piores desempenhos em termos de atrasos, como Azul em Guarulhos (São Paulo), TAM em Brasília ou Avianca em Manaus. Do ponto de vista aeroportuário, Guarulhos em São Paulo é um aeroporto não-pontual para todas as companhias aéreas brasileiras. Assim, provavelmente, a infraestrutura ou alguns dos serviços prestados pelo aeroporto podem estar afetando todos os voos. Por outro lado, Goiânia Vitória, Florianópolis e Fortaleza apresentam performances pontuais para todas as companhias aéreas. Outra diferença observada foi no aeroporto de Campinas, na qual a Azul tende a ter pontualidade neste aeroporto enquanto os voos da Gol têm 15% mais chances de atrasar neste aeroporto.

Tabela 2. Principais motivos de atrasos das companhias aéreas brasileiras

Companhia Aérea	Principal motivo de atraso	Aumento das chances de atraso (%)
GOL	Neblina	148
TAM	Tempestades e chuva	67
Avianca	Aumento de demanda (período de férias)	22

Fonte: Adaptação de Sternberg *et al.* (2016)

Com o intuito de se utilizar dados nacionais, optou-se pelo uso das mesmas variáveis e fonte de dados de Sternberg *et al*. (2016) e a base VRA da ANAC (2016). A base é um conjunto de tabelas com voos de partida e/ou chegada em território nacional, com informações diversas, detalhadas no tópico seguinte, e à disposição do público. Atrasos de voos são monitorados pela ANAC e o público pode ter acesso por meio dos horários de partida e chegada previstos e reais. As variáveis relacionam-se com a origem e destino dos voos, companhia aérea, período do dia e mês, excluindo as condições meteorológicas pela difícil previsão a longo prazo.

3.2 Coleta de dados

A ANAC disponibiliza em seu portal de dados abertos diversos conjuntos de arquivos sobre as viagens que partem ou chegam de território brasileiro. Para maior facilidade de visualização pelo público, a ANAC criou uma ferramenta para visualizar dinamicamente estatísticas sobre voos utilizando-se filtros diversos, como data, natureza (voo doméstico ou internacional), nacionalidade da empresa aérea, empresa aérea e informações sobre origem e destino. Os resultados são exibidos em gráficos de barras e mostram valores absolutos, percentuais e variação em relação a período anterior.

Apesar de útil, a ferramenta não traz uma das informações básicas para o início deste trabalho, pois os filtros não contemplam atrasos ou cancelamentos. Além disso, os tipos de cruzamentos de dados e visualização dos resultados ficam restritos às opções fornecidas pelo

site, o que impediria análises básicas como sazonalidade semanal e mensal e visualização dos dados em histogramas. Foi necessário o uso dos dados brutos fornecidos no mesmo portal, organizados em arquivos de formato CSV separados por ano-mês (ANAC, 2016).

Foram coletados os históricos de voos entre janeiro de 2016 e dezembro de 2018, além dos glossários de códigos e identificadores para realização de "de/para" com os valores usados nos arquivos. Os dados de histórico de voos foram disponibilizados como uma matriz de doze colunas, em que cada linha representa um voo programado - ocorrido ou não, sendo que voos cancelados não trazem dados na partida e chegada reais. As informações das colunas são as seguintes:

Tabela 3. Variáveis originais na base VRA da ANAC

Código da variável	**Descrição da variável**
ICAO Empresa Aérea	Sigla da empresa aérea
Número Voo	Identificador do voo
Código Autorização (DI)	Classificação interna do voo
Código Tipo Linha	Classificação da rota
ICAO Aeródromo Origem	Sigla do aeroporto de origem
ICAO Aeródromo Destino	Sigla do aeroporto de destino
Partida Prevista	Data, hora e minuto
Partida Real	Data, hora e minuto
Chegada Prevista	Data, hora e minuto
Chegada Real	Data, hora e minuto
Situação Voo	Identifica se foi realizado ou cancelado
Código Justificativa	Justificativa para atraso ou cancelamento

Para consultas mais eficientes e aprofundamento de futuras análises, foram criadas variáveis auxiliares a partir das originais. Diferentes faixas de atraso de chegada dos voos também foram determinadas, associando uma variável binária a cada faixa. As faixas foram escolhidas a fim de permitir a aplicação do algoritmo de regressão logística binária e de forma a possibilitar a aplicação do trabalho em seguros contra atrasos:

Tabela 4. Variáveis adicionais criadas para trabalho

Código da variável	Descrição da variável
Rota	Código que concatena origem e destino (exemplo: "SBPS_SBSP")
Atraso partida	Diferença entre partida real e a prevista
Atraso chegada	Diferença entre chegada real e a prevista
Dia da semana	Código do dia da semana da partida prevista
Mês	Mês da partida prevista
Período da Partida	Período do dia em que estava prevista a partida (madrugada, manhã, tarde e noite)
Atraso 1 a 20	É 1, se 1 <= Atraso chegada <= 20; 0 caso contrário
Atraso 21 a 40	É 1, se 21 <= Atraso chegada <= 40; 0 caso contrário
Atraso 41 a 60	É 1, se 41 <= Atraso chegada <= 60; 0 caso contrário
Atraso acima de 60	É 1, se Atraso chegada >= 61; 0 caso contrário
Atraso acima de 20	É 1, se Atraso chegada >= 21; 0 caso contrário

O 'código de autorização' foi removido, pois apenas voos regulares (99% dos voos) estão em estudo. O 'Número do voo' também foi removido, por se tratar de uma variável de uso interno e sem potencial preditivo. As variáveis utilizadas em todo o desenvolvimento do trabalho foram as seguintes.

Tabela 5. Variáveis utilizadas no trabalho

Variável	Tipo
ICAO Empresa Aérea	texto
Código Tipo Linha	caractere
ICAO Aeródromo Origem	texto
ICAO Aeródromo Destino	texto
Rota	texto

Partida Prevista	data (dia, mês, ano, hora, minuto)
Partida Real	data (dia, mês, ano, hora, minuto)
Chegada Prevista	data (dia, mês, ano, hora, minuto)
Chegada Real	data (dia, mês, ano, hora, minuto)
Situação Voo	texto
Código Justificativa	texto
Atraso partida	inteiro; minutos
Atraso chegada	inteiro; minutos
Dia da semana	inteiro; 0 a 6
Mês	inteiro; 1 a 12
Período da Partida	texto
Atraso 1 a 20	binário
Atraso 21 a 40	binário
Atraso 41 a 60	binário
Atraso acima de 60	binário
Atraso acima de 20	binário

3.3 Tratamento dos dados

Após organizar todos os arquivos em pastas, foi realizada uma primeira verificação dos dados históricos por amostragem, em planilhas, a fim de ganhar familiaridade com as informações. Em seguida, foi preciso empilhar os arquivos de voos para aumento do período analisado. Devido ao volume total dos arquivos, com um conjunto que ultrapassava 4,5 milhões de linhas, optou-se pelo uso da ferramenta Pandas de análise e manipulação de dados, disponibilizada gratuitamente na internet e construída com base na linguagem de programação *Python* (PANDAS .... 2020).

Esse empilhamento dos dados em um único grande arquivo não foi completamente satisfatório, pois o acesso na forma de consultas exigia muita memória computacional e frequentemente pequenos erros no meio dos dados ou do código cancelavam os procedimentos inesperadamente. A solução encontrada foi pré processar os dados com o *Dask*, uma biblioteca *Python* de computação paralela utilizada em *Big Data*, e guardá-los em

arquivos à parte para rápido acesso. Com isso, o acesso tornou-se mais eficiente e foi possível prosseguir com análises iniciais (ANACONDA, 2020).

Apesar da otimização obtida com uso das bibliotecas *Python*, a geração de gráficos e a construção de novas consultas exigia muito tempo de programação e edição diretamente no código. A fim de facilitar o trabalho a ser realizado e realizar a remoção de *outliers* e dados incoerentes, decidiu-se pela manipulação dos dados no PostgreSQL, um banco de dados relacional de objetos, que foi a forma final utilizada para manipulação dos dados (THE... 2020).

Assim como a maioria dos estudos citados no tópico de motivação deste trabalho, deu-se foco nos atrasos, retirando-se os cancelamentos das análises. O primeiro procedimento no banco de dados foi remover os voos cancelados, mantendo apenas a análise dos atrasos. Assim, os voos cancelados (8,92% do total) foram eliminados. Também, neste primeiro momento, foram detectados e removidos dados inconsistentes ou *outliers*, como:

- Todas as linhas com partida prevista ou chegada prevista nulos. Se não há previsão, não se podem calcular atrasos ou falar sobre cancelamentos.
- Uma linha em que a situação do voo era “Realizado”, mas a partida ou chegada reais era nulo.
- Voos com atrasos superiores a 24 horas (0,46% dos atrasados).
- Linhas em que o código tipo linha, empresa aérea, aeródromo de origem, aeródromo de destino ou situação do voo eram vazios.

As variáveis categóricas assumem valores em código, cujo significado pôde ser encontrado em Glossários fornecidos pela ANAC juntamente dos dados de voos. Abaixo e em anexo apresentam-se todas as categorias que as variáveis assumiram nos dados analisados.

Tabela 6.Situações de Voos disponíveis no conjunto de dados

Situação do Voo
Realizado
Cancelado

Tabela 7. Empresas Aéreas disponíveis no conjunto de dados

Código ICAO	Nome Empresa
AZU	Azul
FYW	Flyways Linhas
GLO	Gol
ONE	Avianca Brasil
PAM	Map Linhas Aereas
PTB	Passaredo
TAM	Tam
TTL	Total

Tabela 8. Períodos do dia de partida prevista

Períodos dia de partida	Horário
Madrugada	Hora partida < 6:00
Manhã	6:00 <= Hora partida < 12:00
Tarde	12:00 <= Hora partida < 18:00
Noite	18:00 <= Hora partida < 24:00

Tabela 9. Tipos de linha disponíveis no conjunto de dados (8)

Sigla Tipo Linha	Descrição Tipo Linha
C	Cargueiro
E	Especial
G	Cargueiro internacional
I	Internacional
L	Rede Postal

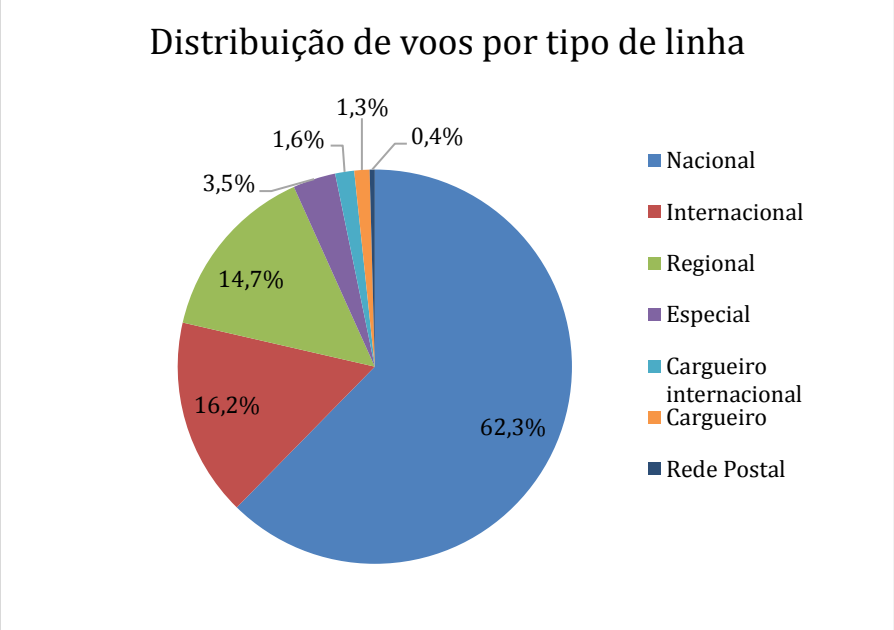
N	Nacional
R	Regional

Tabelas com as descrições das variáveis de Aeroporto, Meses, Dias da semana e Justificativas estão em anexo.

3.4 Análises iniciais

Após o tratamento inicial dos dados, prosseguiu-se com uma análise explanatória, para verificar como as variáveis escolhidas distribuem-se e se relacionam com outras. O tipo de linha é uma variável importante, pois apresentam concentração em poucas categorias:

Figura 4. Distribuição de voos por tipo de linha



Como esperado, a maioria dos voos são classificados como nacionais, por se tratarem de dados de uma base da ANAC. Voos de outros tipos além de nacional, regional e internacional fogem ao escopo do trabalho, por não terem usuários regulares potencialmente compradores de bilhetes de seguro (ANAC, 2016). O tipo internacional de voo não se enquadra na premissa do trabalho, de análise de voos exclusivamente nacionais, uma vez que partem ou chegam de território externos ao Brasil.

Voos nacionais e regionais foram explorados a seguir, para entendimento de seu perfil. Empresas com pouco voos, que juntas totalizam apenas 0,4%, foram agrupadas em outros.

Figura 5. Distribuição de voos nacionais por empresa aérea

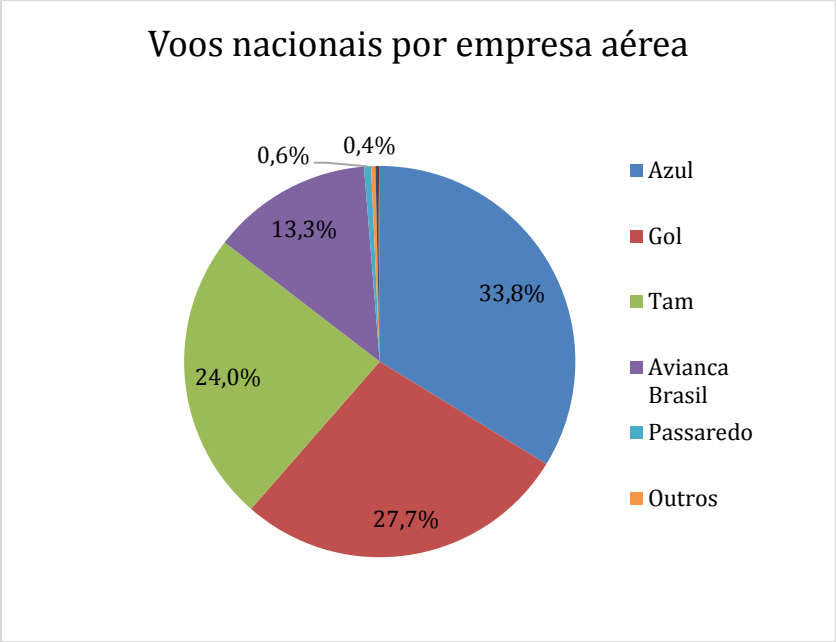
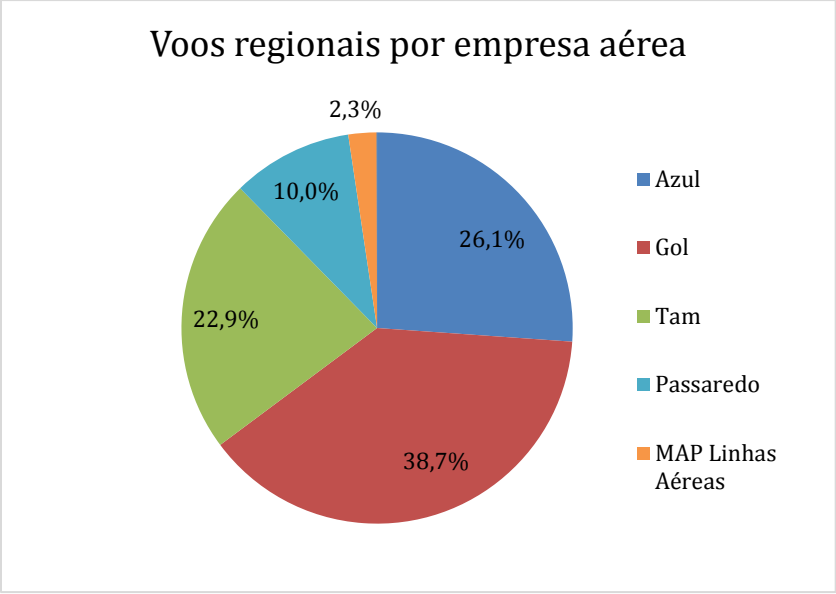


Figura 6. Distribuição de voos regionais por empresa aérea



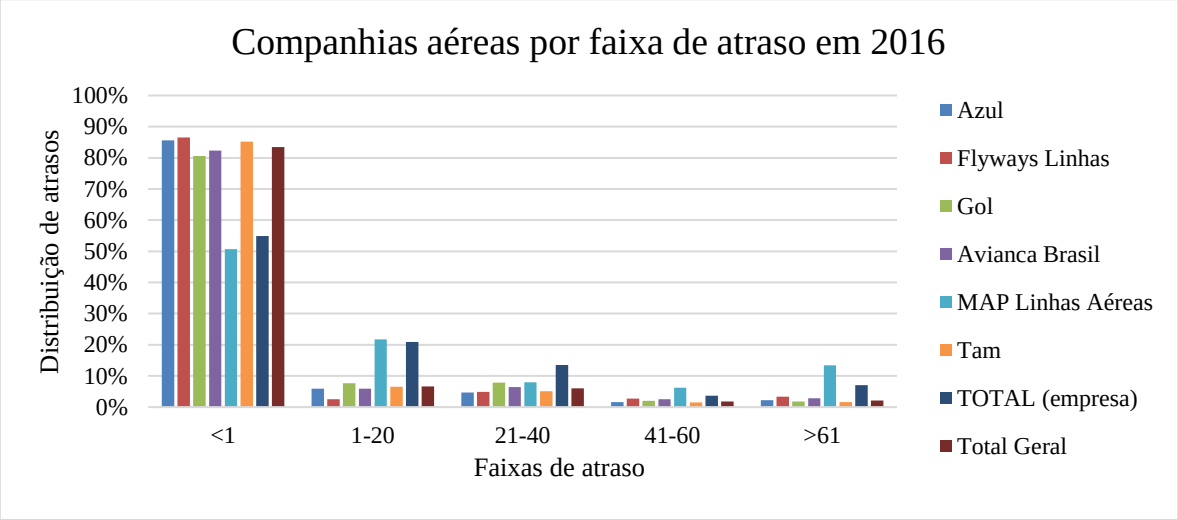
O perfil dos voos nacionais difere dos regionais. Voos nacionais possuem maior duração, possuem rotas mais longas e são realizados por mais empresas. Além disso, os voos regionais são pouco expressivos, representando apenas 14,7% do total. Assim, este trabalho envolverá apenas os voos nacionais para uma análise com dados mais homogêneos, sendo removidos os voos de outros tipos de linha.

Uma informação interessante é a grande quantidade de voos realizados por apenas três das empresas aéreas, que juntas totalizam 80% do total no período considerado. Uma exploração mais profunda foi realizada em amostra menor de voos, com dados de 2016 e 7 empresas presentes. Histogramas com diferentes faixas de atraso foram criados, para se observar como os atrasos se distribuem de maneira tão distinta variando-se as empresas aéreas. Ressalta-se que apenas as companhias Azul, Gol, TAM e Avianca possuem participação expressiva, mas optou-se por exibir todas, em ordem de participação percentual nos voos totais – da empresa com mais voos para a de menos voos.

Tabela 10. Distribuição das companhias aéreas por faixa de atraso em 2016

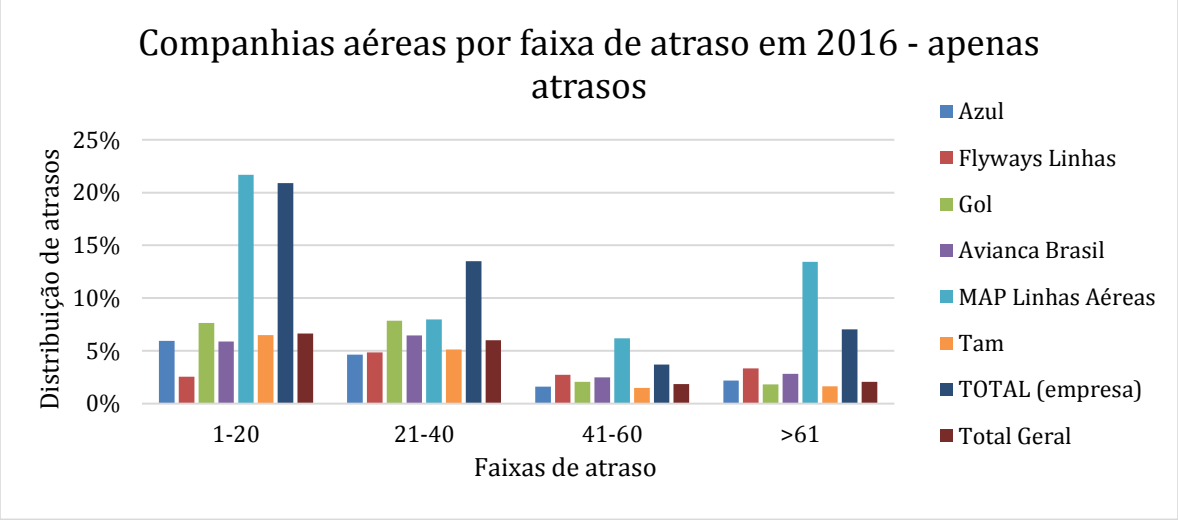
Faixas de atraso	**Azul**	**Gol**	**Tam**	**Avianca Brasil**	**MAP Linhas Aéreas**	**TOTAL (empresa)**	**Flyways Linhas**	**Total Geral**
<1	86%	81%	85%	82%	51%	55%	87%	83%
1-20	6%	8%	7%	6%	22%	21%	3%	7%
21-40	5%	8%	5%	6%	8%	13%	5%	6%
41-60	2%	2%	1%	2%	6%	4%	3%	2%
>61	2%	2%	2%	3%	13%	7%	3%	2%
Total Geral	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	100%

Figura 7. Distribuição das companhias aéreas por faixa de atraso em 2016



Como os voos atrasados são poucos em relação ao total, o gráfico seguinte omite a faixa de voos sem atraso (<1), facilitando visualização das diferenças entre empresas aéreas.

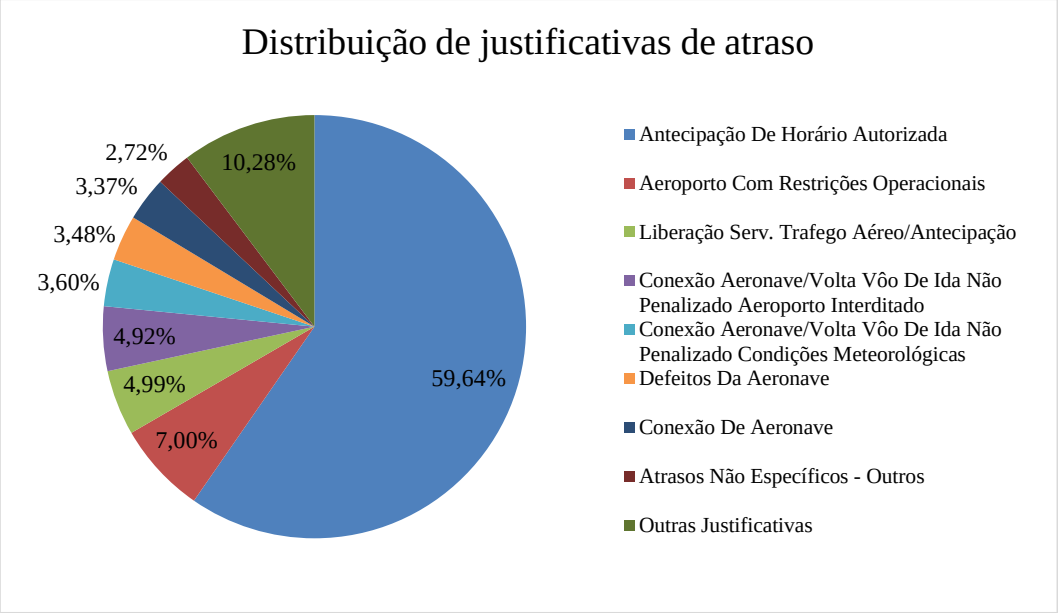
Figura 8. Distribuição das companhias aéreas por faixa de atraso em 2016 – apenas atrasados



Percebe-se grande variação entre faixas de atraso para diferentes empresas aéreas e aquelas com menos voos possuem maior proporção de atrasos. A variável mostra-se significante para a determinação da probabilidade de atrasos e foi melhor explorada nos tópicos seguintes.

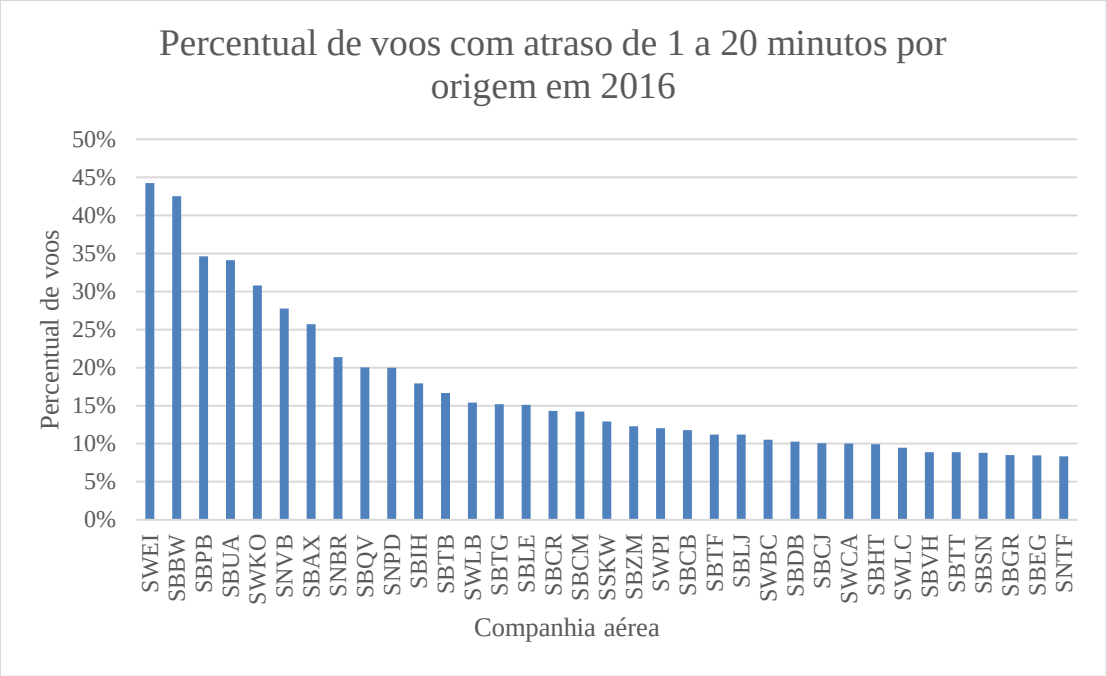
A variável ‘justificativa’ assume diversas categorias e inicialmente apresentava potencial para aplicação nos modelos, mas não foi possível encontrar descrições mais completas do que as registradas na tabela em anexo. O site da ANAC não disponibiliza informações complementares e o tratamento dos dados apontou muitas inconsistências, como justificativas de ‘Atraso não específico’ para voos que não atrasaram e ‘Antecipação de horário autorizada’ para voos cancelados. Assim, a opção de desconsiderá-la da modelagem foi tomada para preservação de resultados confiáveis.

Figura 9. Distribuição de justificativas de atrasos



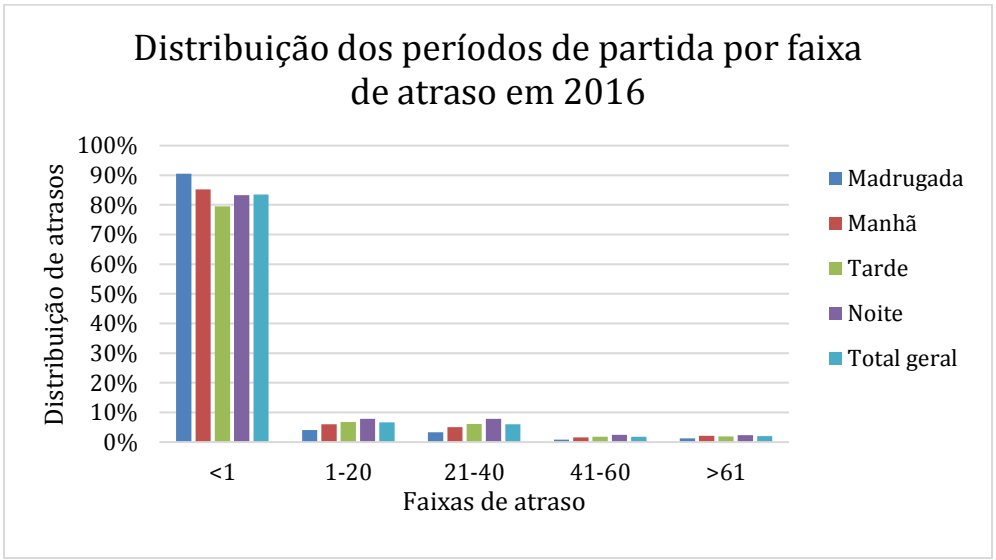
O atraso por aeroporto de origem se distribui em um gráfico de ‘cauda longa’, em que alguns aeroportos possuem grande frequência de atraso em determinada faixa analisada, enquanto a maioria se mantém próxima à média geral. No gráfico a seguir, exibem-se os 35 aeroportos com maior frequência de atraso na faixa de 1 a 20 minutos.

Figura 10. Distribuição de voos com atraso entre 1 e 20 minutos por aeroporto de origem em 2016



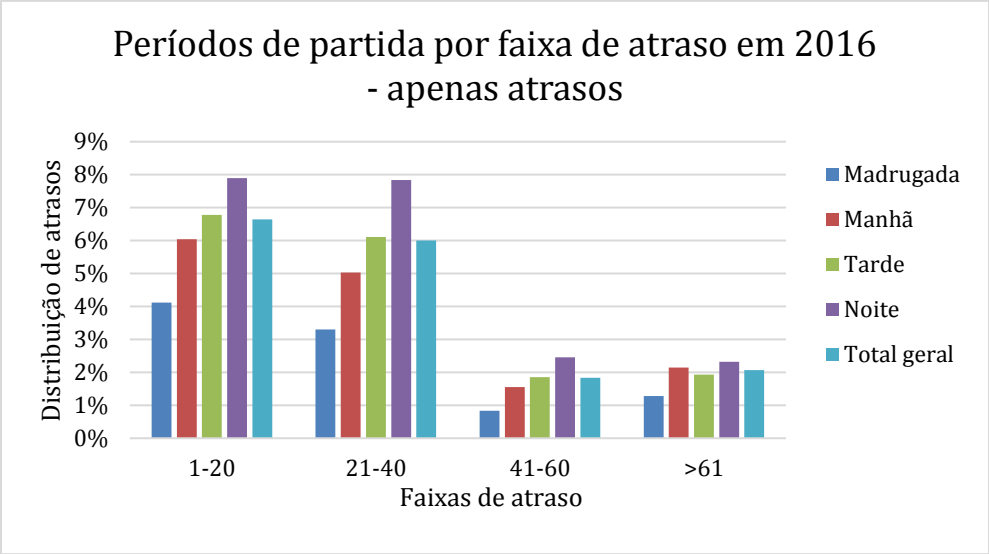
O período de partida também foi analisado com gráficos e, conforme relatado na identificação das variáveis, esperava-se variações nas faixas de atraso.

Figura 11. Distribuição dos períodos de partida por faixa de atraso em 2016



Como os voos atrasados são poucos em relação ao total, o gráfico seguinte omite a faixa de voos sem atraso (<1), facilitando visualização das diferenças entre os períodos de partida.

Figura 12. Distribuição dos períodos de partida por faixa de atraso em 2016 – apenas atrasados



É visível que os atrasos ocorrem com maior frequência conforme o passar do dia e reforça a relevância da variável para a modelagem.

A sazonalidade mensal foi analisada e esperava-se um aumento no número de voos próximo a julho e dezembro (Figura 13), épocas de férias, conforme melhor discutido no item ‘Identificação das variáveis’. Para o dia da semana, conforme também levantado no item ‘identificação de variáveis’, esperava-se maior frequência de atrasos nas sextas-feiras, dia em que há maior volume de voos.

Figura 13. Distribuição dos meses por faixa de atraso em 2016

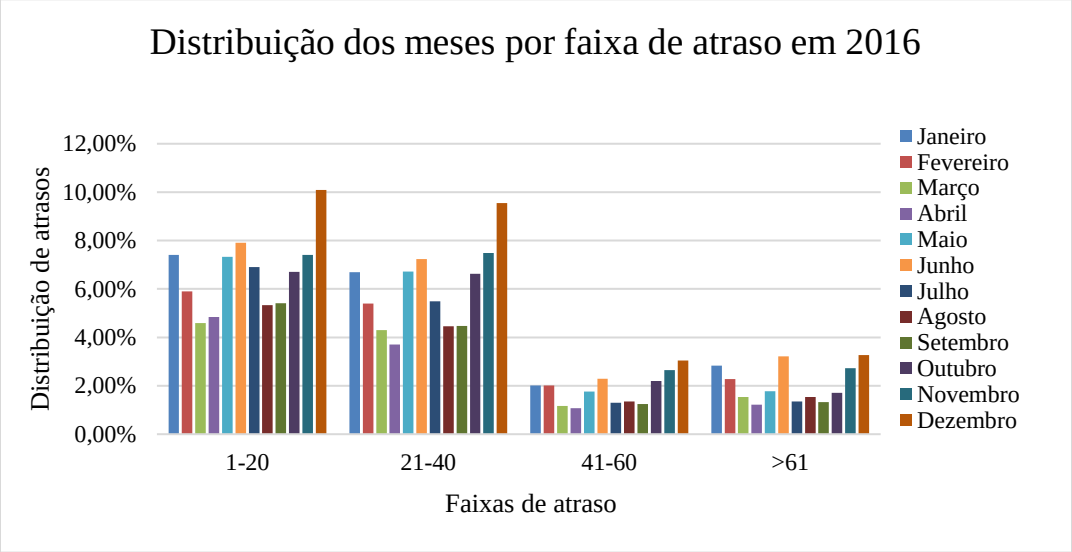
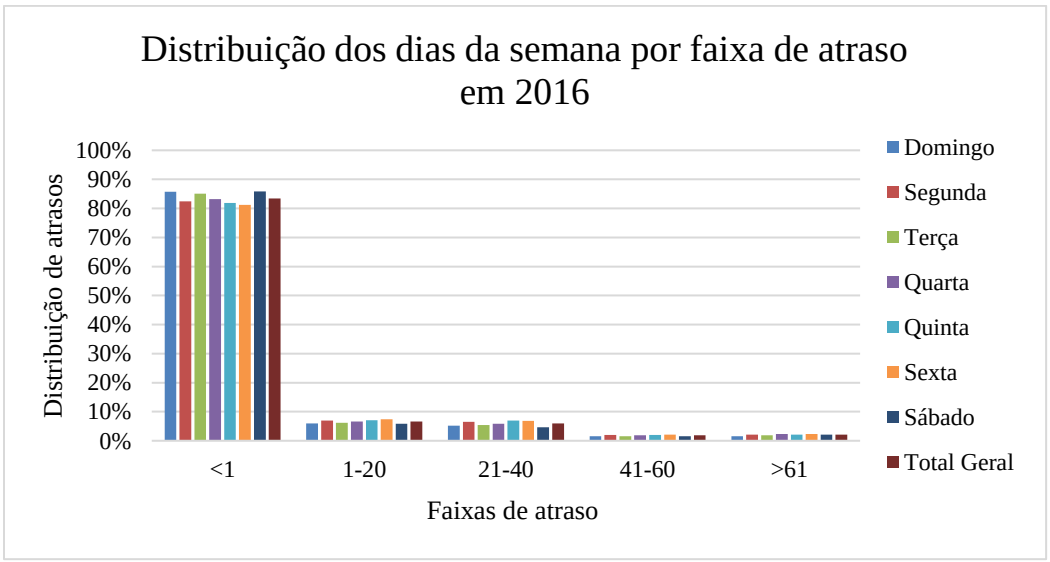
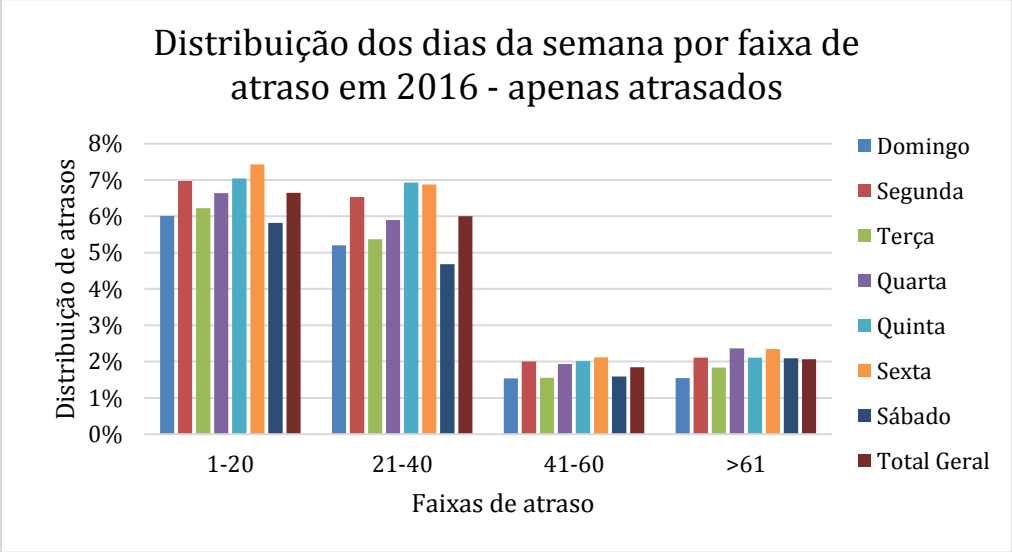


Figura 14. Distribuição dos dias da semana por faixa de atraso em 2016



Para melhor visualização dos voos atrasados, a faixa sem atrasos (<1) foi omitida no gráfico seguinte.

Figura 15. Distribuição dos dias da semana por faixa de atraso em 2016 – apenas atrasados



Há de fato um maior percentual de atrasos nos voos com partida na sexta-feira, mas os percentuais não diferem fortemente entre os dias, com variações menores que 2 pontos percentuais.

4 MODELOS ANALISADOS

4.1 Agrupamentos por *clusters*

Com o objetivo de conhecer quais os fatores determinantes nos atrasos, buscou-se encontrar similaridades entre os dados que permitissem a identificação de padrões úteis à análise dos dados e construção do modelo preditivo. Dessa forma, decidiu-se por usar um método de agrupamento por *clusters*, pois modelos lineares não seriam capazes de lidar com a natureza qualitativa das variáveis da base de dados, que em sua maior parte é categórica ou ordinal, ou seja, não numéricas.

Conforme revisão bibliográfica, o uso do método particional *K-means*, o mais comumente utilizado, não seria viável, devido à dificuldade na adaptação de variáveis categóricas como numéricas. Assim, prosseguiu-se ao uso de *K-medoids*, algoritmo que trabalha agrupando os dados a *medoids*, pontos como os centroides, mas que também fazem

parte do conjunto de dados e são os elementos mais próximos ao centro de gravidade do *cluster*.

A métrica de distância selecionada para agrupar os voos foi o número de variáveis em que duas linhas possuem informações diferentes, a *hamming distance*, explicada na revisão bibliográfica de agrupamentos por *clusters*. Ela permite que dois voos possam ser considerados tão mais diferentes quanto maior o número de variáveis com valores distintos entre os dois. Por exemplo, se entre dois voos a única diferença nas variáveis categóricas é o aeroporto de origem, sua distância é de 1 sobre o número de variáveis categóricas totais.

Foram realizados diversos agrupamentos, alterando-se o número de *clusters* (selecionados *a priori*) e variáveis utilizados, com o objetivo de se obter algum resultado que agrupasse os voos de maneira que os *clusters* possuíssem percentuais de atrasos por faixa distintos – assim, um *cluster* com alto (ou baixo) percentual de voos atrasados poderia ter sua composição explorada, identificando se, e quais, empresas aéreas, rotas e variáveis temporais concentram-se mais no agrupamento do que nos outros.

Os agrupamentos foram feitos com volume de dados reduzido, de um ano (cerca de 600 mil voos), repetindo-se os experimentos para cada ano da base. Foram utilizados 10, 5, 3 e 2 *clusters*. Para cada número de *clusters*, foram usadas diferentes combinações de variáveis, como: apenas variáveis temporais, apenas não temporais, outras combinações diversas e, por fim, o uso de todas em conjunto. As variáveis de faixa de atraso foram incluídas em parte dos experimentos, mas inicialmente esperava-se que não fossem necessárias, pois o intuito do agrupamento era que as similaridades internas dos dados fossem suficientes para agrupar *clusters* com maior ou menor percentual de voos atrasados.

No experimento exibido abaixo foram utilizados 10 *clusters*, distância de *hamming*, dados de 2016 e 7 variáveis: empresa aérea, aeródromo de origem, aeródromo de destino, mês, dia da semana, dia do ano da partida e horário da partida prevista. As faixas de atraso de cada *cluster* são apresentadas na tabela e gráficos abaixo.

Figura 16. Distribuição dos 10 clusters por faixa de atraso

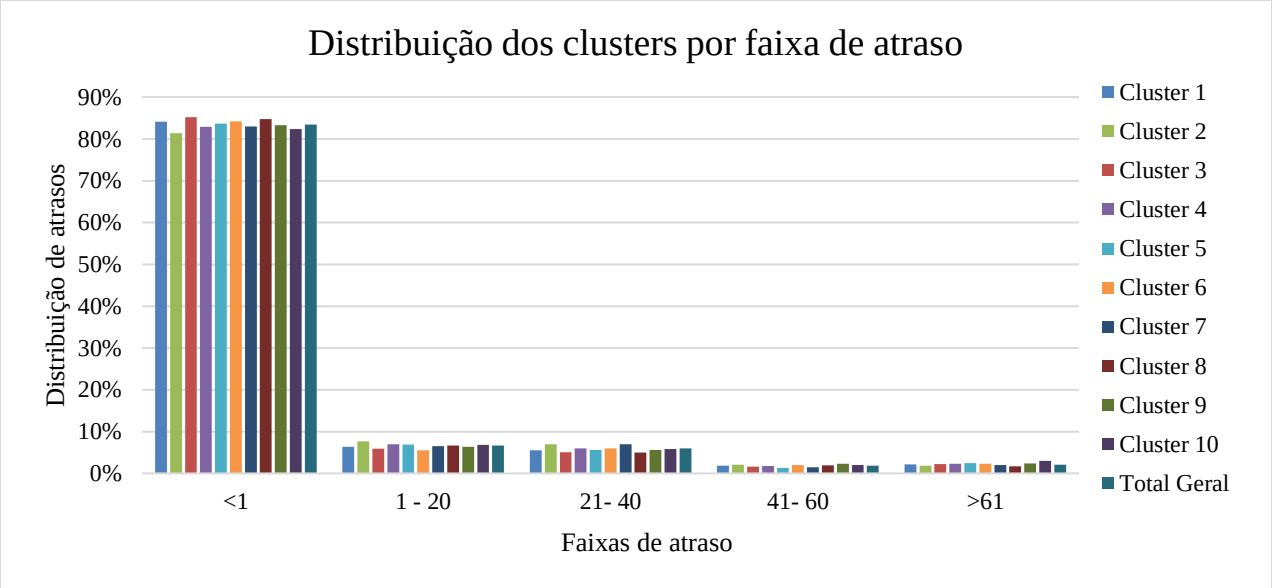
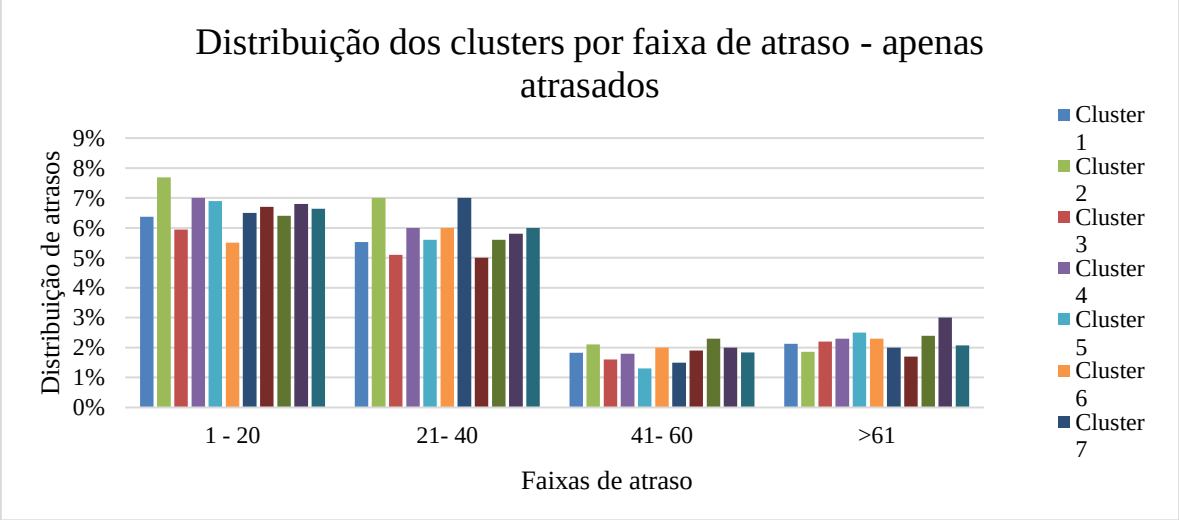


Figura 17. Distribuição dos 10 clusters por faixa de atraso – apenas atrasados



A maioria dos *clusters* varia pouco seus percentuais de atraso, não se distanciando muito da média global. Isso leva à conclusão de que os *clusters* não conseguiram agrupar os dados por atraso.

Outro experimento realizado, desta vez com o uso de quantidade reduzida de *clusters* - apenas 3 -, distância de *hamming*, dados de 2016 e 3 variáveis: rota, empresa aérea e atraso

na chegada. As distribuições dos *clusters* por faixas são apresentadas em tabela e gráfico a seguir.

Tabela 11. Atrasos por faixa por cluster

Faixas de atraso (minutos)	Cluster 1	Cluster 2	Cluster 3	Todos os dados
<1	84,1%	81,3%	85,2%	83,5%
1 - 20	6,4%	7,7%	5,9%	6,6%
21 - 40	5,5%	7,0%	5,1%	6,0%
41 - 60	1,8%	2,1%	1,6%	1,8%
>61	2,1%	1,9%	2,2%	2,1%
Total Geral	**100,0%**	**100,0%**	**100,0%**	**100,0%**

Figura 18. Distribuição dos clusters por faixa de atraso

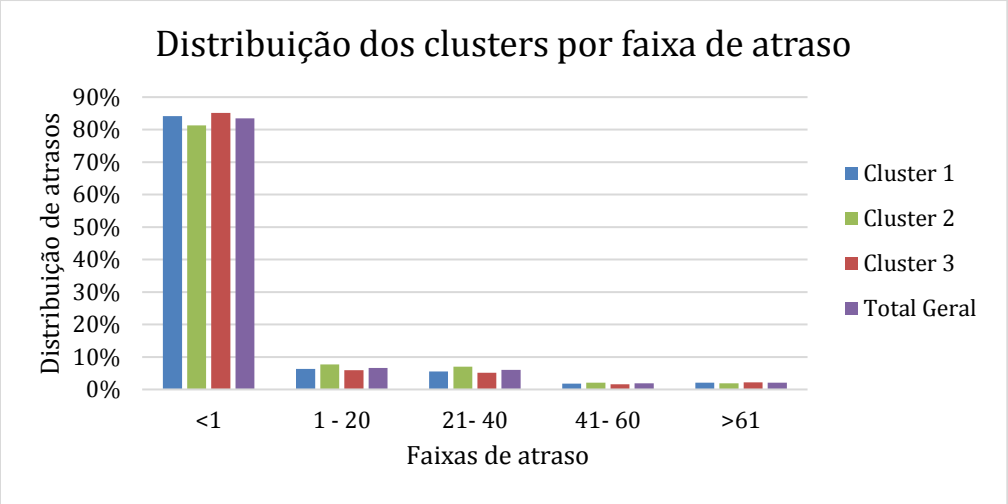
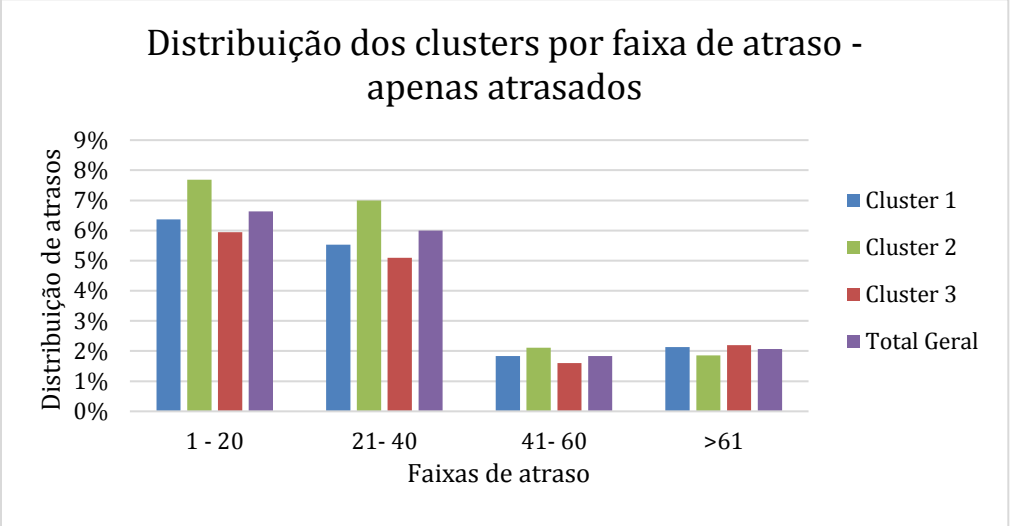


Figura 19. Distribuição dos clusters por faixa de atraso



Os valores não diferiram muito entre os clusters, tampouco da média dos dados. Para complementar a análise, observou-se a distribuição de algumas das variáveis pelos clusters, mas não se identificou nenhum agrupamento claro de empresas mais ou menos causadoras de atrasos, tampouco para a variável rota.

4.2 Discussão dos Agrupamentos por *Clusters*

Os resultados dos agrupamentos realizados não foram o esperado. Os *clusters* demonstraram não ter relação com os atrasos na chegada dos voos na análise por faixas de atraso. Com isso, não foi possível identificar conjuntos de variáveis mais relacionadas aos atrasos, tampouco categorias dentro de cada variável. Assim, não se aprofundou mais no uso de agrupamentos, seguindo diretamente para o uso de regressão logística binária para um modelo preditivo que independa de agrupamentos por *clusters*. O objetivo do uso de *clusters* seria permitir uma melhor seleção das variáveis a serem usadas na regressão, almejando a melhora dos resultados.

Em retrospecto, após a execução dos métodos e avanço no estudo, é possível que os *clusters* não tivessem relação com os atrasos porque a grande quantidade de combinações de categorias das variáveis (apenas de rotas há mais 700), exigissem uma quantidade ainda maior de dados para um bom resultado, talvez de dois ou mais anos. Também é possível que a presença das variáveis resposta de faixas de atraso fossem muito importantes e seu uso foi

prejudicado pelo fato de ser numérica e não categórica, como as outras utilizadas. Assim, o algoritmo *K-Medoids* não identifica, por exemplo, que um atraso de 40 minutos tem mais semelhança com um atraso de 39 minutos do que com outro de 3 minutos. Outras variáveis podem ter sido prejudicadas de maneira similar, como os dias da semana e dias do ano, que possuem natureza mais próxima da ordinal, uma vez que suas categorias se sucedem ao longo do tempo.

4.3 Regressão Logística

Para o entendimento de como cada variável contribui para a probabilidade de atraso e para atingir o objetivo de se desenvolver uma ferramenta preditiva, passou-se para o uso de regressão logística. Inicialmente, optou-se por usar os dados de apenas um ano, 2016.

Uma ferramenta preditiva em planilha foi construída com aplicação da fórmula de regressão logística e os resultados das regressões, sendo usada para previsão da probabilidade de atraso de um determinado voo, bastando a inserção dos valores das variáveis utilizadas na regressão. As previsões foram comparadas com a real média de atrasos ocorrida, utilizando-se gráficos de dispersão para verificação visual.

Para cada variável escolhida para participação no modelo final de regressão logística, houve o uso prévio de regressões para verificar sua capacidade preditiva individualmente. O objetivo foi identificar se todas possuem impacto significativo no percentual de atrasos. Uma variável que isoladamente é capaz de gerar boas previsões é bom indicativo de relevância para um modelo que combina mais variáveis. No entanto, fatores podem causar interferência uns aos outros. Assim, foram realizadas regressões logísticas binárias iniciais com atrasos maiores que 20 minutos como variável independente – por ser um conjunto de faixas com grande potencial de incômodo ao viajante e não ser preciso replicar a regressão para cada faixa individualmente. A variável dependente foi alterada uma a uma.

A primeira regressão utilizou a companhia aérea como variável dependente e a faixa de atrasos maiores que 20 minutos.

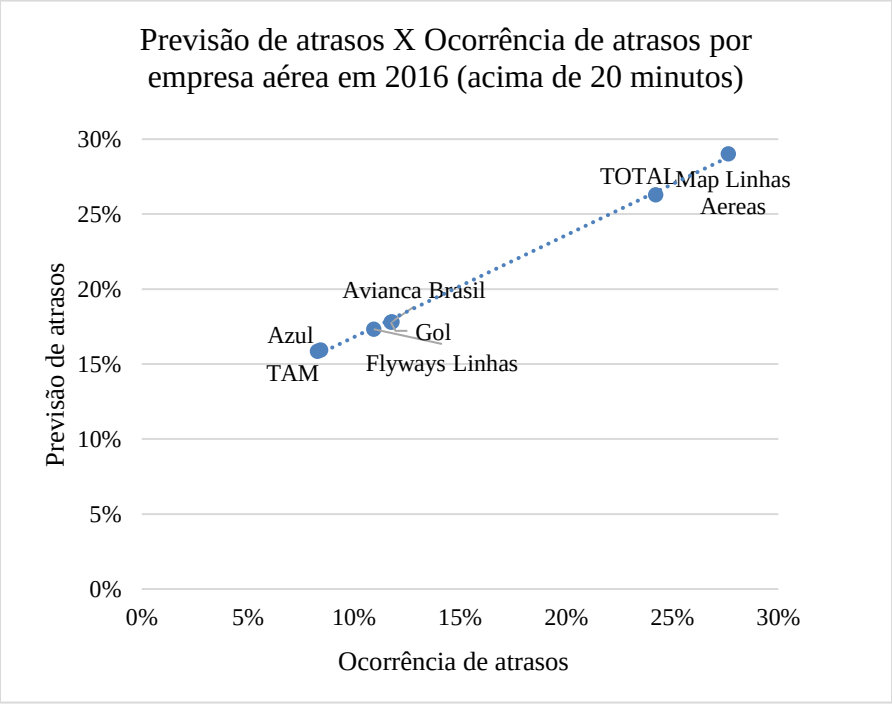
A partir dos coeficientes β, calculou-se para cada empresa aérea a probabilidade média de seus voos atrasarem mais de 20 minutos e comparou-se com a média de atrasos para cada companhia. Na tabela abaixo, a probabilidade de 8,3% para a TAM significa que

a média de todas as probabilidades de voos da TAM atrasarem mais de 20 minutos foi 8,3%. Já o percentual dos voos que realmente atrasarem nessa faixa foi de 15,9%:

Tabela 12. Comparação entre atrasos ocorridos e previstos por empresa aérea em 2016

Empresas	P(atraso>20min)	Atrasos reais
TAM	8,3%	15,9%
Azul	8,4%	15,9%
Flyways Linhas	10,9%	17,3%
Gol	11,7%	17,8%
Avianca Brasil	11,8%	17,8%
TOTAL	24,2%	26,3%
Map Linhas Aereas	27,6%	29,0%

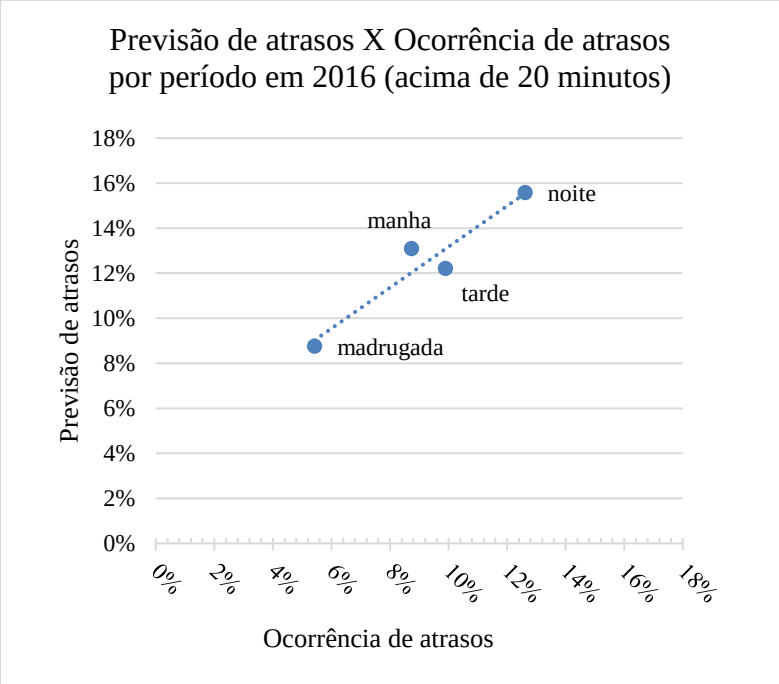
Figura 20. Comparação entre atrasos ocorridos e previstos por empresa aérea em 2016



Verifica-se ótima aderência entre os atrasos previstos e ocorridos por empresa aérea. Porém, as probabilidades de atraso aparecem deslocadas em relação à reta que cruza a origem, de tal forma que o erro para empresas com menores atrasos foi maior que o erro de empresas com maiores atrasos.

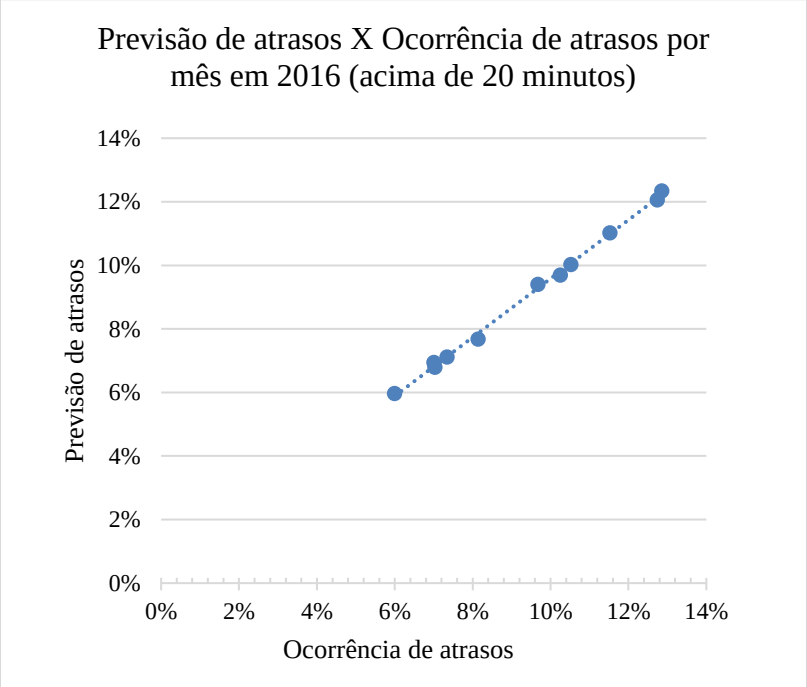
O período de partida foi analisado da mesma forma, com a mesma faixa de atraso superior a 20 minutos.

Figura 21. Comparação entre atrasos ocorridos e previstos por período do dia em 2016 (via regressão logística)



A sazonalidade mensal foi analisada e esperava-se um aumento no número de voos próximo a julho e dezembro, épocas de férias.

Figura 22. Comparação entre atrasos ocorridos e previstos por mês em 2016



O resultado foi coerente com o esperado: os meses com maiores atrasos são aqueles em torno de junho e dezembro. Há uma acentuação nas estações da primavera e verão, mas como os dados referem-se a um único ano, não é possível fazer afirmações antes da regressão final com dados de mais anos.

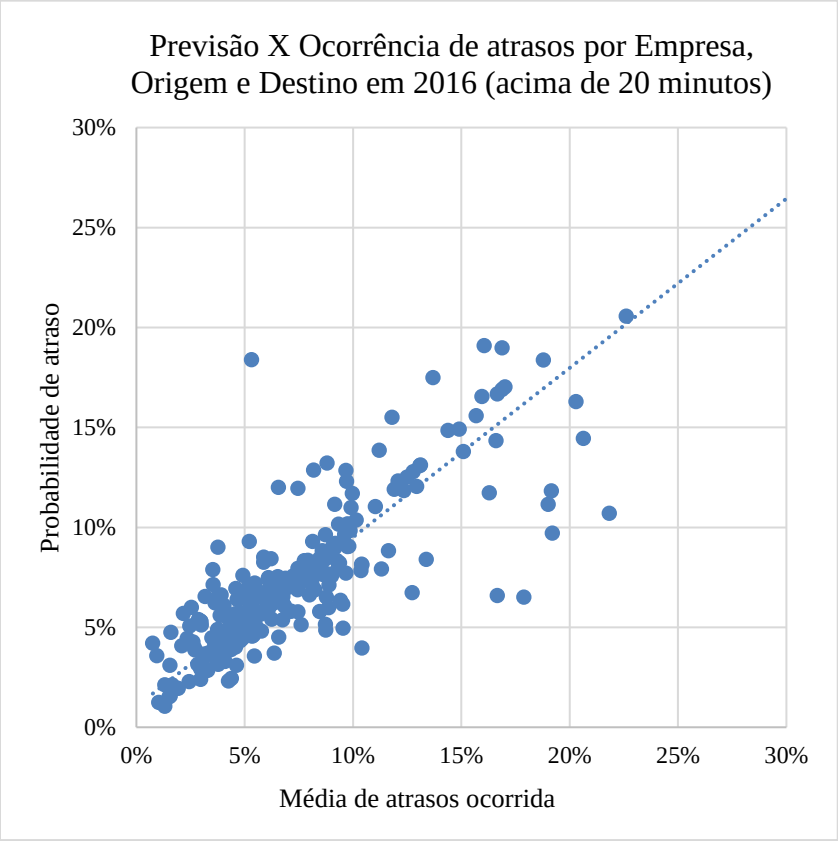
Análises similares foram realizadas para as variáveis Aeroporto de origem e Aeroporto de destino, individualmente. Também escolheu-se a faixa de atraso superior a 20 minutos, com resultados semelhantes .

Após as regressões com pares de variáveis, foi feita uma regressão logística apenas com as variáveis não temporais, seguido de regressão apenas com variáveis temporais, a fim de entender se os componentes de tempo levam a resultados distintos na determinação dos atrasos.

As variáveis não temporais utilizadas na regressão logística com dados de 2016 a 2018 foram ‘Empresa Aérea’, ‘Aeroporto de origem’ e ‘Aeroporto de destino’. A variável independente escolhida foi o atraso maior que 20 minutos. A tabela resultante com os coeficientes foi utilizada para previsão de atrasos. Para averiguação da qualidade da previsão, um gráfico de dispersão foi criado, em que cada ponto representa uma combinação única de

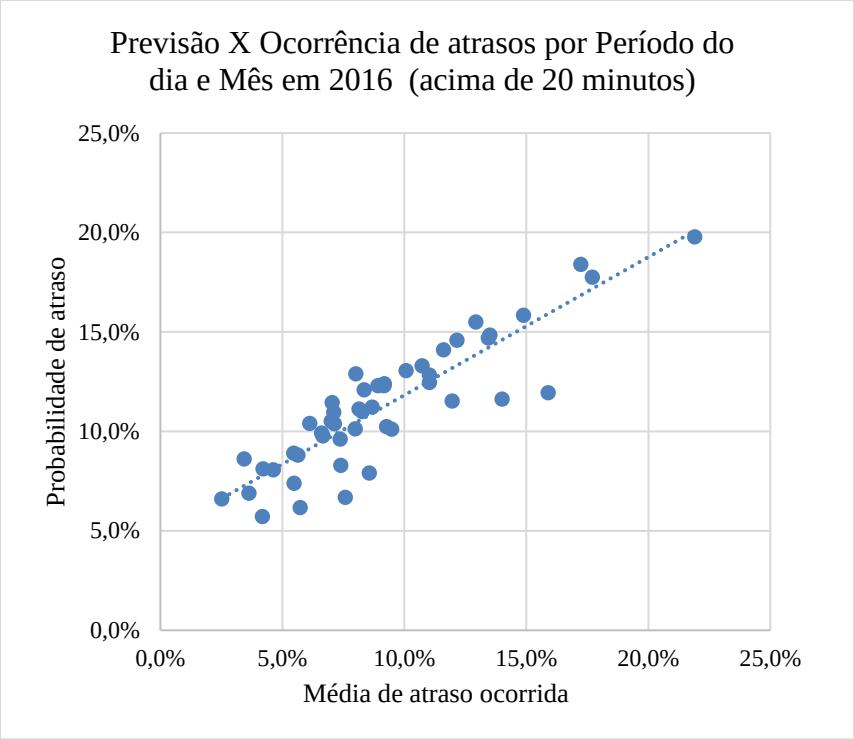
‘Aeroporto de origem’, ‘Aeroporto de destino’ e ‘Empresa Aérea’, com a previsão de atraso no eixo Y e a média de atrasos do grupo no eixo X.

Figura 23. Comparação entre atrasos acima de 20 minutos previstos e ocorridos por grupo de Empresa Aérea, Origem e Destino em 2016



Em seguida, uma regressão logística com dados dos mesmos anos foi realizada com as variáveis temporais ‘Período do dia’ e ‘Mês’. A variável dependente novamente foi a faixa de atraso superior a 20 minutos. A análise do resultado foi feita de maneira similar. Cada ponto do gráfico representa, no eixo X, a média dos atrasos para uma combinação de ‘Período do dia’ e ‘Mês’. No eixo Y tem-se a previsão de atraso do modelo.

Figura 24. Gráfico Regressão logística com Período do dia, Mês e atraso acima de 20 minutos

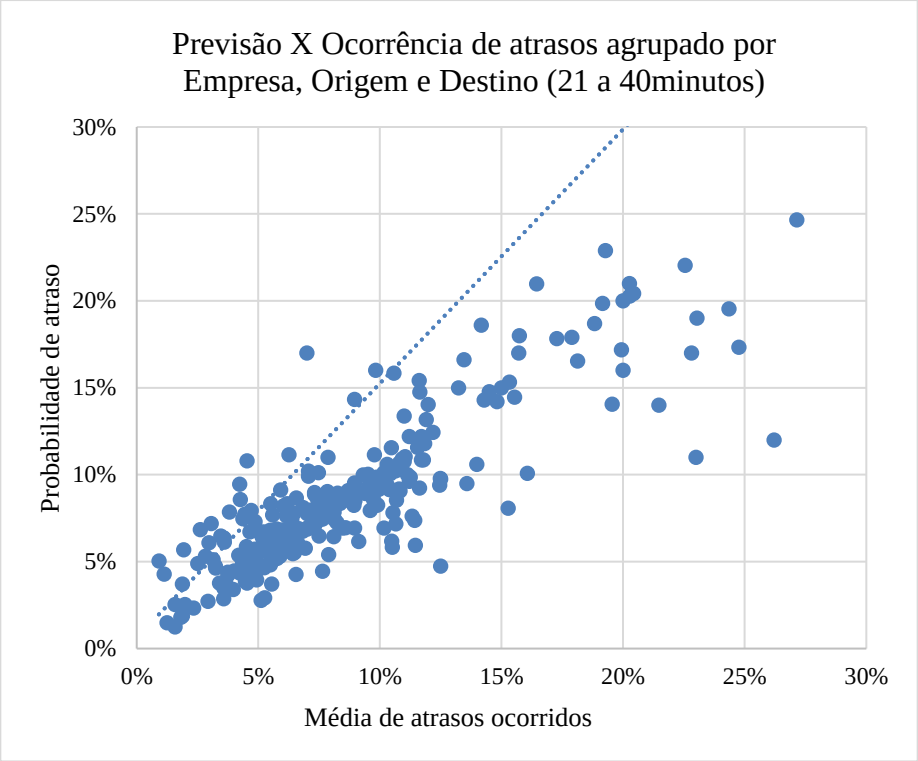


Com os bons resultados obtidos na comparação entre as probabilidades previstas e os atrasos de fato ocorridos, passou-se para a regressão com todas as cinco variáveis. Realizou-se a regressão para cada uma das quatro faixas de atraso. Os resultados novamente foram exibidos com gráficos de dispersão relacionando previsões e ocorrências e agrupando variáveis.

O resultado do modelo de regressão logística para a faixa de 21 a 40 minutos pode ser visto abaixo em um gráfico de dispersão, comparando-se as previsões com os atrasos ocorridos. O eixo X indica a média de atrasos ocorridos na faixa para um agrupamento distinto de empresa, aeroporto de origem e aeroporto de destino. Não foi feito agrupamento por todas as 5 variáveis usadas na construção do modelo, pois em muitos casos sua média de atrasos era de 0%, por não haver voos naquela combinação de empresa, aeroportos, período do dia e mês.

Além disso, mesmo agrupando por 3 variáveis, há grupos de média nula. Esses foram removidos para que fosse possível realizar o cálculo do erro MAPE.

Figura 25. Comparação entre atrasos (21 a 40 minutos) previstos e ocorridos por agrupamento de variáveis



O erro MAPE calculado foi de 18,3%. A princípio, este parece ser um bom valor e indicativo da qualidade da previsão, mas é preciso validar o modelo antes de realizar afirmações.

4.3.1 Análise do modelo de regressão logística

O software utilizado para obter os resultados foi o SPSS, que possui amplo ferramental estatístico. A saída para o método de regressão logística inclui três divisões: Análise descritiva, bloco 0 – inicial – e bloco 1 – final. O bloco 0 modela a regressão sem variáveis independentes, apenas com a constante construída (modelo nulo). Já o bloco 1 adiciona todas as variáveis. Cada uma das etapas gera tabelas com as informações analisadas. Entre outras coisas, analisou-se os dados por máxima verossimilhança, teste de Wald, Hosmer & Lemershow, pseudo R^2 e a significância dos coeficientes. O método de entrada foi a forçada (*enter*), com todas as variáveis inseridas no mesmo passo.

Algumas das tabelas geradas apresentam algumas das seguintes estatísticas:

Tabela 13. Descrição de estatísticas geradas pelo SPSS

Símbolo da estatística	**Descrição**
xi	Variáveis incluídas no modelo
i	coeficiente das variáveis. São utilizados para criar a equação de previsão
Const	Valor da constante
S.E.	Erro padrão dos coeficientes. Serve para testar se os parâmetros são significativamente diferentes de zero. Também formam o intervalo de confiança para o parâmetro
d.f.	Grau de liberdade da variável analisada
Wald e Sig.	Qui-quadrado de Wald testa a hipótese nula de que os coeficientes das variáveis são zero. A hipótese é rejeitada se o p-valor (na coluna 'Sig.') for menor que o valor crítico
Exp()	Usado para calcular os *odds ratio.* Quando seu valor é maior que 1, um aumento no valor da variável, faz com que a probabilidade de sucesso aumente. Quando menor, o oposto ocorre. Se 1, nada se altera
Intervalo de confiança	Intervalo de confiança de 95% de que o valor apontado esteja entre os limites superior e inferior

Fonte: adaptado de Mesquita (2014)

A primeira tabela gerada aponta que todas as linhas foram analisadas no procedimento.

Tabela 14.Casos ponderados na regressão logística com faixa de 21 a 40 minutos de atraso

Casos ponderados		**Número linhas**	**Porcentagem**
Casos selecionados	Incluído na análise	1594568	100,0
	Casos omissos	0	0,0
	Total	1594568	100,0
Casos não selecionados		0	0,0
Total		1594568	100,0

Uma das formas de se verificar a qualidade do resultado de uma regressão logística é estabelecer um valor de corte (0 a 1) a partir do qual a probabilidade de um evento ocorrer é interpretada como sucesso. Usualmente utiliza-se o valor 0,5. Assim, dados atribuídos a

probabilidades iguais ou superiores a 0,5 serão interpretados como sucesso, enquanto às outras será atribuído fracasso. Como a variável resposta binária está inclusa no modelo utilizado pela regressão, é possível comparar o número de casos apontados como sucesso ou fracasso pela previsão e sua real classificação. O percentual de acertos (previsões corretas sobre o total de casos) é indicativo da qualidade de previsão do modelo. Para este trabalho, classificar um voo como atrasado ou não atrasado é de menor relevância. O importante para uma seguradora é possuir uma estimativa do percentual de voos atrasados para certo agrupamento de características (por exemplo, saber que voos de uma empresa A, saindo de B para C, no período D e mês E atrasam em 8% dos casos). Usa-se o valor de corte de 0,08 nesta análise (valor pouco maior que a média de atraso da faixa de 21 a 40 minutos), mas deve-se ter em conta que a tabela de classificação não será a verificação mais importante.

A tabela a seguir indica que se todos os voos com predição de 8% de atraso fossem considerados atrasados, a taxa de acerto global seria de 94%. Esse resultado se deve ao fato de que a maioria dos voos inseridos não atrasaram e o valor de corte é baixo.

Tabela 15. Quadro de Classificação para bloco 0

Quadro de Classificação[a,b]					
Observado			Previsto		
			Atraso de 20 a 40		Porcentagem correta
			0	1	
Etapa 0	Atraso de 20 a 40	0	1498893	0	100,0
		1	95675	0	0,0
	Porcentagem global				94,0
a. A constante está incluída no modelo.					
b. O valor de recorte é ,080					

A tabela seguinte (16) compila informações relacionadas ao passo 0, sem a inclusão de variáveis independentes. Conclui-se que é preciso incluir as demais variáveis independentes para corretas previsões.

Tabela 16. Estatísticas na equação no bloco 0

Variáveis na equação							
			S.E.	Wald	df	Sig.	Exp()
Etapa 0	Constante	-2,751	0,105	686,44	1	0,000	0,064

Em seguida, passou-se para o bloco 1, realizando-se testes do modelo com um todo. Os métodos foram Step, Block e Model, os três com a finalidade de se rejeitar a hipótese nula de que cada coeficiente da equação iguala-se a zero. Segundo Mesquita (2014), o teste Step testa a contribuição individual de cada variável da etapa; Block testa a contribuição de todas que entraram no bloco; e Model testa a contribuição do modelo como um todo. Como a significância de cada teste é menor que p =0,05, pode-se concluir que há boa qualidade no modelo.

Tabela 17. Validade do modelo: teste de Omnibus do Modelo de Coeficientes

Testes de Omnibus do Modelo de Coeficientes				
		Qui-quadrado	df	Sig.
Etapa 1	*Step*	7088,706	231	0,000
	Block	7088,706	231	0,000
	Model	7088,706	231	0,000

Para se verificar a adesão dos valores previstos aos ocorridos, utilizam-se os pseudo R^2 de Cox & Snell e Nagelkerke. Já a verossimilhança é um indicador de quão bem o modelo prediz os resultados: quanto menor, melhor o modelo. O resultado abaixo é negativo nesse sentido, por ser elevado. Os valores dos pseudos R^2 obtidos, 0,22 e 0,52, não são muito altos, o que indica que a aderência das previsões pode não ter sido tão boa.

Tabela 18. Resumo do modelo – testes de verossimilhança e pseudos R^2

Resumo do modelo			
Etapa	Verossimilhança de log -2	R quadrado Cox & Snell	R quadrado Nagelkerke
1	26286,41	0,220	0,520

O teste de Hosmer e Lemeshow identifica se há diferenças significativas entre as classificações do modelo e a ocorrência real dos eventos. O intuito é a não rejeição da hipótese nula de que não há diferenças entre os valores preditos e observados.

A tabela abaixo mostra que o p-valor obtido pela distribuição qui-quadrado e identificado na coluna Sig. é menor que o nível de significância de 5%. Assim, rejeita-se a hipótese nula e o teste aponta que há diferenças entre os valores preditos e observados.

Tabela 19. Testes de Hosmer e Lemeshow

Teste de Hosmer e Lemeshow			
Etapa	Qui-quadrado	df	Sig.
1	94,322	8	0,000

A tabela de classificação do bloco 1 é colocada abaixo. Novamente, para o propósito de estimar probabilidade de atraso com o fim de precificar um seguro, não é esperado que o modelo se saia bem com uso de valores de corte.

Tabela 20. Tabela de classificação do bloco 1

Tabela de Classificação[a]					
Observado			Previsto		
			Atraso de 21 a 40		Porcentagem correta
			0	1	
Etapa 1	Atraso de 21 a 40	0	1206582	292277	80,5
		1	62976	32733	34,2
	Porcentagem global				77,7
a. O valor de recorte é ,080					

4.4 Discussão dos resultados

Os primeiros testes de variáveis com regressão logística, feitos isoladamente para as variáveis Aeroporto de origem, Aeroporto de destino, Empresa Aérea, Mês e Período do dia foram positivos, tendo boa relação entre valores previstos e dados reais de 2016. Um alerta quanto ao Mês foi levantado, pois apenas um ano foi analisado, limitando a qualidade do resultado, por haver apenas um grau de liberdade. A regressão logística foi feita para a faixa de atraso acima de 20 minutos.

Com o intuito de se realizar a última análise por partes, para se identificar se o fator temporal é mais ou menos relevante, a primeira tentativa feita foi de regressão com Aeroporto de origem, Aeroporto de destino e Empresa Aérea. Houve boa aderência dos dados reais do período de 2016 e o resultado previsto pela fórmula de regressão logística com os três anos, conforme verificado pelo gráfico de dispersão e tabela comparativa. A regressão seguinte com todos os dados foi realizada com as variáveis Período do dia, Mês e Atraso acima de 20

minutos. A aderência entre os valores previstos por agrupamento e as médias reais de atraso para voos de 2016 também foi boa, como se verificou no gráfico de dispersão.

Quanto ao modelo final com as variáveis empresa aérea, aeroporto de origem, aeroporto de destino, mês e período do dia, o resultado inicialmente mostrava-se positivo com o cálculo do MAPE, mas foi contrariado pelo encontrado na análise detalhada da saída do SPSS. A verossimilhança, o teste de Hosmer e Lemeshow e os pseudo R^2 apontam que o modelo não é plenamente confiável para previsões.

Uma das explicações para a diferença entre o apontado pelo MAPE e o gráfico de dispersão é ter sido utilizado um agrupamento de apenas 3 variáveis para visualização dos dados e cálculo do erro. O fato de haver muitos grupos com média de 0% para atrasos (removidos na montagem do gráfico) pode ser outro motivo para que o modelo falhe. Isso também explicaria por que as regressões anteriores, feitas com menos variáveis, pareciam caminhar para um bom resultado final.

Considerando que o método de agrupamento por *K-Medoids* também falhou, levanta-se a hipótese de que as variáveis não tenham sido adequadamente escolhidas ou que sejam insuficientes para a previsão. Como o resultado obtido por Sternberg *et al* (2016) de que as condições meteorológicas são as que mais geram impacto nos atrasos de voos brasileiros, é possível que essas sejam fatores essenciais na construção de um modelo preditivo válido para atrasos.

5. CONCLUSÕES

Os resultados de correlação entre o modelo de predição e os dados reais incialmente pareciam suficientes para que as seguradoras pudessem aplicar a ferramenta preditiva com quantidade reduzida de variáveis, não dependendo de fatores de difícil previsão e pouco controle, como os meteorológicos. No entanto, na última etapa, de análise com modelo completo e todas as cinco variáveis escolhidas, os resultados não foram o esperado. Os testes realizados pelo *software* não confirmaram a validade do modelo e algumas hipóteses para o ocorrido já foram levantadas. Assim, não houve motivos para se prosseguir e comparar os diferentes modelos utilizados – agrupamento por *clusters* e a regressão logística.

É possível que com uso de outras variáveis não exploradas a validade do modelo regressivo seja concluída e a qualidade das predições seja melhorada. Para as variáveis utilizadas, também é possível que o uso de redes neurais com treinamento com uma quantidade maior de voos possibilite resultados adequados - vale notar que o site da ANAC possui décadas de dados à disposição.

Além dos fatores meteorológicos, outro possível fator de atraso foi levantado ao final das análises e pode ser de utilidade para trabalhos futuros: a volumetria de voos por aeroporto. Os dados coletados na base VRA da ANAC permitem que se crie uma variável que contabilize o número de voos que partiram do mesmo aeroporto de origem nas últimas horas. Voos que partem de aeroportos com grande fluxo poderiam ser penalizados com maior probabilidade de atrasos. Essa variável, que funciona como um grande otimizador na precisão das previsões, é um valor de difícil previsão a longo prazo, uma vez que o número de voos programados para saída do aeroporto pode não estar finalizado, podendo haver alterações conforme se aproxime a data da viagem. Mesmo assim, seguradoras podem se beneficiar da inclusão da variável de volumetria para o cálculo de previsões com prazos menores e mais confiáveis.

Para a empresa em que se desenvolveu o trabalho, o levantamento bibliográfico e a exploração dos dados e criação da ferramenta preditiva já se demonstraram úteis, elevando as possibilidades de utilizar o trabalho como base para o projeto de precificação para seguradoras. Os próximos passos a serem tomados incluem a compreensão mais profunda dos motivos de falha na validade do modelo, levantamento de mais hipóteses para atrasos, avanço da ferramenta para que se torne um produto comercial, extensão do volume de voos

utilizados e estudo da precificação para desenvolvimento de uma ferramenta completa para venda a seguradoras.

REFERÊNCIAS BIBLIOGRÁFICAS

ANAC (Brasil). **Boletim de monitoramento do consumidor. gov. br**: **transporte aéreo.** 2018. Disponível em: <https://www.anac.gov.br/noticias/2018/anac-divulga-boletim-de-monitoramento-de-reclamacoes-de-consumidores-do-primeiro-trimestre/boletim-trimestral-de-monitoramento-do-consumidor-gov-br_1o-semestre-2018.pdf>. Acesso em: 01 jun. 2019.

ANAC (Brasil). **Boletim de monitoramento do consumidor. gov. br**: **transporte aéreo.** 2018. Disponível em: <https://www.anac.gov.br/noticias/2018/anac-divulga-boletim-de-monitoramento-de-reclamacoes-de-consumidores-do-primeiro-trimestre/boletim-trimestral-de-monitoramento-do-consumidor-gov-br_1o-semestre-2018.pdf>. Acesso em: 01 jun. 2019.

ANAC (Brasil). **Histórico de Voos**. 2016. Disponível em: <https://www.anac.gov.br/assuntos/dados-e-estatisticas/historico-de-voos>. Acesso em: 10 mar. 2019.

ANAC (Brasil). **Resolução nº 141, de 9 de março de 2010.** Dispõe sobre as Condições Gerais de Transporte aplicáveis aos atrasos e cancelamentos de voos e às hipóteses de preterição de passageiros e dá outras providências. 2010. Disponível em: <http://www.anac.gov.br/assuntos/legislacao/legislacao-1/resolucoes/resolucoes-2010/resolucao-no-141-de-09-03-2010/@@display-file/arquivo_norma/A2010-0141.pdf>. Acesso em: 14 maio 2019.

ANAC (Brasil). **Resolução nº 400, de 13 de dezembro de 2016.** Dispõe sobre as Condições Gerais de Transporte Aéreo. 2016. Disponível em: <https://www.anac.gov.br/assuntos/legislacao/legislacao-1/resolucoes/resolucoes-2016/resolucao-no-400-13-12-2016/@@display-file/arquivo_norma/RA2016-0400%20-%20Retificada.pdf>. Acesso em: 14 maio 2019.

ANAC, 2015. Agância Nacional de Aviação Civil. Technical Report. <http://www.anac.gov.br/>

ANACONDA. **Dask Documentation.** Disponível em: https://docs.dask.org/en/latest/. Acesso em: 11 jan. 2020.

ANÁLISE DE REGRESSÃO LOGÍSTICA. Disponível em: http://www.portalaction.com.br/analise-de-regressao/45-predicao. Acesso em: 10 abr. 2020.

APENAS 35% dos brasileiros contratam seguro viagem quando vão ao exterior. 2017. Disponível em: https://www.revistaapolice.com.br/2017/02/seguro-viagem-ao-exterior/. Acesso em: 14 mar. 2020

AVEN, Terje et al. Society For Risk Analysis. **SRA glossary**. 2015. Disponível em: htt0070s://www.sra.org/sites/default/files/pdf/SRA-glossary-approved22june2015-x.pdf. Acesso em: 15 mar. 2020.

AXA goes blockchain with fizzy. 2017. Disponível em: https://www.axa.com/en/magazine/axa-goes-blockchain-with-fizzy. Acesso em: 13 fev. 2020.

BRASIL. **Circular SUSEP n.º 535, de 28 de abril de 2016.** Estabelece a codificação dos ramos de seguro e dispõe sobre a classificação das coberturas contidas em planos de seguro, para fins de contabilização. Disponível em : <http://www2.susep.gov.br/bibliotecaweb/docOriginal.aspx?tipo=1&codigo=37965>. Acesso em: 01 jun. 2019.

BRASIL. Lei nº 7565, de 19 de dezembro de 1986. : Subchefia para Assuntos Jurídicos. Brasil, Disponível em: http://www.planalto.gov.br/ccivil_03/leis/L7565.htm. Acesso em: 08 abr. 2020.

CORRAR, LUIZ J. Análise Multivariada: para os cursos de administração, ciências contábeis e economia/ **FIPECAPI –Fundação Instituto de pesquisas Contábeis, Atuariais e Financeiras**; Luiz J. Corrar, Edílson Paulo, José Maria Dias Filho (coordenadores).1 ed. –São Paulo: Atlas, 2011.

COX, D.R.; SNELL, E.J. **Analysis of Binary Data**. London: Chapman & Hall, 2ª Edição, 1989.

CRAMER, J.S. **Logit models from economics and other fields**. Cambridge: Cambridge University, 2003.

DOI, Anderson. **Gerenciamento de riscos corporativos em pequenas e médias empresas**: análise de uma empresa nacional do setor de TI. 2017. Dissertação (Mestrado em Empreendedorismo) - Faculdade de Economia, Administração e Contabilidade, Universidade de São Paulo, São Paulo, 2017. doi:10.11606/D.12.2017.tde-07122017-113323. Acesso em: 15 mar. 2020.

DU, K.-l.. Clustering: a neural network approach. **Neural Networks**, [s.l.], v. 23, n. 1, p. 89-107, jan. 2010. Elsevier BV. http://dx.doi.org/10.1016/j.neunet.2009.08.007.

DUDA, R. O.; HART, P. E. **Pattern Classification and Scene Analysis**. New York: JohnWiley & Sons Inc, 1973.

EVERITT, B. S. **Cluster Analysis**. New York: John Wiley & Sons, Inc., 1993.

GRACE, Martin F. *et al*. The Value of Investing in Enterprise Risk Management. **Journal Of Risk And Insurance**, [s.l.], v. 82, n. 2, p. 289-316, 16 jan. 2014. Wiley. http://dx.doi.org/10.1111/jori.12022.

GUPTA, S. K. *et al*. K-means Clustering Algorithm for Categorical Attributes. **Data Warehousing and Knowledge Discovery**. Alemanha: Springer, 1999. p. 203-208. Disponível em: https://link.springer.com/content/pdf/10.1007%2F3-540-48298-9.pdf. Acesso em: 1 fev. 2020.

GUZHVA, Vitaly S. *et al*. Aviation Insurance and the Impact on Risk Management. **Aircraft Leasing And Financing**, [s.l.], p. 141-168, 2019. Elsevier. http://dx.doi.org/10.1016/b978-0-12-815285-0.00005-5.

HAIR, J.R. *et al*. (1998). **Multivariate analyses data**. New Jersey: Princeton University Press, 1998.

HARIKUMAR ITT, B. S. **Cluster Analysis**. New York: John Wiley & Sons, Inc., 1993.

HOSMER, D.W.; LEMESHOW, S. **Applied Logistic Regression**. New York: John Wiley, 2ª Edição, 2000.

INSTITUTE FOR DIGITAL RESEARCH & EDUCATION STATISTICAL CONSULTING. **What is the difference between categorical, ordinal and numerical variables?** Disponível em: https://stats.idre.ucla.edu/other/mult-pkg/whatstat/what-is-the-difference-between-categorical-ordinal-and-numerical-variables/. Acesso em: 04 abr. 2020.

ISO - INTERNATIONAL ORGANIZATION OF STANDARDISATION. **Gestão de riscos:** princípios e diretrizes, 2009. Disponível em: <http://www.iso.org>. Acesso em: 15 mar. 2020.

JAPIASSU, H.; MARCONDES, D. (1989). **Pequeno dicionário de filosofia**. São Paulo: Jorge Zahar Ed., 1989.

JENNINGS, D.E. **Judging Inference Adequacy in Logistic Regression**. Journal of the American Statistical Association, Vol. 81, 471-476, 1986.

KHANMOHAMMADI, Sina *et al*. A New Multilevel Input Layer Artificial Neural Network for Predicting Flight Delays at JFK Airport. **Procedia Computer Science**, [s.l.], v. 95, p. 237-244, 2016. Elsevier BV. http://dx.doi.org/10.1016/j.procs.2016.09.321.

LIKAS, Aristidis *et al*. The global k-means clustering algorithm. **Pattern Recognition**, [s.l.], v. 36, n. 2, p. 451-461, fev. 2003. Elsevier BV. http://dx.doi.org/10.1016/s0031-3203(02)00060-2.

MATOS, Renata Assis de. **COMPARAÇÃO DE METODOLOGIAS DE ANÁLISE DE AGRUPAMENTOS NA PRESENÇA DE VARIÁVEIS CATEGÓRICAS E CONTÍNUAS**. 2007. 156 f. Dissertação (Mestrado) - Curso de Estatística, Universidade Federal de Minas Gerais, Belo Horizonte, 2007

MESQUITA, Paulo Sérgio Belchior. **UM MODELO DE REGRESSÃO LOGÍSTICA PARA AVALIAÇÃO DOS PROGRAMAS DE PÓS-GRADUAÇÃO NO BRASIL**. 2014. 107 f. Dissertação (Mestrado) - Curso de Engenharia de Produção, Centro de Ciências e Tecnologia, Universidade Estadual do Norte Fluminense, Campos dos Goytacazes, 2014.

Ministério da Infraestrutura. **Novo Guia do Passageiro**: direitos do passageiro. Direitos do Passageiro. Disponível em: https://www.infraestrutura.gov.br/novoguiadopassageiro/direitos-do-passageiro. Acesso em: 10 maio 2020.

MINUSSI, João Alberto et al. **Um Modelo de Previsão de Solvência Utilizando Regressão Logística**. RAC, v. 6, n. 3, p. 109-128, set. 2002. Trimestral. Disponível em: http://www.anpad.org.br/periodicos/arq_pdf/a_380.pdf. Acesso em: 1 jan. 2020.

MIRANDA, Isabela Pagani Heringer de. **Comparação de diferentes Métodos de Previsão em Séries Temporais com valores discrepantes**. 2014. 31 f. TCC (Graduação) - Curso de Estatística, Ufjf, Juiz de Fora, 2014.

MOREIRA, Dilmo Bantim. **Gestão de Riscos de Seguros de Pessoas**. Disponível em: http://www.anspnet.org.br/opiniao-academica/gestao-de-riscos-de-seguros-de-pessoas/. Acesso em: 15 mar. 2020.

NELDER, J.A.; WEDDERBURN, R.W.N. Generalized Linear Models. **Journal of the Royal Statistical Society**, A, V.135, p.370-384, 1972

OLIVA, Fábio Lotti *et al*. A maturity model for enterprise risk management. **International Journal Of Production Economics**, [s.l.], v. 173, p. 66-79, mar. 2016. Elsevier BV. http://dx.doi.org/10.1016/j.ijpe.2015.12.007

PANDAS DEVELOPMENT TEAM. **Pandas Documentation**. Disponível em: https://pandas.pydata.org/docs/. Acesso em: 08 jan. 2020.

PARK, Hae-sang; JUN, Chi-hyuck. A simple and fast algorithm for K-medoids clustering. **Expert Systems With Applications**, [s.l.], v. 36, n. 2, p. 3336-3341, mar. 2009. Elsevier BV. http://dx.doi.org/10.1016/j.eswa.2008.01.039.

REGRESSÃO LOGÍSTICA. [201-]. Disponível em: https://edisciplinas.usp.br/pluginfile.php/3769787/mod_resource/content/1/09_RegressaoLogistica.pdf. Acesso em: 11 abr. 2020.

REGRESSÃO LOGÍSTICA. Disponível em: https://edisciplinas.usp.br/pluginfile.php/3769787/mod_resource/content/1/09_RegressaoLogistica.pdf. Acesso em: 10 abr. 2020.

ROMANO, Rogério Tadeu. **O SINISTRO**: o artigo discute sobre o sinistro no contrato de seguro. O ARTIGO DISCUTE SOBRE O SINISTRO NO CONTRATO DE SEGURO. 2019. Disponível em: https://jus.com.br/artigos/75637/o-sinistro. Acesso em: 02 abr. 2020.

SAKHALKAR, Soumitra *et al*. Effective SAR image segmentation and sea-ice floe distribution analysis via kernel graph cuts based feature extraction and fusion. **Science And Technology Publications**, fev. 2015. 10.5220/0005408100280037.

SAXENA, Amit *et al*. A review of clustering techniques and developments. **Neurocomputing**, [s.l.], v. 267, p. 664-681, dez. 2017. Elsevier BV. http://dx.doi.org/10.1016/j.neucom.2017.06.053.

SILVA, Fabiana Lopes Da; CHAN, Betty L.; "Análise da Demanda e Sinistralidade do Seguro Prestamista", p. 233 -254. In: **Aportes ao Desenvolvimento da Economia Brasileira**. São Paulo: Blucher, 2015. ISBN: 978-85-8039-123-7, DOI 10.5151/9788580391237-12

SRA. **Specialty Groups**. [2019]. Disponível em: https://www.sra.org/specialty-group. Acesso em: 15 mar. 2020

STERNBERG, Alice et al. An analysis of Brazilian flight delays based on frequent patterns. **Transportation Research Part e**: Logistics and Transportation Review, [s.l.], v.95, p. 282-298, nov. 2016. Elsevier BV. http://dx.doi.org/10.1016/j.tre.2016.09.013.

(SUSEP), Superintendência de Seguros Privados **6º Relatório de Análise e Acompanhamento dos Mercados Supervisionados.** Janeiro de 2018. Disponível em: <http://www.susep.gov.br/menuestatistica/SES/6b0%20Relat_Acomp_Mercado_2018.pdf> . Acesso em 01 jun. 2019.

(SUSEP), Superintendência de Seguros Privados **Anuário Estatístico da SUSEP 1997**. Disponível em : <http://www.susep.gov.br/menu/a-susep/historia-do-seguro>. Acesso em: 01 jun. 2019.

TEIXEIRA, Lívia *et al*. **"Desenvolvimento de material de estudo dos princípios de meteorologia e meio ambiente para estudantes, professores e meios de comunicações"**. [2018]. Disponível em: https://www.cptec.inpe.br/glossario.shtml. Acesso em: 10 fev. 2020.

Testes De Hipóteses E Intervalos De Confiança Em Modelos Lineares Generalizados. 2020. Disponível em: https://docs.ufpr.br/~taconeli/CE225/Aula9.pdf. Acesso em: 14 abr. 2020.

THE POSTGRESQL GLOBAL DEVELOPMENT GROUP. **PostgreSQL Documentation**. Disponível em: https://www.postgresql.org/. Acesso em: 12 jan. 2020.

WIELAND, F. "Limits to growth: results from the detailed policy assessment tool [air traffic congestion]," **16th DASC. AIAA/IEEE Digital Avionics Systems Conference. Reflections to the Future. Proceedings**, Irvine, CA, USA, 1997, pp. 9.2-1, doi: 10.1109/DASC.1997.637296.

ZANINI, Alexandre. **REGRESSÃO LOGÍSTICA E REDES NEURAIS ARTIFICIAIS:UM PROBLEMA DE ESTRUTURA DE PREFERÊNCIA DO CONSUMIDOR E CLASSIFICAÇÃO DE PERFIS DE CONSUMO**. 2007. 15 f. Dissertação (Mestrado) - Curso de Economia Aplicada, Ufjf, Juiz de Fora, 2007.

ANEXOS

ANEXO A - Códigos de dias da semana nos dados

Código Dia da semana	**Dia da semana**
0	Domingo
1	Segunda
2	Terça
3	Quarta
4	Quinta
5	Sexta
6	Sábado

ANEXO B - Códigos de meses nos dados

Código Mês	**Mês**
1	Janeiro
2	Fevereiro
3	Março
4	Abril
5	Maio
6	Junho
7	Julho
8	Agosto
9	Setembro
10	Outubro
11	Novembro
12	Dezembro

ANEXO C - Aeroportos disponíveis no conjunto de dados (122)

Código ICAO	**Nome Aeroporto**	**Cidade**
SBAE	Bauru/Arealva	Arealva
SBAR	Santa Maria	Aracaju
SBAT	Deputado Benedito Santiago	Alta Floresta
SBAU	Estadual Dario Guarita	Araçatuba
SBAX	Romeu Zema	Araxã
SBBE	Val De Cans	Belém
SBBG	Comandante Gustavo Kraemer	Bagé
SBBH	Pampulha - Carlos Drummond De Andrade	Belo Horizonte
SBBR	Presidente Juscelino Kubitschek	Brasília
SBBV	Atlas Brasil Cantanhede	Boa Vista

SBBW	Barra Do Garças	Barra Do Garças
SBCA	Adalberto Mendes Da Silva	Cascavel
SBCB	Cabo Frio	Cabo Frio
SBCF	Tancredo Neves	Confins
SBCG	Campo Grande	Campo Grande
SBCH	Serafin Enoss Bertaso	Chapecã
SBCJ	Carajás	Parauapebas
SBCM	DiomãCio Freitas	Forquilhinha
SBCN	Nelson Rodrigues Guimarães	Caldas Novas
SBCP	Bartolomeu Lisandro	Campos Dos Goytacazes
SBCR	Corumbã	Corumbã
SBCT	Afonso Pena	São José Dos Pinhais
SBCX	Regional Hugo Cantergiani	Caxias Do Sul
SBCY	Marechal Rondon	Várzea Grande
SBCZ	Cruzeiro Do Sul	Cruzeiro Do Sul
SBDB	Bonito	Bonito
SBDN	Presidente Prudente	Presidente Prudente
SBDO	Aeroporto Muncipal De Dourados	Dourados
SBEG	Eduardo Gomes	Manaus
SBFI	Cataratas	Foz Do Iguaçu
SBFL	Hercílio Luz	Florianópolis
SBFN	Fernando De Noronha	Fernando De Noronha
SBFZ	Pinto Martins	Fortaleza
SBGL	Aeroporto Internacional Do Rio De Janeiro/Galeão – Antonio Carlos Jobim	Rio De Janeiro
SBGO	Santa Genoveva	Goiânia
SBGR	Guarulhos - Governador André Franco Montoro	Guarulhos
SBGV	Coronel Altino Machado	Governador Valadares
SBHT	Altamira	Altamira
SBIH	Itaituba	Itaituba
SBIL	Bahia - Jorge Amado	Ilhéus
SBIP	Usiminas	Santana Do Paraíso
SBIZ	Prefeito Renato Moreira	Imperatriz
SBJA	Regional Sul	Jaguaruna
SBJI	Ji-Paraná	Ji-Paraná
SBJP	Presidente Castro Pinto	Bayeux
SBJU	Orlando Bezerra De Menezes	Juazeiro Do Norte
SBJV	Lauro Carneiro De Loyola	Joinville
SBKG	Presidente João Suassuna	Campina Grande
SBKP	Viracopos	Campinas
SBLE	Horácio De Mattos	Lençóis
SBLJ	Correia Pinto	Lages
SBLO	Governador José Richa	Londrina

SBMA	Marabá	Marabá
SBMG	Regional De Maringá - Sílvio Name Júnior	Maringã
SBMK	Mário Ribeiro	Montes Claros
SBML	Frank Miloye Milenkovich	Marília
SBMO	Zumbi Dos Palmares	Rio Largo
SBMQ	Macapá	Macapá
SBMS	Dix-Sept Rosado	Mossorá
SBNF	Ministro Victor Konder	Navegantes
SBNM	Santo Ângelo	Santo Ângelo
SBPA	Salgado Filho	Porto Alegre
SBPB	Prefeito Doutor João Silva Filho	Parnaíba
SBPC	Embaixador Walther Moreira Salles	Poços De Caldas
SBPF	Lauro Kurtz	Passo Fundo
SBPJ	Brigadeiro Lysias Rodrigues	Palmas
SBPK	Pelotas	Pelotas
SBPL	Senador Nilo Coelho	Petrolina
SBPS	Porto Seguro	Porto Seguro
SBPV	Governador Jorge Teixeira De Oliveira	Porto Velho
SBQV	Vitória Da Conquista	Vitória Da Conquista
SBRB	Plácido De Castro	Sena Madureira
SBRD	Rondonópolis	Rondonópolis
SBRF	Guararapes - Gilberto Freyre	Recife
SBRJ	Santos Dumont	Rio De Janeiro
SBRP	Leite Lopes	Ribeirão Preto
SBSG	São Gonçalo Do Amarante	São Gonçalo Do Amarante
SBSJ	Professor Urbano Ernesto Stumpf	São José Dos Campos
SBSL	Marechal Cunha Machado	São Luís
SBSM	Santa Maria	Santa Maria
SBSN	Maestro Wilson Fonseca	Santarém
SBSP	Congonhas	São Paulo
SBSR	Professor Eriberto Manoel Reino	São José do Rio Preto
SBSV	Deputado Luís Eduardo Magalhães	Salvador
SBTB	Trombetas	Oriximiñã
SBTE	Senador Petrônio Portella	Teresina
SBTF	Tefé	Tefé
SBTG	Aeroporto Regional Plínio Alarcon	Três Lagoas
SBTT	Tabatinga	Tabatinga
SBTU	Tucuruã	Tucuruã
SBUA	São Gabriel Da Cachoeira	São Gabriel Da Cachoeira
SBUF	Paulo Afonso	Paulo Afonso
SBUG	Rubem Berta	Uruguaiana
SBUL	Tenente-Coronel Aviador César Bombonato	Uberlândia
SBUR	Mário De Almeida Franco	Uberaba

SBUY	Urucu	Coari
SBVG	Major Brigadeiro Trompowsky	Varginha
SBVH	Brigadeiro Camarão	Vilhena
SBVT	Eurico De Aguiar Salles	Vitória
SBZM	Regional Da Zona Da Mata	Goianá
SJRG	Rio Grande	Rio Grande
SNBR	Barreiras	Barreiras
SNDT	Diamantina	Diamantina
SNDV	Brigadeiro Cabral	Divinópolis
SNPD	Patos De Minas	Patos De Minas
SNPJ	Patrocínio	Patrocínio
SNTF	Teixeira De Freitas	Teixeira De Freitas
SNTO	Juscelino Kubitscheck	Teófilo Otoni
SNUI	Araçuaí	Araçuaí
SNVB	Valença	Valença
SNZA	Pouso Alegre	Pouso Alegre
SSKW	Cacoal	Cacoal
SSZR	Santa Rosa	Santa Rosa
SWBC	Barcelos	Barcelos
SWCA	Carauari	Carauari
SWEI	Eirunepé	Eirunepé
SWGN	Araguaína	Araguaína
SWKO	Coari	Coari
SWLB	Lárea	Lábrea
SWLC	General Leite De Castro	Rio Verde
SWPI	Júlio Bélem	Parintins
SWSI	Presidente João Batista Figueiredo	Sinop

ANEXO D - Justificativas disponíveis no conjunto de dados (8)

Sigla OACI	**Justificativa**
AA	Atraso Aeroporto De Alternativa - Ordem Técnica
AF	Facilidades Do Aeroporto - Restrições De Apoio
AG	Migração/Alfândega/Saúde
AI	Aeroporto De Origem Interditado
AJ	Aeroporto De Destino Interditado
AM	Atraso Aeroporto De Alternativa - Condições Meteorológicas
AR	Aeroporto Com Restrições Operacionais

AS	Segurança/Pax/Carga/Alarme
AT	Liberação Serv. Trafego Aéreo/Antecipação
DF	Avaria Durante Operações Em Vôo
DG	Avaria Durante Operações Em Solo
FP	Plano De Vôo - Aprovação
GF	Abastecimento/Destanqueio
HA	Autorizada
HD	Antecipação De Horário Autorizada
HI	Antecipação De Horário Autorizada Específico Vôos Internacionais
IR	Inclusão De Etapa (Aeroporto De Alternativa) Devido A Um Vôo Especial Retorno
MA	Falha Equipo Automotivo E De Atendimento De Pax
MX	Atrasos Não Específicos - Outros
OA	Autorizado
RA	Conexão De Aeronave
RI	Conexão Aeronave/Volta Vôo De Ida Não Penalizado Aeroporto Interditado
RM	Conexão Aeronave/Volta Vôo De Ida Não Penalizado Condições Meteorológicas
TC	Troca De Aeronave
TD	Defeitos Da Aeronave
VR	Vôo Especial De Retorno (Exclusivo Para Retorno Ao Aeroporto De Origem)
WA	Alternativa Abaixo Dos Limites
WI	Degelo E Remoção De Neve E/Ou Lama Em Aeronave
WO	Aeroporto Origem Abaixo Dos Limites
WR	Atraso Devido Retorno - Condições Meteorológicas

WS	Remoção Gelo/Água/Lama/Areia - Em Aeroporto
WT	Aeroporto Destino Abaixo Dos Limites