CENTRO FEDERAL DE EDUCACAÇÃO TECNOLÓGICA
CELSO SUCKOW DA FONSECA

Análise da Propagação de Atrasos na Malha Aérea Brasileira Usando Padrões Frequentes

Lara Mello e Luana Fragoso

Prof. Orientador:
Eduardo Soares Ogasawara, D.Sc.

Rio de Janeiro,
Dezembro de 2016

CENTRO FEDERAL DE EDUCACAÇÃO TECNOLÓGICA
CELSO SUCKOW DA FONSECA

Análise da Propagação de Atrasos na Malha Aérea Brasileira Usando Padrões Frequentes

Lara Mello e Luana Fragoso

Projeto final apresentado em cumprimento às normas do Departamento de Educação Superior do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ, como parte dos requisitos para obtenção do título de Bacharel em Ciência da Computação.

Prof. Orientador:
Eduardo Soares Ogasawara, D.Sc.

Rio de Janeiro,
Dezembro de 2016

Ficha catalográfica elaborada pela Biblioteca Central do CEFET/RJ

M527 Mello, Lara
Análise da propagação de atrasos na malha aérea brasileira usando padrões frequentes / Lara Mello, Luana Fragoso.—2016.
x, 45f. : il. (algumas color.) , grafs. , tabs. ; enc.

Projeto Final (Graduação) Centro Federal de Educação Tecnológica Celso Suckow da Fonseca , 2016.
Bibliografia : f. 43-45
Orientador : Eduardo Soares Ogasawara

1. Computação. 2. Percepção de padrões. 3. Aeroportos – Brasil. 4. Mineração de dados. 5. Aeronáutica – Vôos. I. Fragoso, Luana. II. Ogasawara, Eduardo Soares (Orient.). III. Título.

CDD 004

AGRADECIMENTOS

Agradece-se as contribuições de Alice Sternberg que deu início as pesquisas sobre o tema abordado.

RESUMO

Esse trabalho tem como objetivo analisar o cenário da propagação de atrasos entre os aeroportos brasileiros, verificando a identificação dos aeroportos mais influentes. Entender melhor a propagação dos atrasos no país tem grande importância pois pode ajudar tanto as empresas e aeroportos quanto os passageiros a tomarem melhores decisões. Para esse fim, foi aplicado um processo de mineração de dados, utilizando o método de mineração de dados de Padrões Frequentes e o algoritmo Apriori, nos dados provenientes da Agência Nacional de Aviação Civil (ANAC), o qual utilizamos informações sobre os voos domésticos brasileiros desde Janeiro de 2009 até Fevereiro de 2015. A análise da propagação dos atrasos foi possível por meio das regras de associação geradas pelo algoritmo Apriori. Por fim, os resultados obtidos são interessantes, como o fato dos aeroportos Guarulhos e Galeão, situados em São Paulo e Rio de Janeiro, respectivamente, serem os que mais influenciam outros aeroportos da rede. Ambos aeroportos são importantes na malha aérea brasileira, sendo Guarulhos o maior aeroporto em escala nacional e o Galeão, o mais movimentado do Rio de Janeiro.

Palavras-chave: atraso de voo; regras de associação; padrões frequentes; propagação de atrasos

ABSTRACT

This work aims to analyze the scenario of delay propagation between Brazilians airports, highlighting the identification of the most influential airports. Better understanding of delays propagation in the country is important because it can help both businesses, airports and also passengers to make better decisions. For this purpose, we automated a mining data process performing the Frequent Pattern method and the algorithm Apriori, using the data from the National Civil Aviation Agency (ANAC), which we use information about brazilian domestic flights from January 2009 to February 2015. The analysis of delays propagation was possible through the association rules generated by the Apriori algorithm. The final results are interesting: like the fact that Guarulhos and Galeão airports, located in São Paulo and Rio de Janeiro, respectively, are the ones that most influence other airports in the network. Both major airports in Brazil, with Guarulhos being the largest airport on a national scale and Galeão, the busiest of the Rio de Janeiro city.

Keywords: flight delays; associations rules; frequent pattern; delay propagation

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

Capítulo 1

Introdução

Voos atrasados têm se mostrado bastante relevante no cenário mundial, independente da grandeza da malha aérea. Em 2013, 36% dos voos atrasaram mais de cinco minutos na Europa, 31.1% dos voos atrasaram mais de 15 minutos nos Estados Unidos e 16.3% dos voos foram cancelados ou sofreram atrasos maiores que 30 minutos no Brasil [EUROCONTROL, 2015].

Atrasos em voos têm consequências econômicas negativas para todos os envolvidos no processo: passageiros, companhias aéreas e aeroportos. Passageiros geralmente planejam viajar várias horas antes de seus compromissos, aumentando o custo de sua viagem para garantir sua chegada pontual e, às vezes, são obrigados a planejar novamente seus itinerários. Por outro lado, companhias aéreas sofrem tanto com multas e custos adicionais de operação quanto com a retenção de aeronaves e funcionários nos aeroportos [Britto et al., 2012].

Observando o cenário de atrasos no Brasil, tem-se em média 155.000 voos atrasados por ano, o que corresponde a aproximadamente 30% do total da quantidade média de voos anuais. É um valor significativo, o que torna o estudo dos atrasos e sua propagação no Brasil interessante. Sem citar que um dos principais aeroportos do mundo e também referência na América Latina está presente na malha aérea brasileira. Inaugurado em 20 de janeiro de 1985, o Aeroporto Internacional de São Paulo/Guarulhos - Governador André Franco Montoro (GRU), hoje em dia, se destaca no cenário mundial e consequentemente se apresenta como o aeroporto mais importante quando mencionamos de voos nacionais. Encontrar a exata relação que o gigante GRU tem com os demais aeroportos é bastante motivador, o que também geraria conhecimento sobre a capacidade de pequenos aeroportos suportarem alguma eventualidade que venha a causar atrasos.

Basicamente, uma operação típica de um voo comercial começa em áreas terminais e pistas de decolagem do aeroporto de partida, passa pelo espaço aéreo e termina nas pistas e áreas terminais dos aeroportos de chegada, sendo suscetível a diferentes tipo de atrasos [Reynolds-Feighan and Button, 1999; Hunter et al., 2007; AhmadBeygi et al., 2008]. Podemos citar como exemplo problemas mecânicos, condições climáticas e as filas de pista.

As operações são repetidas várias vezes ao longo do dia para cada voo no sistema. Devido

aos repousos legais e planos de manutenção, aeronaves, pilotos e comissários de bordo podem seguir itinerários diferentes. Assim, quando alguma rotina se altera por algum motivo, o fluxo pode ser prejudicado e atrasar os voos subsequentes da companhia aérea. Além disso, interrupções podem gerar congestionamento no espaço aéreo ou em outros aeroportos, atrasando alguns voos de outras companhias aéreas também [Xu et al., 2005; Pyrgiotis et al., 2013]. A propagação de atrasos entre aeroportos é um tema importante e abordado principalmente na Europa e nos EUA, com aplicação de diversos métodos a fim de analisar esse fenômeno.

Portanto, todo o ecossistema de voos precisa ser entendido de uma forma melhor. Para isso ser possível, grande volume de dados comerciais é coletado a cada momento e armazenado em várias bases de dados. A fim de extrair informações úteis desses dados, analistas e cientistas de dados estão se esforçando em intensificar suas habilidades computacionais e de gestão de dados.

Uma técnica bastante conhecida para extração de conhecimento é a mineração de Padrões Frequentes, utilizada principalmente para procurar padrões na grande quantidade de dados. A descoberta de novos e diferenciados padrões possibilita um novo e interessante conhecimento sobre os dados. Basicamente, esse método de mineração procura por relações recorrentes em um grande conjunto de dados. Esses dados podem ser divididos em conjuntos distintos tendo como base alguma característica em comum ou algum atributo que queira ser melhor analisado. Já as relações são obtidas através das regras de associação, geradas durante esse processo de mineração.

Há diversas técnicas para mineração de Padrões Frequentes. Os diversos algoritmos se diferem na forma como procuram pelos padrões em tempo de execução [Han et al., 2011], porém acabam chegando ao mesmo resultado, ou seja, os mesmos conjuntos frequentes. Como exemplo podemos citar: Apriori [Agrawal et al., 1994], FP-growth [Han et al., 2000] e Eclat [Zaki et al., 1997]. Quando consideramos regras de associação, o Apriori é um método iterativo bastante conhecido para mineração de conjuntos frequentes [Sternberg et al., 2016] e se baseia na seguinte propriedade: dado um conjunto frequente, todos seus subconjuntos não vazios também devem ser frequentes.

Nossa pesquisa busca uma análise de propagação de atrasos nos aeroportos brasileiros com a intenção de melhor entender o impacto de atrasos entre os aeroportos e neles mesmos, sendo assim, o método de mineração de Padrões Frequentes foi escolhido pois atende as nossas necessidades em encontrar correlações (decorrente das regras de associação) entre os aeroportos e

neles mesmos, a partir da ocorrência de um atraso. Através dessa pesquisa queremos responder as seguintes perguntas: (i) Os aeroportos estão interligados quando um atraso ocorre? (ii) Por quanto tempo um atraso se propaga entre os aeroportos? (iii) Como é o comportamento da propagação dentro do próprio aeroporto?

Além dessa introdução, o trabalho se divide em mais cinco outras seções. A seção 2 apresenta a revisão da literatura. A seção 3 apresenta os trabalhos relacionados. A metodologia propriamente dita é apesentada na seção 4. Os resultados encontrados são analisados na seção 5. Finalmente, a seção 6 encerra a dissertação desse trabalho.

Capítulo 2

Fundamentação Teórica

Para desenvolver essa pesquisa, precisamos entender o cenário da propagação de atrasos nos voos, os métodos de pré-processamento utilizados e os conceitos de janela deslizante e Padrões Frequentes. Nas seções a seguir, a partir da seção 2.1 até a 2.4, a base teórica para cada item citado é apresentada respectivamente.

2.1 Cenário da Propagação de Atraso nos Voos

O sistema comercial de aviação é bastante complexo porque envolve, diariamente, vários recursos, como aviões e tripulação, e responsabilidade, como segurança e confiabilidade. Além disso, é um sistema amplo, pois é composto de vários aeroportos e companhias aéreas, ambos interligados entre si. Por exemplo, 65 aeroportos brasileiros são responsáveis por 98% do tráfego aéreo nacional [Secretaria de Aviação Civil, 2015], sendo as principais companhias que realizam esses voos a TAM, GOL, Avianca e Azul.

Para um voo acontecer, várias operações devem ser realizadas pelos aeroportos, companhias e passageiros. A Figura 1 ilustra uma operação típica de um voo comercial com etapas em terminais, pistas ou no espaço aéreo, estando, portanto, suscetível a diversos tipos de atraso.

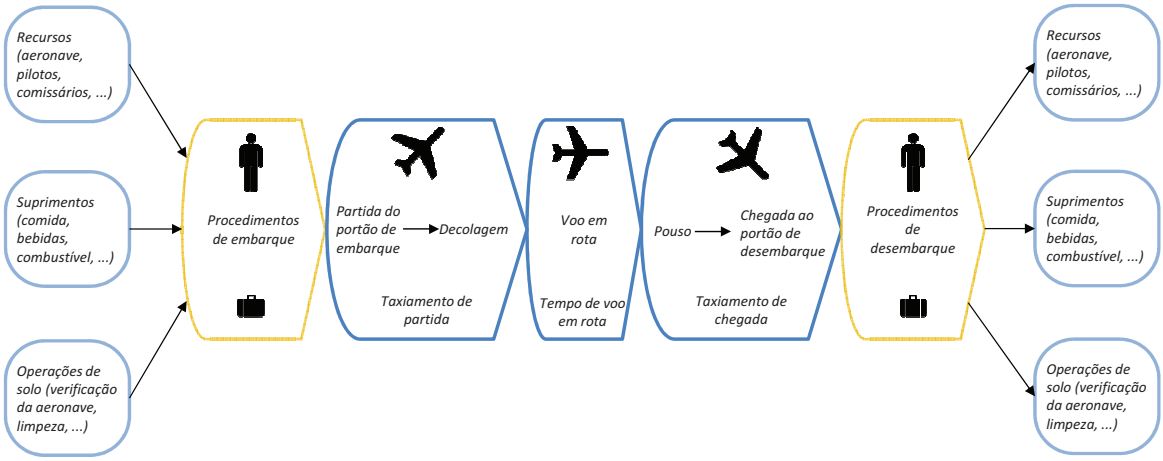

Figura 1: Operações típicas de um voo comercial

Atrasos ou cancelamentos causados por vários motivos podem afetar um voo. Tempestades, neve, neblina, entre outras causas meteorológicas podem originar os atrasos ou cancelamentos. Outro cenário é quando o aeroporto está lotado e um voo precisa aterrissar. Esse voo irá esperar até que um espaço fique disponível. Tal fenômeno pode aumentar o número de embarques e desembarques em um período, gerando problemas de capacidade e de fila de espera.

Por outro lado, recursos críticos das companhias aéreas, como as aeronaves e a tripulação, costumam ser alocados para mais de um voo em um mesmo dia. Ou seja, o esquema da Figura 1 se repete várias vezes ao longo do mesmo dia. Também deve-se considerar que tais recursos não são compartilhados apenas por uma companhia, mas também para outras companhias e aeroportos. Esse modo de compartilhamento de recursos é a principal causa para gerar a propagação do atraso em vários aeroportos.

Em relação à propagação de atrasos, presume-se que um atraso já ocorreu num ponto da rede. O principal estudo da propagação de atraso nos voos é entender a conexão entre os aeroportos diante de um atraso ocorrido. Outro ponto a ser considerado é que um atraso pode influenciar outro atraso horas depois como também pode continuar influenciando durante um determinado período de tempo.

Portanto, a compreensão dos atrasos e da sua propagação é um fator importante para o suporte na tomada de decisão das companhias e aeroportos, que podem direcionar seus investimentos na melhoraria do planejamento dos seus recursos críticos e no auxílio dos passageiros, na organização de suas viagens.

2.2 Pré-Processamento de Dados

A etapa de pré-processamento de dados é importante pois lida com a redundância, incompletude e inconsistência dos dados que são geralmente encontrados em bases de dados reais. Além disso, as fases do pré processamento devem também preparar os dados para a implementação de um modelo mais eficiente e para facilitar a compreensão de seus resultados [Han et al., 2011]. Os próximos parágrafos descrevem as atividades de limpeza, seção 2.2.1, transformação dos dados, seção 2.2.2, e materialização, seção 2.2.3, que foram utilizadas durante a pesquisa.

2.2.1 Limpeza dos Dados

Limpeza de dados é o processo que detecta e corrige dados com valores faltantes e discrepantes. Esse processo possui várias técnicas para tratar esses tipos de dados, que serão discutidas com mais detalhes nos próximos parágrafos dessa subseção. Trabalhar com dados incorretos ou inconsistentes pode resultar em falsas conclusões, podendo causar sérios prejuízos em escala pública e privada.

Por exemplo, empresas registram dados de seus clientes acerca do nome, endereço, e-mail, entre outras informações. Se um desses dados estiver inconsistente por algum motivo, contatar esse cliente pode ser uma tarefa árdua e até mesmo impossível, podendo ocasionar a perda desse contato. A empresa poderia ser prejudicada em suas vendas, pois não conseguiria informar sobre promoções, pagamentos e avaliações.

Já no ponto de vista do cenário dos atrasos aéreos, dados inconsistentes, incorretos ou discrepantes como o horário previsto de partida ou o horário real de partida seria um grande problema, pois é na diferença entre esses horários que o atraso é calculado. Logo, é importante tratar esses dados para evitar influências negativas nos resultados.

O primeiro passo do processo de limpeza de dados é encontrar valores discrepantes [Han et al., 2011]. Valores discrepatantes (ou *outliers*) são dados que possuem valores atípicos em relação às demais observações da base de dados. Essas discrepâncias podem ser ocasionadas por diversos motivos como uma modelagem ruim dos dados ou erro humano na entrada do dado.

A detecção de discrepância é o processo de encontrar valores que não são permitidos pelo modelo. Sua presença acaba interferindo negativamente na análise dos dados. Para detectar os *outliers*, deve-se usar qualquer conhecimento prévio dos dados como ponto incial. Analisar os dados estatisticamente é uma das técnicas para conhecer os registros que se está trabalhando. Portanto, a partir de um levantamento estatístico, como média, mediana, moda, desvio padrão, análise de quartis e etc, pode ser possível verificar as tendências dos dados e identificar anomalias. Usando a análise por quartil, um *outlier* é um valor entre $Q1 - 3IQR$ ou acima de $Q3 + 3IQR$, onde IQR é a variedade do intervalo do quartil e $Q1$ e $Q3$ são o primeiro e terceiro quartis, respectivamente. Uma maneira comum de lidar com as tarefas de limpeza é remover da base de dados as tuplas que contém discrepâncias [Han et al., 2011].

Também existem vários métodos para tratar os valores faltantes, dentre eles ignorar as tu-

plas. Neste método, como o nome já diz, as tuplas com valores faltando são ignoradas, ou seja, não são usadas para realizar a análise. Além disso, os atributos restantes da tupla não são considerados quando a mesma é ignorada, o que pode ser ruim para alguns casos, já que tais informações podem ser úteis. Contudo, é um método mais eficaz quando se trata de muitas tuplas apresentando dados faltantes [Han et al., 2011].

2.2.2 Transformação dos Dados

Transformação de dados é o processo de converter dados ou informações de um formato para outro na intuição de consolidar e preparar esses dados em um formato apropriado para a mineração. Existem várias técnicas para realizar a transformação, dependendo do atributo a ser transformado e também dos resultados esperados. *Binning* e agregação temporal são algumas das técnicas existentes para esse processo.

Agregação temporal é aplicada para estudar uma série temporal num nível agregado [Tiao, 1972]. Um exemplo seria o registro de vendas por trimestre desde o ano de 2008 até 2010. Porém, a análise a ser realizada em cima desses dados está interessada em vendas anuais, em vez de trimestrais. Logo, os dados poderiam ser agregados para representar as vendas por ano. Essa agregação é ilustrada na Figura 2. O conjunto de dados resultantes é menor em tamanho mas não tem perda de informação para a análise desejada [Han et al., 2011].

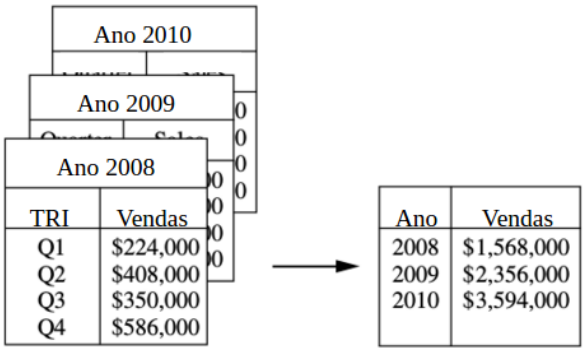

Figura 2: Dados de vendas entre os anos de 2008 e 2010. Na esquerda, as vendas estão representadas por trimestre (atributo TRI). Na direita, os dados são agregados para fornecer as vendas por ano. Fonte: adaptado de [Han et al., 2011]

No cenário da aviação, pode-se ter de exemplo uma série temporal na qual os dados são armazenados a cada minuto. Uma transformação de agregação é aplicada para transformar essa série em uma série temporal com observações a cada hora. Logo, seja x_t uma série das porcentagens de atrasos num determinado aeroporto a cada minuto, essa série temporal pode ser

agregada em y_t, a qual representa uma série das porcentagens de atrasos de voos num aeroporto a cada hora.

Outra forma de transformação de dados, além da agregação temporal, é a técnica de discretização *binning*. Antes de ententer o conceito de *binning*, vale ressaltar o que é a técnica de discretização. Discretização é a transformação de um atributo, por exemplo, idade, em intervalos, como 0-10, 11-20, etc., ou em categorias conceituais (jovem, adulto e idoso) [Han et al., 2011].

Por fim, *binning* é uma técnica de discretização na qual um conjunto de dados é separado em intervalos denominados *bin*, onde cada valor desse conjunto pertence a apenas um *bin*. Assim, esses valores podem ser representados e substituídos pelo rótulo desse *bin* [Sternberg et al., 2016]. Nesse caso, o número de *bins* é um ponto sensível para a transformação [James et al., 2013]. Para resolver esse problema, o número de *bins* é encontrado através da curvatura máxima de uma regressão *spline* sobre o erro de cada observação em relação à média do seu respectivo *bin* [Burden et al., 2015].

2.2.3 Materialização

Uma forma simples de aplicar várias operações, sequencialmente, de uma consulta do banco de dados é montar tabelas intermediárias como resultado de cada operação. A criação de uma tabela intermediária no banco se chama materialização. Em sua grande maioria, essa tabela é escrita no disco rígido, exceto nos casos em que as tabelas são pequenas [Silberschatz et al., 2010].

A Figura 3 mostra uma consulta com várias operações. A consulta é executada a partir da folha da árvore (sua base) até a raíz (o topo da árvore). Nessa consulta, a primeira operação é uma seleção. O resultado dessa seleção é uma tabela intermediária que será usada no nível logo acima da árvore, no qual os *inputs* podem ser outra tabela intermediária ou relações da base de dados [Silberschatz et al., 2010]. No caso do exemplo da Figura 3, é executado uma junção entre a tabela intermediária *departamento* e a relação do banco *instrutor*. O resultado dessa junção é outra tabela intermediária. Esse processo de criação de tabelas intermediárias, no intuito de realizar várias operações de uma consulta, é executado até a raíz da árvore, provendo o resultado final da consulta.

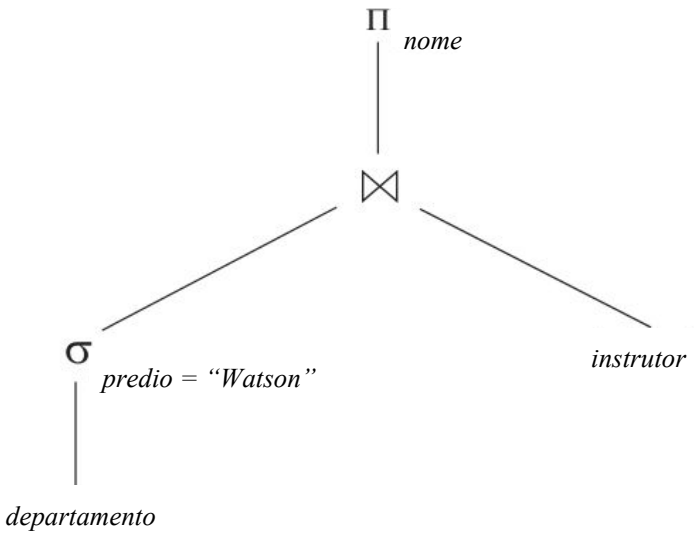

Figura 3: Consulta num formato de árvore. Fonte: adaptado de [Silberschatz et al., 2010]

2.3 Janela Deslizante

Um grupo de modelos bastante usado para lidar com grande fluxo de dados são os modelos de janela de tempo. São eles: *damped window*, *landmark window* e a janela deslizante. Os três modelos se apresentam como eficientes, porém cada um deve ser usado para uma determinada característica. No modelo *landmark*, a mineração é feita considerando todos os dados de um intervalo de tempo, ou seja, a partir de um ponto fixo escolhido no tempo até o tempo atual, sendo assim, a janela tende a ir aumentando. No modelo *damped window*, pesos diferenciados são atribuídos aos dados tendo em consideração a ordem de aparecimento, os dados mais recentes ganham pesos maiores do que os dados antigos. Já nas janelas deslizantes, é considerado no processo de mineração apenas os dados dentro do comprimento fixo da janela naquele momento. À medida que novas transações chegam, as transações mais antigas na janela deslizante expiram [Ahmed et al., 2012].

O modelo de janela deslizante tenta lidar com a mudança de conceito considerando somente os dados recebidos recentemente [Deypir et al., 2012]. Devido a ilimitada quantidade de transações recentes e da quantidade de memória ser limitada, a janela deslizante deve ser limitada. Janela deslizante é um dos principais modelos para processamento e mineração de dados *stream*, no qual um tamanho fixo dos dados recebidos recentemente é considerado [Deypir and Sadreddini, 2011].

No ponto de vista para a aplicação do método de mineração de Padrões Frequentes, a janela deslizante não representa corretamente os *itemsets* se o tamanho da janela for maior ou menor

que o ideal. Em uma janela maior que o ideal, se o suporte de um *itemset* que aparece mais vezes em transações recentes for baixo, esse *itemset* deve ser reconhecido como não-frequente, devido a sua baixa frequência nas transações antigas da janela [Deypir et al., 2012]. Já o suporte de um *itemset* que não aparece nas transações recentes deve se manter como frequente devido ao seu alto suporte nas transações mais antigas da janela. Por outro lado, em uma janela menor que o ideal, quando os dados de entrada tem um conceito estável, o resultado da mineração é uma pobre aproximação dos *itemset* frequentes mais recentes, desde que o resultado seja obtido a partir de uma limitada quantidade de transações. Dito isso, podemos concluir que a definição do tamanho da janela a ser utilizada é extremamente importante.

2.4 Padrões Frequentes

Mineração de Padrões Frequentes é um método de extração de conhecimento que busca encontrar padrões que se repetem nos dados [Sternberg et al., 2016]. Com a enorme quantidade de dados sendo armazenada nos bancos de dados de empresas, elas desenvolveram um interesse muito grande no uso de Padrões Frequentes para minerar esses dados, com o objetivo de encontrar correlações nos registros. A descoberta de correlações interessantes entre os dados pode ajudar em várias tomadas de decisão.

Uma aplicação popular de padrões frequentes está em estabelecimentos de compra onde há o interesse em identificar produtos que são comprados juntos frequentemente [Deypir and Sadreddini, 2011]. Esse processo analisa o hábito das compras do cliente [Han et al., 2011]. Suponha que a execução de padrões frequentes tenha obtido como resultado que um cliente que compra leite tende a comprar pão também. A partir dessa análise, os varejistas podem elabrorar planos estratégicos mais eficientes. Como, por exemplo, posicionar produtos comprados em sequência, como o leite e o pão, mais próximos um do outro para aumentar suas vendas.

Um conjunto de itens que aparecem juntos frequentemente num *dataset* transacional é denomidado de *itemset* [Han et al., 2011]. Um *k-itemset* é um conjunto que contém *k* itens. O conjunto {leite, pão} é um *2-itemset*. Um *k-itemset* é frequente na base de dados se o número de sua ocorrência na mesma for igual ou maior ao suporte e a confiança. Ou seja, *k-itemset* deve satisfazer o mínimo do suporte e da confiança para ser um conjunto frequente. Um suporte de 7% significa que 7% de todas as transações do banco de dados mostram que o leite e o pão são comprados juntos. Em outras palavras, o suporte é a probabilidade (frequência) de um *itemset* ocorrer na base de dados. A confiança é a contagem das transações que tem algum relaciona-

mento entre seus itens. Uma confiança de 70% mostra que 70% dos clientes que compraram leite também compraram pão. Assim, confiança pode ser interpretada como uma probabilidade condicional. Tanto o suporte quanto a confiança são determinados pelo usuário do processo de mineração [Deypir et al., 2012]. Os *k-itemsets* que satisfazem os mínimos do suporte e da confiança são chamados de regras de associação fortes.

Mesmo com o suporte e a confiança definidos, nem todas as regras resultantes são interessantes para a análise. Por exemplo, uma transação com 10000 registros mostra que 6000 transações do cliente incluem compra do leite, enquanto 7500 incluem a compra do pão e 4000 com a compra dos dois itens. Suponha que a mineração de dados de padrões frequentes é executada em cima dessas transações com um suporte mínimo definido em 30% e uma confiança de 60%. As regras geradas analisando a compra do leite e em seguida a compra do pão são fortes. O suporte de quem compra leite e pão é de $\frac{4000}{10000} = 40\%$ e a confiança de quem compra leite também compra pão é de $\frac{4000}{6000} = 66\%$, satisfazendo o mínimo de ambas variáveis. Porém, a compra do pão representa 75% dos dados transacionais, que é maior do que 66% em relação ao cliente que compra leite também compra pão. Nesse caso, a compra do leite diminui a frequência da compra do pão. Ou seja, gerou um efeito negativo ao invés de um positivo, como é do interesse da análise.

Portanto, apenas com o uso das medidas de suporte e confiança não é suficiente para selecionar regras interessantes. Para evitar que essa seleção de regras interessantes seja feita pelos usuários da mineração, tornando-a algo subjetiva, algumas medidas de correlação entre os elementos existem, sendo uma delas a medida *lift*. Essa medida é usada para identificar a correlação entre os *itemsets*. Se o *lift* for maior do que um, indica que os *itemsets* tem relacionamentos positivos entre eles. Ou seja, a aparição de um *itemset* aumenta a aparição do outro.

Em resumo, a mineração de padrões frequentes procura por relações recorrentes em um conjunto de dados. Essas relações são obtidas através das regras de associação geradas na execução do método. Elas são representadas conforme a Equação 2.1, que informa que a ocorrência do *itemset* frequente A = $\{a_1, a_2, ..., a_n\}$ leva a ocorrência de um outro *itemset* frequente B = $\{b_1, b_2, ..., b_n\}$. As regras de associação são geradas a partir de um suporte e uma confiança mínimos estabelecidos, Equação 2.2 e Equação 2.3, respectivamente. Note que na Equação 2.3, a notação $P(A \cup B)$ indica a probabilidade da transação conter a união do conjuntos A e B. Ou seja, de conter todos os itens de A e B. Isso não pode ser confundido com *P(A ou B)*, o qual indica a probabilidade da transação conter os itens de A ou B [Han et al., 2011]. Além disso,

precisa-se saber o quanto esses itens estão correlacionados entre si, usando o conceito de *lift*, representado na Equação 2.4. As fórmulas estão listadas a seguir:

$$regra(A \rightarrow B)[suporte, confiana, lift] \tag{2.1}$$

$$suporte(A \rightarrow B) = P(A \cup B) \tag{2.2}$$

$$confianca(A \rightarrow B) = P(B|A) = \frac{suporte(A \cup B)}{suporte(A)} \tag{2.3}$$

$$lift(A,B) = \frac{confianca(A \rightarrow B)}{suporte(B)} \tag{2.4}$$

Existem vários algoritmos para executar o método de Padrões Frequentes, Apriori é um deles. Ao final de sua execução, produz as regras de associação. O algoritmo se desenvolve identificando os *k-itemset* frequentes, começando com *k=1* na base de dados, e estendendo-os para *(k + 1)-itemsets*, até que a aparição desses *itemsets* seja suficientemente frequente na base de dados. Quando *k=1*, temos grupos como { A }, { B }, { C }. Para *k=2*, temos { A, B }, { A, C }, { B, C } e assim por diante. O processo executado pelo Apriori é explicado com mais detalhes no próximo parágrafo.

Primeiramente, o conjunto de *1-itemset* frequente é encontrado através de uma varredura na base de dados com a intenção de acumular a contagem de cada item do conjunto. O conjunto formado antes da verificação do suporte mínimo é chamado de conjunto candidato C_k. Os itens de C_1 que satisfazem o suporte mínimo compõem o conjunto denotado de L_1. No segundo passo, é feita uma combinação 2 a 2 de L_1, formando conjuntos de *2-itemsets*, C_2. Uma varredura no banco é realizada novamente para encontrar a contagem desses *2-itemsets*. Os *itemsets* que satisfazem o suporte formam L_2, que através do mesmo processo é usado para encontrar L_3 e assim por diante, até que nenhum outro *k-itemset* possa ser encontrado.

Para encontrar cada L_k é necessário uma varredura completa na base de dados. Contudo, Apriori possui uma propriedade importante para melhorar a eficiência desse processo. Essa propriedade define que cada sub-conjunto de um *itemset* frequente tem que ser frequente também. Assim, o espaço de busca é reduzido [Han et al., 2011].

O processo do algoritmo Apriori, com suporte mínimo igual a 20%, é ilustrado a partir da

Tabela 1 e da Figura 4. A tabela contém as transações com seus respectivos itens e a figura ilustra a execução do algoritmo em cima dessas transações. O processo decorre como foi explicado nos parágrafos anteriores. Note que C_3 não possui todas as combinações de L_2, por conta da propriedade do Apriori mencionada. O conjunto {I2, I4, I5}, por exemplo, não está incluído em C_3 porque o *itemset* {I4, I5} não é frequente, ele não pertence à L_2. Assim, a varredura na base de dados para a contagem desse conjunto é evitada.

Tabela 1: Transações com seus respectivos itens. Fonte: adaptado de [Han et al., 2011]

TID	Lista de itens
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

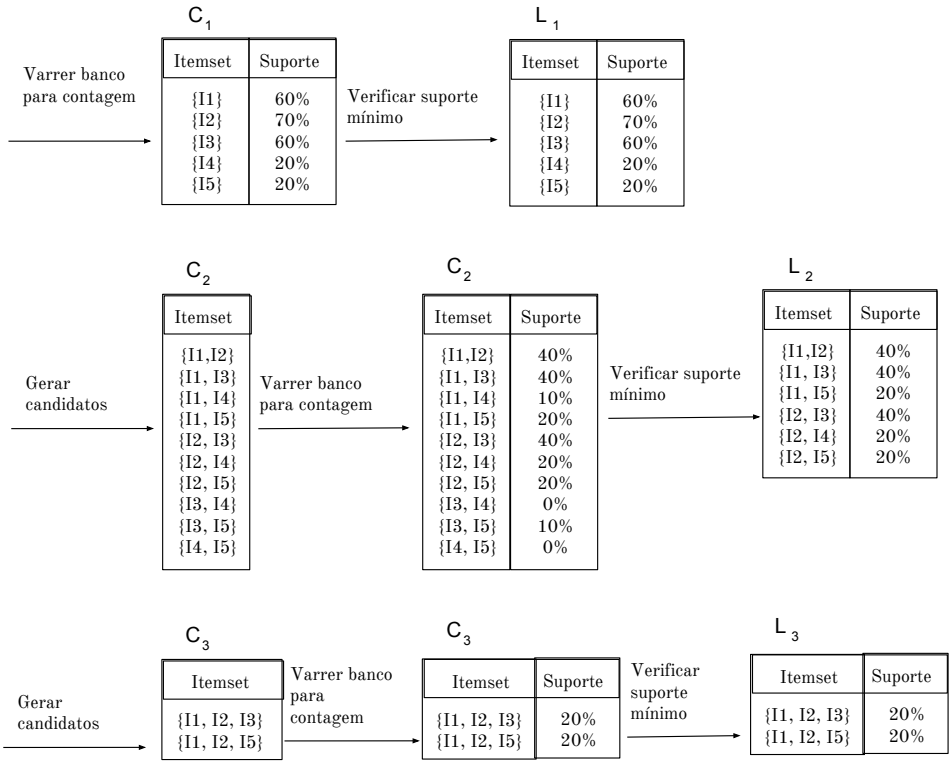

Figura 4: Geração dos *itemsets* candidatos e frequentes com suporte mínimo igual a 2, a partir das transações da Tabela 1. Fonte: adaptado de [Han et al., 2011]

Capítulo 3

Trabalhos Relacionados

Diversos trabalhos a respeito do problema de atraso de voo tem sido desenvolvidos atualmente. Para explorar esse tema na literatura, fizemos uma busca sistemática no Scopus com a seguinte string: *"flight delay"AND "propagation"* e obtivemos em torno de 140 resultados. Após a filtragem dos resumos, pudemos observar as características mais relevantes nas abordagens como introduzidas a seguir. Alguns trabalhos focam na análise desses atrasos, identificando causas, padrões no cenário e eventuais interferências a fim de minificar os danos através de estratégias personalizadas. Diferentes métodos e técnicas são utilizados para tentar melhorar os resultados, como também há trabalhos comparando algumas dessas técnicas. Também há diversos trabalhos que focam na previsão de atrasos ou em sua propagação.

A análise de atrasos de voos é um assunto frequente em artigos sobre aviação comercial. Essa análise visa ajudar a diminuir os impactos dos atrasos, seja identificando padrões de atrasos de determinada rede ou até mesmo prevendo atrasos. Extrair padrões dos atrasos é bastante importante, como concluído em [Meng and Peng, 2015], pois traz referências para tomada de decisão tanto dos aeroportos como das companhias aéreas e, consequentemente, para a prevenção de atrasos.

Como consequências de múltiplos atrasos, temos a propagação e o congestionamento aéreo. Como exemplo de abordagem da última consequência citada, podemos citar [Xu and Li, 2015] que, considerando conceitos de formação de congestionamento e de efeitos posteriores, aplica métodos de identificação de congestionamento baseado em diferentes escalas de tempo e métodos de previsão com base em algoritmos matemáticos. Diversos artigos trabalham a questão da propagação, seja dentro de um país ou até mesmo propagação entre países, como desenvolvido em [Baspinar et al., 2016], analisando os aeroportos europeus mais movimentados e suas ligações com outros aeroportos da rede, conseguindo assim investigar o efeito dos distúrbios locais no tráfego aéreo da Europa.

Devido a sua importância, há diferentes métodos sendo aplicados para melhor entender ou prever a propagação de atrasos de voos. [Zhao et al., 2016] utiliza o conceito de *virtual queue*, baseado da tomada de decisão colaborativa, comparando várias métricas e cenários a fim de

reordenar voos de uma forma que diminua o tempo de espera do passageiro. Já [Qiu et al., 2015] aplica a função *Copula* para estudar a correlação entre uma sequência de atrasos causados pelo mesmo atraso inicial. Também é encontrado na literatura diversos métodos matemáticos e estatísticos para prever ou estimar atrasos, [Tu et al., 2008] utiliza algoritmos de otimização com algumas ideias de algoritmos genéticos.

A maioria dos artigos encontrados tem como alvo as redes de tráfego aéreo da Europa, EUA ou China. Também há trabalhos comparando essas grandes potências, como [Campanelli et al., 2016] que faz uma comparação do modelo de propagação de atrasos entre as redes europeias e norte americanas. Analisando a rede brasileira, temos apenas um artigo, [Sternberg et al., 2016], que faz uma análise sobre o cenário de atrasos levando em consideração diversos atributos como os metereológicos, porém sem focar no estudo da propagação de atrasos.

As tabelas a seguir apresentam os trabalhos relacionados e suas relações com duas características. Tabela 2 indica as técnicas utilizadas pelos trabalhos citados e a Tabela 3 a localização geográfica em qual se baseiam os estudos. Podemos observar uma maioria de trabalhos analisando a malha aerea da Europa e da China, o que sugere o Brasil como um potencial de exploração. Nosso trabalho, além de explorar esse cenário de propagação na malha aérea brasileira, se diferencia pelo uso de Padrões Frequentes.

Tabela 2: Comparação dos trabalhos relacionados em relação às técnicas aplicadas

Trabalho	Algoritmos Matemáticos	Virtual Queue	Copula
[Xu and Li, 2015]	X		
[Tu et al., 2008]	X		
[Zhao et al., 2016]		X	
[Qiu et al., 2015]			X

Tabela 3: Comparação dos trabalhos relacionados em relação à área geográfica estudada

Trabalho	Europa	EUA	China	Brasil
[Baspinar et al., 2016]	X			
[Campanelli et al., 2016]	X	X		
[Xu and Li, 2015]			X	
[Zhao et al., 2016]			X	
[Sternberg et al., 2016]				X

Capítulo 4

Metodologia

4.1 Base de Dados

O Brasil possui um sistema aéreo extenso, com mais de 2000 aeródromos e por volta de 200 milhões de embarques e desembarques por ano [Secretaria de Aviação Civil, 2015]. Existem algumas instituições responsáveis por armazenar e gerenciar esses dados provenientes do sistema aéreo brasileiro, como a Agência Nacional de Aviação Civil (ANAC), uma das agências federais reguladoras do Brasil. Foi criada para regulamentar e supervisionar as atividades da aviação civil, aeronáutica e a infraestrutura dos aeroportos brasileiros.

Por conta da ANAC ser responsável pelo armazenamento dos dados operacionais do sistema aéreo brasileiro, como horários de partida prevista e real, e por consequência, os atrasos ocorridos, utilizamos seu banco de dados para realizar esse trabalho. Esses dados são disponíveis mensalmente num banco de dados público chamado VRA [ANAC, 2015]. Os dados utilizados englobam os anos de 2009 até 2015. O ano de 2009 foi escolhido como sendo o inicial por conta das grandes companhias TAM, Gol, Avianca e Azul terem começado a operar juntas. Além disso, para consolidar nossa análise na relação entre os aeroportos, levamos em consideração os 17 principais aeroportos do país, que foram responsáveis por 80% das partidas realizadas entre os anos de 2009 e 2015 [Sternberg et al., 2016], conforme Tabela 4.

Porém, os dados provenientes da base VRA não contém informações meteorológicas. Apesar de não utilizarmos esses dados para essa pesquisa no momento, temos intenção de incorporar tais informações em trabalhos futuros. Esses dados são obtidos pela empresa *The Weather Company* a partir de seu banco de dados público chamado *Weather Underground* (WU). Além dos dados meteorológicos, também temos interesse em trabalhar na identificação de dias úteis, por conta disso, a classe Feriado é necessária. Logo, o modelo de dados utilizado nesse trabalho está ilustrado na Figura 5. As classes amarelas são as utilizadas nessa pesquisa e as em branco já estão preparadas para futuros trabalhos.

A parte de pré-processamento dos dados foi realizada pela Alice Stenberg [Sternberg et al., 2016]. Nessa base de dados foi aplicada a limpeza dos dados, detectando os *outliers* através

do uso estatístico dos quartis e removendo essas tuplas do banco. Além disso, registros com a hora prevista de partida e chegada em branco, ou seja, com valores faltantes, também foram removidos da base.

Tabela 4: Aeroportos brasileiros analisados. O código ICAO foi utilizado como identificador dos aeroportos nesse trabalho. Fonte: adaptado de [Sternberg et al., 2016]

Código ICAO	Aeroporto	Cidade	Frequência Acumulativa
SBSP	Congonhas	São Paulo	10%
SBGR	Guarulhos - Gov. A. F. Montoro	São Paulo	20%
SBBR	Presidente Juscelino Kubitschek	Brasília	29%
SBGL	Galeão – Antonio Carlos Jobim	Rio de Janeiro	35%
SBRJ	Santos Dumont	Rio de Janeiro	41%
SBCF	Tancredo Neves	Belo Horizonte	47%
SBKP	Viracopos	Campinas	52%
SBSV	Dep. L. E. Magalhães	Salvador	57%
SBCT	Afonso Pena	Curitiba	61%
SBPA	Salgado Filho	Porto Alegre	65%
SBRF	Guararapes - Gilberto Freyre	Recife	69%
SBFZ	Pinto Martins	Fortaleza	72%
SBVT	Eurico de Aguiar Salles	Vitória	74%
SBFL	Hercilio Luz	Florianópolis	76%
SBBE	Val De Cans	Belém	78%
SBGO	Santa Genoveva	Goiânia	79%
SBEG	Eduardo Gomes	Manaus	80%

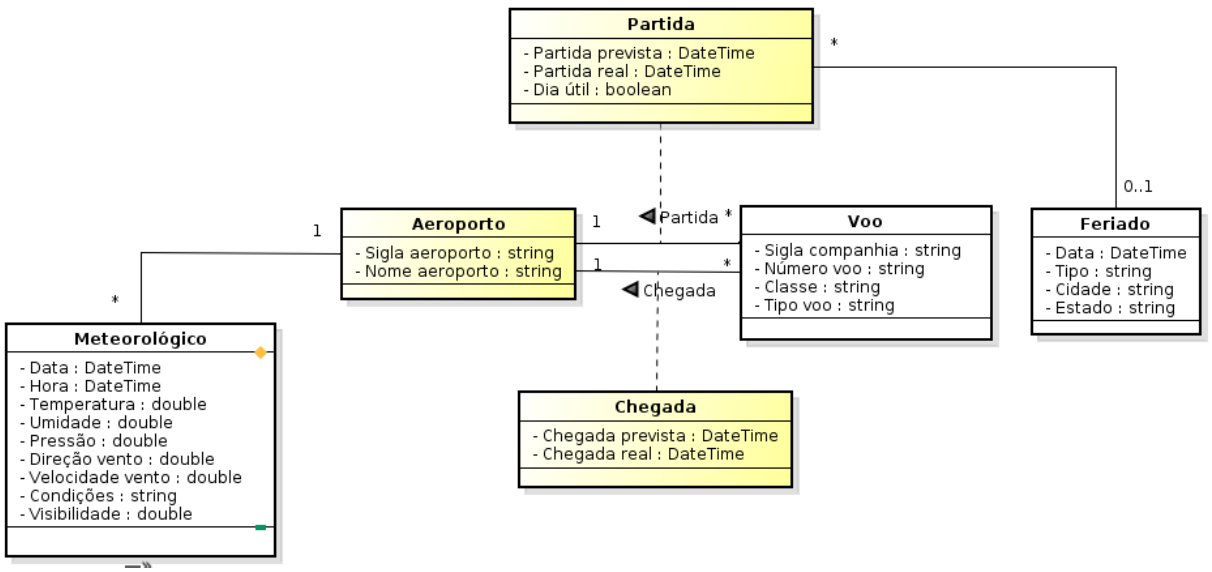

Figura 5: Diagrama de classes com integração dos dados meteorológicos e operacionais. Fonte: adaptado de [Sternberg et al., 2016]

4.2 Processo de Mineração de Dados

O processo de mineração de dados aplicado para esse trabalho é composto por quatro partes principais: (i) indexação dos dados, (ii) materialização das janelas deslizantes, (iii) geração das regras e (iv) análise das regras. Exceto o item (iv), todos os restantes foram executados de forma automática.

O Algoritmo 1 ilustra a metodologia geral aplicada para produzir as regras de associação. Inicialmente, a função *regraAssociacaoAtrasoVoo* é invocada, recebendo como parâmetro um *data warehouse dw*. Essa função executa *indexacaoDados* primeiro e, em seguida, para cada aeroporto da Tabela 4, chama a função *materializacaoJanelaDeslizante* e a *geracaoRegras*.

Algorithm 1 Processo de Mineração de Dados para a Geração de Regras de Associação de Atrasos de Voos

```
1: function REGRAASSOCIACAOATRASOVOO(DW dw)
2:     ids ← indexacaoDados(dw)
3:     for each arpt ∈ dw do
4:         j ← materializacaoJanelaDeslizante(ids, arpt)
5:         return geracaoRegras(j)
6:     end for
7: end function
1: function INDEXACAODADOS(DW dw)
2:     return agregacaoTemporal(dw)
3: end function
1: function MATERIALIZACAOJANELADESLIZANTE(IDW ids, Aeroporto arpt)
2:     r ← rotas(ids, arpt)
3:     p ← percAtrasos(ids, r, arpt)
4:     b ← binning(p)
5:     return b
6: end function
1: function GERARACAOREGRAS(IDW ids)
2:     arp ← parametrosHiper(ids)
3:     ar ← apriori(ids, arp)
4:     iar ← filtracaoRegras(ar)
5:     return iar
6: end function
```

4.3 Indexação de Dados

Como o suporte necessita da frequência de padrões no banco de dados, analisar os atrasos de voos por segundos ou minutos iria dificultar a aparição de regras interessantes, já que a ocor-

rência de voos por minuto ou segundo é bem pequena em comparação a cada hora. Logo, resolvemos agrupar os voos por hora. A função correspondente no Algoritmo 1 é a *indexacaoDados*, que recebe um *data warehouse* como parâmetro. Aplicamos uma transformação de dados nesse *data warehouse* utilizando a técnica de agregação temporal. Selecionamos o atributo partida prevista para realizar essa indexação. O dado original consiste de datas de 01/01/2009 0:00:00 até 28/02/2015 23:59:59. Logo, a estrutura da data do dado original é composta de *dia/mês/ano hora:minuto:segundo*. A partir dessa técnica foi possível agrupar os dados por hora (minutos e segundos foram zerados), mantendo a informação do dia, mês e ano. Assim, a data resultante possui a estrutura *dia/mês/ano hora:00:00*. Para cada intervalo de tempo, é associado aeroportos de partida junto com a quantidade total de voos realizados e com a quantidade de atrasos correspondente, conforme Tabela 5. Consideramos como atraso, voos com saídas depois do horário previsto de partida e voos cancelados. Vale ressaltar que estamos analisando a rede de atrasos apenas na partida.

Tabela 5: Resultado da agregação temporal

dia_hora	aeroporto	voos	atraso
2009-01-01 00:00:00	SBSV	1	0
2009-01-01 00:00:00	SBVT	1	1
2009-01-01 01:00:00	SBEG	1	0
2009-01-01 01:00:00	SBFZ	3	0

4.4 Materialização da Janela Deslizante

A janela deslizante é a etapa em que deixa os dados preparados para executar o algoritmo Apriori. Para conseguirmos encontrar regras interessantes no contexto da propagação de atrasos, criamos uma janela deslizante para cada aeroporto, ou seja, foram formadas 17 janelas deslizantes durante o trabalho.

Os valores da janela correspondem às porcentagens de atrasos de um determinado aeroporto junto com a informação de quanto tempo depois a propagação do atraso ocorreu. Tanto a informação do aeroporto quanto do tempo da propagação são apresentadas no nome dos atributos da janela. Essas porcentagens foram calculadas a partir da Tabela 5 e são agrupadas pela data de partida a cada hora, começando em 01/01/2009 0:00:00 até 28/02/2015 23:00:00.

O primeiro atributo da janela corresponde aos atrasos ocorridos no aeroporto de partida no

horário de partida do voo. O aeroporto de partida é o aeroporto que está sendo analisado no momento. Como dito antes, são geradas 17 janelas no total, uma para cada aeroporto. Representamos a primeira coluna como: *aeroportoPartida_0*. Os atributos restantes correspondem aos aeroportos de destino, ou ao próprio, juntamente com a informação de quanto tempo depois da partida está sendo considerado, o que representa a propagação do atraso. Representamos como: *aeroportoDestino_horasDepois*. Por exemplo, considere que o primeiro atributo da janela é *sbcf_0* e o terceiro, *sbvt_1*. Isso significa que as porcentagens de atraso apresentadas na coluna *sbvt_1* são do aeroporto SBVT depois de 1 hora dos atrasos de SBCF, melhor ilustrado na Tabela 6. Assim, é possível analisar a propagação dos atrasos na geração das regras de associação.

Tabela 6: Janela deslizante com percentual de atrasos do aeroporto SBCF

dia_hora	sbcf_0	sbcf_1	sbvt_1
2009-01-01 05:00:00	0.0000	0.8750	0.5000
2009-01-01 06:00:00	0.8750	0.6000	0.0000
2009-01-01 07:00:00	0.6000	0.0000	0.0000
2009-01-01 08:00:00	0.0000	0.5555	0.0000

Além disso, analisamos as rotas entre os aeroportos e definimos a propagação do atraso em até no máximo 4 horas. Ou seja, para cada aeroporto destino da janela, pode-se considerar um intervalo de 1 a 4 horas depois do atraso na partida. Esse limite foi definido porque a maior rota do banco de dados da ANAC encontrada leva em média 4 horas para ser realizada, SBGL $\rightarrow$ SBEG com duração de 04:06:16 em média. A existência de rotas foi analisada, pois nem todos os aeroportos apresentados na Tabela 4 possuem rotas entre si e, além disso, foi importante considerar o tempo de viagem entre eles porque, por exemplo, em uma rota que tem 2 horas de duração, não faria sentido analisarmos a propagação de atraso dessa rota 1 hora depois da saída do avião, já que nem teve tempo do mesmo chegar no aeroporto de destino. Logo, esses dados na composição da janela deslizante seriam enganosos e exigiria um filtro no fim da análise. Por conta disso, levamos em consideração essas informações e não incluímos na nossa janela aeroportos que não possuem rotas entre si e nem a propagação de atraso menor do que o tempo de duração da rota.

Portanto, a execução da janela deslizante primeiramente executa a função *rotas(ids, arpt)*, a qual recebe como parâmetro a tabela agregada *ids* e o aeroporto que está sendo analisado no momento. Essa função retorna os aeroportos de destino e o tempo médio de duração de viagem

em relação ao aeroporto que está sendo analisado. Em outras palavras, retorna as rotas com os seus tempos de duração. A partir das rotas e da tabela agregada, pode-se criar a janela com os percentuais de atrasos obedecendo os critérios apresentados nos parágrafos anteriores desta seção. No caso, essa função corresponde a *percAtraso(ids, r, arpt)* do Algoritmo 1.

Por fim, a técnica *binning* é aplicada na janela deslizante para classificar as porcentagens em *low, medium e high* (baixa, média e alta). Dessa forma, mais padrões podem ser encontrados, já que estamos transformando os dados contínuos da porcentagem em intervalos. Para encontrar os valores para cada intervalo, primeiro aplicamos a curvatura máxima para saber qual seria o número ideal de intervalos para os dados que estamos trabalhando. O resultado é 4, conforme Figura 6. Com quatro intervalos entre 0 a 1 (0% a 100%), temos os seguintes valores: 0.00, 0.25, 0.50, 0.75 e 1.00. A partir desses quatro intervalos queremos formar três. No caso, consideramos metade ou mais de atrasos uma taxa alta. Ou seja, 50% ou mais da quantidade de atrasos corresponde à taxa *high*. Logo, definimos os intervalos conforme a Tabela 7.

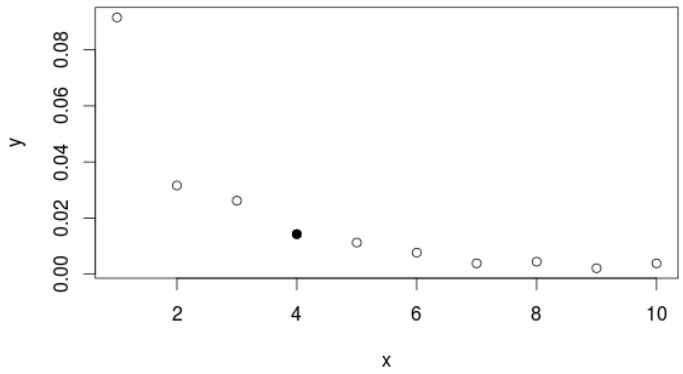

Figura 6: Curvatura máxima para a análise do número de intervalos *bin*

Tabela 7: Intervalos definidos a partir do *binning*

Intensidade do Atraso	Porcentagem Correspondente
low	0% até < 25%
medium	$\geq$ 25% até < 50%
high	$\geq$ 50% até 100%

Portanto, substituindo as porcentagens pelos intervalos definidos, a janela resultante da função *materializacaoJanelaDeslizante(ids, arpt)* ficou conforme ilustrada na Tabela 8. O atributo *dia_hora* foi retirado pois não estamos interessados no horário em que o atraso ocorreu especificamente, e sim nos aeroportos impactados e em quanto tempo depois esse atraso se propaga. As

janelas, resultante dos 17 aeroportos, foram gravadas em disco, ou seja, foram materializadas para aplicar o algoritmo Apriori em seguida.

Tabela 8: Resultado da janela deslizante para o Aeroporto SBCF

sbcf_0	sbcf_1	sbvt_1
low	high	high
high	high	low
high	low	low
low	high	low

4.5 Geração das Regras

Todo o processo de geração das regras corresponde à função *geracaoRegras(ids)* do Algoritmo 1. Primeiramente, invoca-se a função *parametrosHiper(ids)*, que é responsável pelo processo de encontrar os valores para os parâmetros do algoritmo Apriori, como suporte e confiança. Após a definição desses parâmetros, a função *apriori(ids, arp)* é executada, na qual as regras de associação são geradas. Por fim, uma filtragem em cima dessas regras é realizada na função *filtrarRegras(ar)*. Detalhes sobre o processo realizado em cada função são apresentados nos próximos parágrafos dessa seção.

O algoritmo Apriori foi executado para gerar as regras através do pacote Arules da linguagem R [R Core Team, 2014]. Aplicamos esse algoritmo para as 17 janelas deslizantes, similares à apresentada na Tabela 8. Alguns parâmetros são exigidos para que a execução desse algoritmo seja possível, como o suporte, a confiança e os tamanhos das regras. Além disso, também deve-se definir os valores para os antecessores e consequentes das regras.

A confiança foi calculada a partir da aplicação da técnica *binning* na classificação dos atrasos em *low, medium e high*, descrita na subseção anterior. Definimos a confiança como sendo o menor intervalo resultante do *binning*. Os intervalos foram definidos conforme Tabela 7. Logo, o tamanho do menor intervalo refere-se a 25%. Definir a confiança nesse valor permite englobar todas as aparições de qualquer atraso, seja ele, *low, medium* ou *high*.

Já o suporte, foi calculado atráves da curvatura máxima em relação a quantidade de regras com o suporte variando de 1% a 10%. Para calcular a quantidade das regras, as mesmas foram limitadas a um lift maior do que 1, já que são as regras que nos interessam. O valor encontrado aplicando a curvatura máxima foi de 7%, como mostrado na Figura 1.

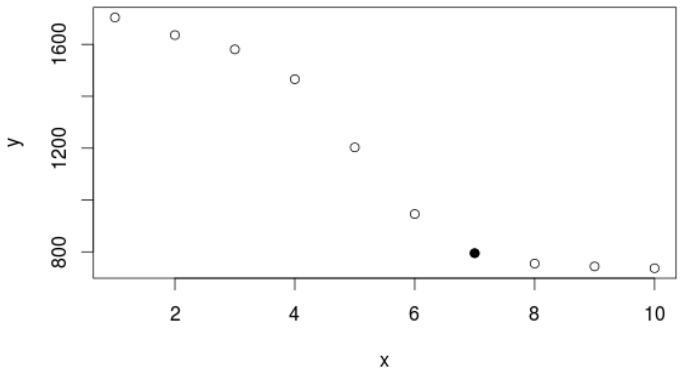

Figura 7: Curvatura máxima para a análise do valor do suporte

Outro parâmetro definido para a execução do Apriori é o tamanho das regras. O tamanho mínimo das regras deve representar o número mínimo de atributos necessários para obter regras significativas. Se não definir nenhum valor para o tamanho mínimo, regras do tipo $\{\} \rightarrow$ B são geradas. No caso do nosso trabalho, o tamanho mínimo é igual a 2, tendo que as regras devem ter pelo menos um atributo do lado esquerdo (LHS), que corresponde ao aeroporto que origina o atraso, e um do lado direito (RHS), que representa o aeroporto impactado pela propagação do atraso. Em outras palavras, sem qualquer atributo no LHS, não se pode obter condições que levam a uma propagação de atraso de voo.

O tamanho máximo também foi definido em 2, porque estamos interessadas em analisar a propagação dos atrasos entre aeroportos diretamente. Queremos encontrar regras em que podemos explicar que um atraso no aeroporto A impacta em um aeroporto B tantas horas depois. Logo, o tamanho máximo definido em 2 é o suficiente para explicarmos essa propagação entre aeroportos.

Por fim, falta definir os valores para os antecessores (LHS) e consequentes (RHS) das regas. Para o antecessor, restringimos a aparição das frequências de atrasos em *high*. Já que o nosso foco está na propagação de atrasos com impactos significativos, atributos equivalentes a *low* e *medium* no lado LHS das regras foram ignorados. Para os consequentes, deixamos o valor *default* do Apriori, já que seria mais complexo restringir. Essa configuração foi suficiente para gerar regras do nosso interesse.

Com isso, regras do tipo $\{sbgr\_0{=}high\} \rightarrow \{sbgl\_2{=}high\}$ são geradas. O que significa que quando a taxa de atraso em Guarulhos (SBGR) é alta, o aeroporto Galeão (SBGL) tem um reflexo com taxa alta de atraso 2 horas depois. Ou seja, uma alta taxa de atraso em SBGR se

propaga em SBGL duas horas depois.

Para finalizar, no último passo do Algoritmo 1, desenvolvemos uma filtragem das regras para a nossa análise. Essa filtragem consiste em considerar apenas as regras com o lift maior do que 1. Além disso, também realizamos uma filtragem no RHS das regras para selecionar apenas as que possuem taxas *high* de atraso, que são as taxas de nosso interesse. Isso foi necessário porque as taxas *low* e *medium* aparacem nas regras, pois não restrigimos nenhum valor para o RHS na execução do Apriori. Logo depois, salvamos todas essas regras filtradas em disco para poder realizar a análise apresentada a seguir.

Capítulo 5

Avaliação Experimental

Nossa análise foi feita em cima das regras de associação *iar* do Algoritmo 1. Ou seja, em cima dos resultados obtidos após a geração das regras. Com essa análise, conseguimos responder as perguntas anunciadas na introdução, a serem discutidas nas próximas seções.

Vale ressaltar que além de apresentar uma análise geral, também efetuamos uma análise detalhada para cada pergunta em relação aos aeroportos Guarulhos (SBGR), Galeão (SBGL) e Presidente Juscelino Kubitschek (SBBR) nas cidades de São Paulo, Rio de Janeiro e Brasília, respectivamente. O aeroporto de Brasília foi escolhido para essa análise detalhada porque sua localização no Distrito Ferederal, em Brasília, é bem estratégica, tornando-o um dos principais *hubs* sul-americanos, transportando mais de 19,8 milhões de passageiros anualmente [Aeroporto de Brasília, 2016]. O de Guarulhos foi selecionado por ser o maior complexo aeroportuário do país [Aeroporto Internacional de São Paulo, 2015]). Ou seja, é o principal aeroporto do Brasil. Já o Galeão foi selecionado pois é o aeroporto mais movimentado do Rio de Janeiro, um dos principais estados do Brasil economica e politicamente. Além disso, a cidade do Rio de Janeiro é uma das que mais recebem turistas normalmente e no ano de 2014, excepcionalmete, por conta da ocorrência de um grande evento como a Copa do Mundo de Futebol.

5.1 Os aeroportos estão interligados quando um atraso ocorre?

Os resultados das regras mostram que os atrasos ocorridos em um determinado aeroporto se propagam sim para outros aeroportos da malha brasileira. Obtivemos 164 regras mostrando a propagação em diversos aeroportos brasileiros. No total, 15 dos 17 aeroportos provocam a propagação de atrasos no Brasil.

O Brasil é divido em 5 regiões: Norte, Nordeste, Centro-Oeste, Sudeste e Sul. Quase todas essas regiões possuem pelo menos um aeroporto que provoca a propagação em outras regiões. Apenas a região Norte do país não possui nenhum aeroporto influente nessa análise pois nenhuma regra foi produzida para os aeroportos dessa região. Para facilitar o estudo do impacto dos atrasos, dividimos os resultados conforme as regiões brasileiras.

Os resultados são mostrados em tabelas dividas por região, como listadas a seguir: Tabela 9, Tabela 10, Tabela 11 e Tabela 12. Nas tabelas, mostramos os 5 principais aeroportos impactados por essa região. Esses aeroportos estão listados em ordem descrescente em relação ao *lift*. Desta maneira, pode-se visualizar melhor o aeroporto mais impactado para o menos impactado. Essas tabelas também informam a quantidade de regras que esse aeroporto gerou, que corresponde ao número de propagações ocorridas, e a quantidade de aeroportos impactados. A diferença entre a quantidade de regras e o número total de aeroportos pode ser explicada pelo fato de que podemos ter regras do tipo $\{sbgr\_0{=}high\} \rightarrow \{sbgl\_1{=}high\}$ e $\{sbgr\_0{=}high\} \rightarrow \{sbgl\_2{=}high\}$, que mostram a propagação do atraso de SBGR em SBGL em 1 e 2 horas depois. Portanto, há 2 propagações, porém, em apenas um aeroporto, SBGL. Além disso, cada uma dessas regras possui um *lift*. Logo, quando acontece esse cenário descrito anteriormente, o *lift* apresentado na tabela é a média dos *lifts* do aeroporto impactado.

Vamos começar com a análise da região Sudeste. Essa região possui 7 aeroportos dos 17 que estamos considerando. Pelos resultados, mostrou ser a região que mais influencia na propagação de atrasos na malha brasileira porque, além de observarmos que essa região influencia atrasos em todo o restante do Brasil, possui 82 regras, o que representa metade do total das regras geradas. Além disso, apenas os aeroportos Guarulhos (SBGR) e Galeão (SBGL), juntos, possuem 46 regras, as quais correspondem a quase 30% do total. Portanto, só esses dois aeroportos do Sudeste têm um papel impactante no restante da rede aérea. Esses números podem ser observados na Tabela 9.

Apenas analisando Guraulhos (SBGR), os resultados mostraram que quando esse aeroporto está com a taxa de atrasos alta, impacta 9 aeroportos diferentes no pais. Os cinco que sofrem mais influencia de SBGR são: {SBPA, SBCF, SBBR, SBSV, SBCT}, ordenados respectivamente pelo maior *lift* até o menor. Os quatro restantes são: {SBGL, SBVT, SBRF, SBFZ}, também em ordem descrencente do *lift*, variando entre 1.3950 e 1.4344. Logo, apenas o aeroporto SBGR consegue impactar quase todas as regiões do Brasil, exceto a região Norte. Na própria região Sudeste já se pode observar reflexos dos atrasos de Guarulhos, pelos aeroportos SBCF, SBGL e SBVT. No sul, os aeroportos SBPA e SBCT sofrem reflexo. No nordeste, SBSV, SBRF e SBFZ. Por fim, na região Centro-Oeste, o grande aeroporto de Brasília, SBBR, também sofre influências dos atrasos de SBGR.

Outro aeroporto bastante importante na propagação dos atrasos é o SBGL. Esse aeroporto, junto com SBGR, mostrou ser o que mais influencia outros aeroportos na malha aérea brasileira,

Tabela 9: Análise geral da propagação originada na região Sudeste do Brasil

Aeroporto Analisado	Regras Geradas	Número de Aeroportos Impactados	Aeroportos Impactados e seus Lifts
SBGR	24	9	SBPA: 1.5893 SBCF: 1.5511 SBBR: 1.5076 SBSV: 1.4795 SBCT: 1.4711
SBGL	22	9	SBCF: 1.4238 SBBR: 1.4222 SBVT: 1.4136 SBGR: 1.3861 SBSV: 1.3686
SBVT	14	6	SBCF: 1.5577 SBSV: 1.4486 SBBR: 1.4309 SBGL: 1.4235 SBCT: 1.3696
SBCF	10	5	SBBR: 1.5719 SBGL: 1.4546 SBGR: 1.4527 SBSV: 1.3952 SBCT: 1.3873
SBKP	7	3	SBCF: 1.5723 SBGL: 1.3718 SBCT: 1.3680
SBSP	4	2	SBCF: 1.9747 SBGL: 1.7950
SBRJ	1	1	SBGR: 1.6351

impactando 9 aeroportos no total. Os cinco com maior *lift* são: {SBCF, SBBR, SBVT, SBGR, SBSV}. Os restantes, também em ordem decrescente do *lift*, são: {SBKP, SBCT, SBFZ, SBRF, SBSV}, com o *lift* variando entre 1.3074 até 1.3660. Os aeroportos em que SBGL propaga seu atraso, em sua maioria, são os mesmos que SBGR propaga. Esse compartamento é explicado pelo fato de que os dois aeroportos estão localizados próximos um do outro e estão entre os cinco principais aeroportos do Brasil. Logo, podemos concluir que quando a taxa de atraso está alta em SBGL, vários aeroportos de diferentes regiões do Brasil sentem o reflexo dessa condição.

Por fim, os aeroportos restantes da região Sudeste somam em 36 regras geradas e influenciam, juntos, 17 aeroportos brasileiros. As informações individuais para cada aeroporto pode ser vista na Tabela 9. Podemos concluir, a partir dessa tabela, que a concentração da propagação dos atrasos causados pela região Sudeste está nos aeroportos SBGR e SBGL.

Já na região Centro-Oeste, podemos deduzir claramente que a propagação está concentrada no aeroporto de Brasília, SBBR, conforme Tabela 10. Dos 6 aeroportos impactados por essa região, SBBR impacta 5, com um total de 13 propagações. Os cinco principais aeroportos que sofrem impacto do atraso de SBBR são: {SBCF, SBGR, SBSV, SBCT, SBGL}. Grandes aeroportos, como SBGR e SBGL, sentem a propagação gerada pelo aeroporto de SBBR. Além disso, SBBR impacta todas as demais regiões do Brasil (exceto a região Norte): Sul (SBCT), Nordeste (SBSV) e Sudeste (SBCF, SBGR e SBGL). Disso, podemos concluir que sua importância no cenário da propagação de atrasos na malha brasileira é bastante significativa.

O aeroporto SBGO tem um impacto muito pequeno na quantidade de aeroportos na mallha brasileira. Apenas influencia o aeroporto de Guarulhos (SBGR) em dois tempos de propagação. Sendo assim, comparado com os aeroportos já analisados como SBGR, SBGL e SBBR, esse aeroporto não tem um papel significativo em relação a quantidade de aeroportos em que seu atraso pode propagar.

Tabela 10: Análise geral da propagação originada na região Centro-Oeste do Brasil

Aeroporto Analisado	Regras Geradas	Número de Aeroportos Impactados	Aeroportos Impactados e seus Lifts
SBBR	13	5	SBCF: 1.5469 SBGR: 1.4506 SBSV: 1.3840 SBCT: 1.3745 SBGL: 1.3688
SBGO	2	1	SBGR: 1.4824

Tendo como foco a região Sul brasileira, conforme Tabela 11, observamos como destaque o aeroporto de Curitiba, SBCT. Possui um total de 15 regras geradas e influencia 7 aeroportos diferentes, correspondendo a 58% dos aeroportos impactados pela região. Entre esses aeroportos, os cinco mais impactados são: {SBPA, SBCF, SBBR, SBGR, SBKP}, montrando uma forte influência na região Sudeste pois dos cinco aeroportos apresentados, três pertencem a essa região. Os dois aeroportos que não estão na tabela são {SBGL, SBSV}, com os *lifts* correspondentes de 1.3616 e 1.3164, respectivamente. Portanto, SBCT propaga seu atraso principalmente nos aeroportos da região Sudeste. Contudo, também propaga para demais regiões, exceto Norte.

Os demais aeroportos do sul, SBPA e SBFl, juntos, impactam menos aeroportos do que SBCT, totalizando em 5 aeroportos contra 7 de SBCT. Desses 5 aeroportos, 2 deles, SBGR e SBGL, sofrem influência de ambos aeroportos, mostrando, novamente, uma grande influência nos atrasos ocorridos no Sul do país na região Sudeste.

Tabela 11: Análise geral da propagação originada na região Sul do Brasil

Aeroporto Analisado	Regras Geradas	Número de Aeroportos Impactados	Aeroportos Impactados e seus Lifts
SBCT	15	7	SBPA: 1.6395 SBCF: 1.5213 SBBR: 1.4822 SBGR: 1.4509 SBKP: 1.3963
SBPA	9	3	SBCT: 1.6219 SBGR: 1.5827 SBGL: 1.5381
SBFL	5	2	SBGR: 1.4060 SBGL: 1.2981

Por fim, a análise da região Nordeste, que possui 3 aeroportos dos 17 que consideramos para essa pesquisa. O principal aeroporto dessa região é SBSV, conforme Tabela 12, que impacta SBCF, SBBR, SBRF, SBGR e SBGL. Além desses cinco aeroportos, SBSV também impacta SBCT com *lift* em 1.3518. Logo, podemos concluir que SBSV tem maior influencia nos aeroportos da região Sudeste, mas também impacta outros aeroportos importantes como SBCT e SBBR. Portanto, a análise dos aeroportos da região Nordeste mostraram que SBSV é o aeroporto mais influente dessa região, já que metade dos aeroportos que são influenciados pelo Nordeste tem como influenciador o SBSV.

Tabela 12: Análise geral da propagação originada na região Nordeste do Brasil

Aeroporto Analisado	Regras Geradas	Número de Aeroportos Impactados	Aeroportos Impactados e seus Lifts
SBSV	11	6	SBCF: 1.5101 SBBR: 1.4714 SBRF: 1.4384 SBGR: 1.4064 SBGL: 1.3972
SBFZ	8	4	SBSV: 1.3451 SBRF: 1.2848 SBGL: 1.2384 SBGR: 1.2016
SBRF	4	2	SBGL: 1.2192 SBGR: 1.2040

Com isso, a análise realizada nessa seção mostrou que os aeroportos brasileiros estão interligados quando um atraso ocorre em um aeroporto da rede. Foi possível responder a pergunta tema da seção apresentando detalhadamente as relações dos aeroportos de cada região com os

demais. A partir dessa análise, podemos identificar os 3 principais aeroportos em relação a quantidade de aeroportos diferentes influenciados e também em relação a quantidade de propagações geradas, conforme indica a Tabela 13.

Tabela 13: Aeroportos mais influentes na propagação de atrasos na malha aérea brasileira

Aeroporto	Aeroportos Impactados	Regras Geradas
SBGR	9	24
SBGL	9	22
SBCT	7	15

5.2 Por quanto tempo um atraso se propaga entre os aeroportos?

Como foi explicado na Seção 4.4, a janela deslizante materializada nesse trabalho se extende em até 4 horas depois de um atraso ocorrido. Ou seja, na análise da pergunta dessa subseção, os atrasos podem se propagar em até, no máximo, 4 horas após um atraso ocorrido na rede.

Assim como na seção anterior, montamos tabelas para cada região do Brasil com as informações de cada aeroporto pertencente a mesma. Assim, o detalhamento dos resultados decorre de uma maneira mais organizada. Uma das informações mostrada nessas tabelas é o aeroporto que está sendo analisado, ou seja, o aeroporto no qual o atraso tem sua origem. Além dessa informação, é mostrada a quantidade de propagações que ocorreram em relação às horas corridas após o atraso.

A região Sudeste possui aeroportos que impactam outros em várias horas depois, como os grandes aeroportos SBGR e SBGL, e tem outros que impactam somente em uma hora, como SBRJ. Essa diferença se dá pelo fato da movimentação dos aeroportos e na influência dos mesmos na propagação de atrasos na malha aérea brasileira, discutida na seção anterior. Os dados do tempo de propapagação dessa região podem ser visualizados na Tabela 14.

Quando a taxa de atrasos em SBGR está alta, a propagação desse cenário é sentida principalmente depois de 2 e 3 horas nos outros aeroportos. Em 2 horas depois, 7 aeroportos diferentes são influenciados pela alta taxa de atrasos em SBGR. O mesmo acontece 3 horas depois. Já em SBGL, a concentração está entre 1 e 2 horas depois dos atrasos ocorridos. Na primeira hora, 6 aeroportos são impactados pelos atrasos do aeroporto do Rio de Janeiro, Galeão (SBGL). Na segunda hora, 8 são impactados, o que registra a maior quantidade de aeroportos impactados num intervalo de tempo na região Sudeste.

Tabela 14: Número de aeroportos impactados a cada hora corrida da propagação, originada no Sudeste do Brasil

Aeroporto	1 Hora	2 Horas	3 Horas	4 Horas
SBGR	5	7	7	5
SBGL	6	8	5	3
SBVT	6	4	2	2
SBCF	5	2	2	1
SBKP	3	2	1	1
SBSP	2	1	1	0
SBRJ	0	1	0	0

Os aeroportos que SBGR e SBGL impactam em cada intervalo de tempo podem ser observados na Figura 8. Essa figura mostra diferentes tempos referentes à propagação de atrasos em cada um dos dois aeroportos. Aeroportos como SBVT sentem o reflexo do atraso de SBGL e SBGR entre 1 hora a 2 horas depois. Já SBCF extende até 3 horas. Outros aeroportos como SBFZ e SBRF só sentem a propagação de SBGR entre 3 a 4 horas depois. Já em SBGL, esses mesmos aeroportos sentem o atraso a partir de 2 horas depois. Em ambos aerportos, podemos perceber uma maior influência nas regiões Sudeste e Nordeste do país.

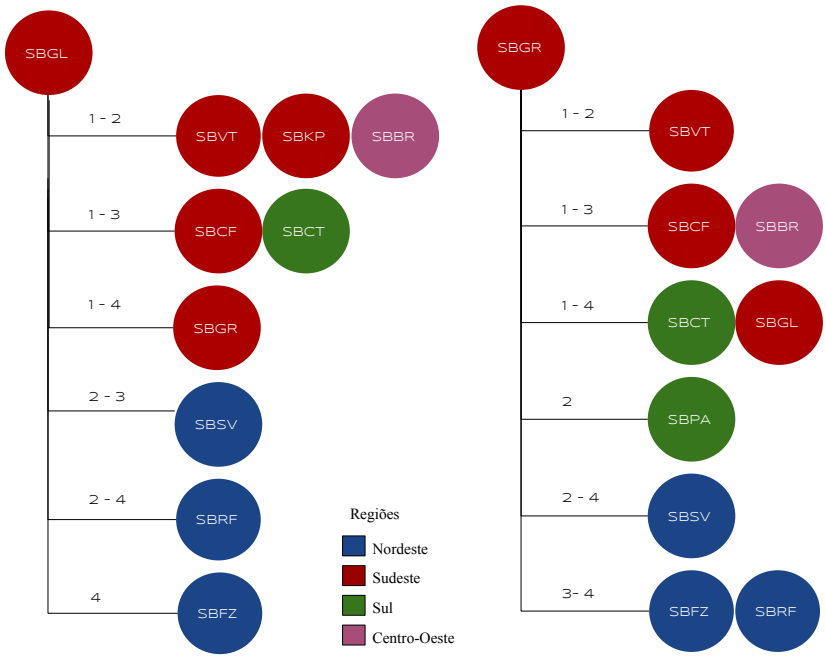

Figura 8: Propagação de atrasos originada pelos aeroportos SBGL e SBGR. Os números indicam quanto tempo depois o atraso é propagado e quanto tempo dura. Exemplo: 1-4, significa que o atraso se propaga 1 hora depois e se extende em até 4 horas depois do atraso ocorrido

A Tabela 15 mostra detalhes do tempo da propagação na região Centro-Oeste brasileira. Nessa região, o principal aeroporto é o de Brasília (SBBR), como já vimos em seções anteriores.

Podemos perceber, pela tabela, que SBBR influencia vários aeroportos em diversos momentos após seus atrasos ocorridos. Como SBBR impacta 5 aeroportos diferentes, mostrados na seção anterior, todos os aeroportos que ele impacta sentem a propagação 1 hora depois do atraso ocorrido. A segunda hora apresenta 4 aeroportos afetados. Já na terceira e na quarta hora, 2 aeroportos são afetados.

Tabela 15: Número de aeroportos impactados a cada hora corrida da propagação, originada no Centro-Oeste do Brasil

Aeroporto	1 Hora	2 Horas	3 Horas	4 Horas
SBBR	5	4	2	2
SBGO	1	1	0	0

Como dito no parágrafo anterior, todos os aeroportos impactados por SBBR, ilustrados na Figura 9, sentem o reflexo do atraso de SBBR na primeira hora decorrida. Essa propagação pode se extender em duas ou em até quatro horas depois do atraso iniciado em SBBR. SBCF e SBCT sentem o reflexo do atraso de SBBR 1 hora depois, podendo permanecer por mais 1 hora. Ou seja, pode-se sentir o reflexo nesses dois aeroportos 2 horas depois do atraso ocorrido em SBBR. Já os dois grandes aeroportos SBGL e SBGR são impactados pela propagação em 1 hora a 4 horas depois do atraso inicial em SBBR.

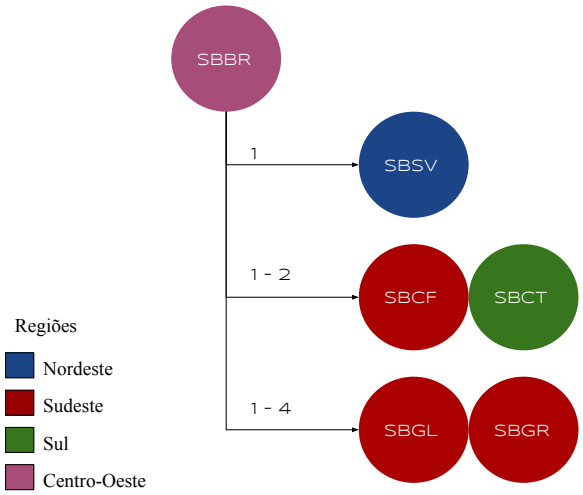

Figura 9: Propagação de atrasos originada pelo aeroporto SBBR

Na região Sul, podemos ver claramente que SBCT já tem um grande impacto na primeira hora corrida depois de seus atrasos, totalizando em 6 aeroportos diferentes, conforme Tabela 16. Na segunda hora, consegue influenciar 4 aeroportos diferentes e esse número descresce de uma unidade até a quarta hora corrida. Já o aeroporto SBPA, na primeira e segunda hora impacta

3 aeroportos. Na terceira hora, 2, e na primeira apenas 1 aeroporto sofre influencia de SBPA. Essa análise mostra novamente que SBCT tem maior influência na região Sul, pois em todas as horas após os atrasos ocorridos, SBCT influencia mais aeroportos diferentes do que os demais da região.

Tabela 16: Número de aeroportos impactados a cada hora corrida da propagação, originada no Sul do Brasil

Aeroporto	1 Hora	2 Horas	3 Horas	4 Horas
SBCT	6	4	3	2
SBPA	3	3	2	1
SBFL	1	2	1	1

Por fim, a região Nordeste mostra uma influência significante com o aeroporto SBSV na segunda hora depois dos atrasos terem acontecido. Nesse momento, SBSV impacta 5 aeroportos diferentes. Nas restantes horas, impacta somente 2. Podemos observar que após 3 horas dos atrasos, SBFZ tem maior influência do que SBSV, pois impacta 3 aeroportos contra 2. Contudo, na segunda hora, SBSV mostra ser cinco vezes mais influente do que SBFZ.

Tabela 17: Número de aeroportos impactados a cada hora corrida da propagação originada no Nordeste do Brasil

Aeroporto	1 Hora	2 Horas	3 Horas	4 Horas
SBSV	2	5	2	2
SBFZ	2	1	3	2
SBRF	0	0	2	2

Assim, podemos classificar os aeroportos mais influentes a cada hora transcorrida após os atrasos, conforme Tabela 18. Percebe-se uma grande influência da região Sudeste, pois impacta aeroportos em todos os intervalos de tempo. Na primeira hora da Tabela 18, os três aeroportos influenciam 6 outros. Na segunda, SBGL impacta 8 aeroportos diferentes. Na terceira hora, SBGR influencia 7 aeroportos. E na quarta hora, SBGR impacta 5 aeroportos diferentes. Portanto, em vista a região Sudeste, SBGL mostra ser mais impactante nas primeiras 2 horas e SBGR nas últimas 2 horas.

Tabela 18: Aeroportos mais influentes na rede em relação ao tempo transcorrido após o atraso

1 Hora	2 Horas	3 Horas	4 Horas
SBGL SBCT SBVT	SBGL	SBGR	SBGR

5.3 Como é o comportamento da propagação dentro do próprio aeroporto?

Há regras geradas do tipo *{sbgr_0=high}* $\rightarrow$ *{sbgr_1=high}* que torna possível estudar a propagação do atraso dentro do próprio aeroporto, no caso do exemplo, a autopropagação em SBGR. Porém, alguns aeroportos mostraram não propagar os atrasos dentro dos mesmos, como SBGO e SBFZ, conforme Tabela 19. Essa tabela apresenta a duração da autopropagação para cada aeroporto, exceto os da região Norte porque não tiveram nenhum resultado.

Tabela 19: Duração da propagação dentro dos aeroportos. As divisórias indicam a qual região os aeroportos pertencem, sendo Sudeste, Centro-Oeste, Sul e Nordeste em sequência

Aeroporto	Duração (em horas)
SBGR	1 a 4
SBGL	1 a 4
SBVT	1 a 2
SBCF	1 a 2
SBKP	1 a 2
SBSP	1 a 2
SBRJ	1 a 2
SBBR	1 a 2
SBGO	-
SBCT	1 a 3
SBPA	-
SBFL	-
SBSV	1 a 3
SBFZ	-
SBRF	1 a 2

A duração da propagação em SBGR mostra ser de 1 a 4 horas depois dos atrasos ocorridos. Podemos interpretar essa duração de SBGR como sendo a dificuldade desse aeroporto se recuperar de atrasos passados. Tanto ele quanto SBGL extendem suas propagações em até 4 horas. Ou seja, demoram mais de 4 horas para se recuperarem. Como nossa janela deslizante é limitada em 4 horas, não é possível ter uma ideia melhor de quanto tempo esses dois aeroportos costumam conseguir diminuir sua taxa de atraso.

Outros aeroportos como SBBR, SBVT, SBCF, SBKP, SBSP, SBRJ e SBRF, ou seja, a maioria dos 17 aeroportos considerados nessa pesquisa, permanecem com a taxa alta de atrasos depois de 1 a 2 horas de terem atingido a mesma. Portanto, mostram conseguir se recuperar da taxa de atrasos *high* depois de 2 horas que essa taxa se estabeleceu. Por fim, os únicos aeroportos apontando uma duração de 1 a 3 horas para se recuperar de seus atrasos são SBCT e

SBSV.

Para entender um pouco melhor o motivo do tempo de duração que os aeroportos levam para se recuperar de outros atrasos, analisamos também a quantidade de aeroportos que impactam um determinado aeroporto. Em outras palavras, uma tabela com uma visão inversa das informações apresentadas anteriormente, em relação a quantidade de aeroportos que um aeroporto impacta, representadas entre as Tabelas 14 e 17. Por exemplo, um aeroporto A impacta 10 outros aeroportos, porém, 7 aeroportos propagam seus atrasos para esse aeroporto A. Com isso, entre as Tabelas 20 e 23, que serão apresentadas adiante, estamos analisando o último caso. Essas tabelas apresentam o aeroporto A que está sendo analisado e a quantidade de aeroportos Bs que o impacta. Essa quantidade de aeroportos impactantes está dividida em relação às horas transcorridas após o aeroporto A apresentar uma taxa *high* de atrasos.

Os resultados da Tabela 20 mostram que a quantidade de aeroportos que impactam SBGR e SBGL nos 4 intervalos de tempo é bastante alta em comparação aos outros aeroportos. Logo, esses dois aeroportos demoram mais para se recuperarem de atrasos passados, pois a quantidade de atrasos propagados neles é grande e constante nas 4 horas da janela que estabelecemos. Isso explica uma duração da autopropagação de 1 a 4 horas, maior do que qualquer outro aeroporto da malha brasileira. Logo, é bem provável que SBGR e SBGL acumulem bastante atrasos de outros aeroportos e deles mesmos.

Tabela 20: Número de aeroportos que impactam, a cada hora corrida de seus atrasos, os aeroportos do Sudeste do Brasil

Aeroporto	1 Hora	2 Horas	3 Horas	4 Horas
SBGR	8	10	10	9
SBGL	8	10	11	8
SBVT	2	2	0	0
SBCF	8	6	2	0
SBKP	2	1	0	0
SBSP	0	0	0	0
SBRJ	0	0	0	0

Nos outros aeroportos da região Sudeste, SBVT, SBCF e SBKP, podemos observar uma maior concentração de aeroportos impactando neles entre 1 a 2 horas depois dos atrasos. Essa concentração faz com que esses aeroportos tenham dificuldades de se recuperarem dos atrasos nesses intervalos, mostrando então, uma duração da autopropagação de 1 a 2 horas. Os aeroportos SBSP e SBRJ não possuem nenhum aeroporto impactando eles significamente. Porém, na Tabela 19 temos a informação de que a duração da autopropagação deles é de 1 a 2 horas.

Isso pode ser explicado através dos atrasos internos, sem interferência dos atrasos de aeroportos externos. Apenas os atrasos dentro desses aeroportos já bastam para criar uma dificuldade deles se recuperarem da taxa alta de atrasos.

Aprofundamos a análise para visualizar quais aeroportos impactam em SBGR e SBGL detalhadamente, Figuras 10 e 11, respectivamente. Podemos ver que ambos são impactados por 12 aeroportos, a maioria sendo os mesmos. Novamente, essa situação pode ser explicada por conta desses dois aeroportos estarem próximos geograficamente e por terem quase a mesma importância na escala nacional. Portanto, com essa quantidade de aeroportos provenientes de todas as regiões (exceto a região norte) influenciando os atrasos em SBGR e SBGL entre 1 e 4 horas, esses dois aeroportos ficam com mais dificuldade de se recuperarem de atrasos passados pois estão sofrendo reflexos ao mesmo tempo desses aeroportos. Logo, o tempo para eles se recuperarem é longo e ultrapassa as 4 horas corridas, já que nossa janela foi fixada em até esse período.

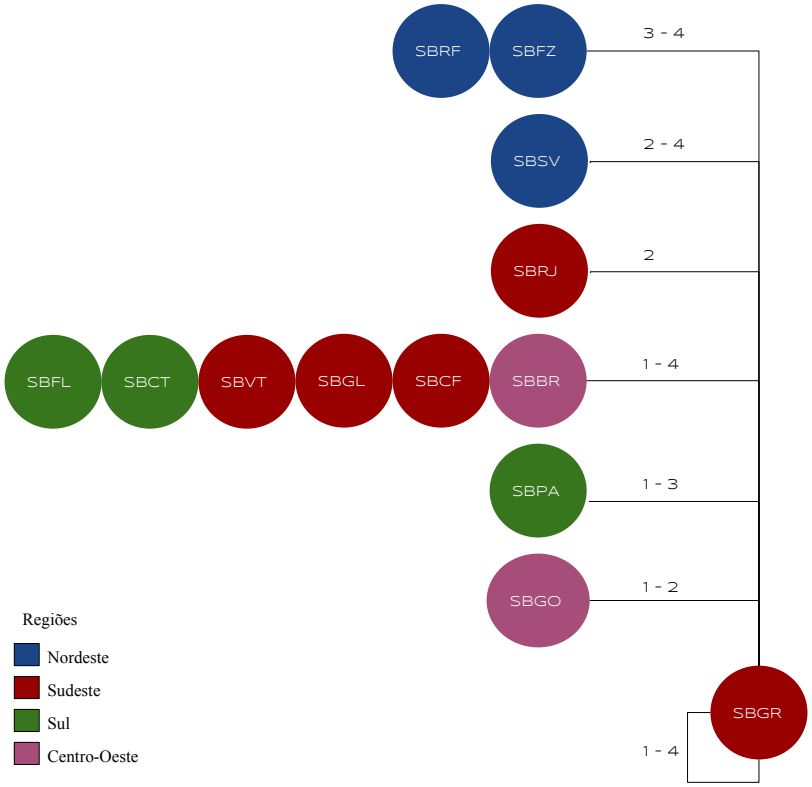

Figura 10: Propagação de atrasos que impactam o aeroporto SBGR

Na região Centro-Oeste, Tabela 21, SBBR apresenta uma maior concentração de atrasos na primeira hora, totalizando em 6 aeroportos impactando SBBR nesse momento. Contudo, na Tabela 19, SBBR apresenta uma duração de 1 a 2 horas. Podemos entender essa duração por

conta da grande influência de atrasos dos outros aeroportos na primeira hora e por conta disso, dificulta a recuperação de SBBR até a segunda hora, que, além disso, também tem 2 aeroportos propagando atrasos nesse momento. Já SBGO não apresenta autopropagação na Tabela 19 e nem aeroportos influentes nos atrasos na Tabela 21.

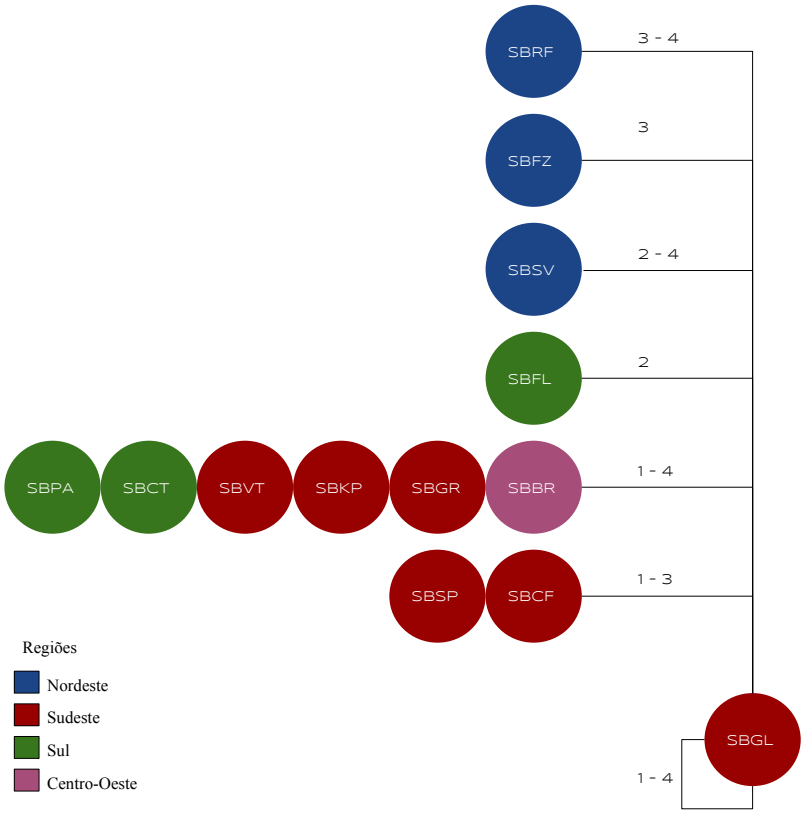

Figura 11: Propagação de atrasos que impactam o aeroporto SBGL

Tabela 21: Número de aeroportos que impactam, a cada hora corrida de seus atrasos, os aeroportos do Centro-Oeste do Brasil

Aeroporto	1 Hora	2 Horas	3 Horas	4 Horas
SBBR	6	2	1	0
SBGO	0	0	0	0

Também realizamos um estudo mais detalhado dos aeroportos que podem estar influenciando SBBR e assim, contribuindo para a permanência de sua autopropagação. A Figura 12 mostra que a maioria dos aeroportos que prejudicam SBBR pertencem à região Sudeste, são eles: SBGR, SBGL, SBCF e SBVT. Esses aeroportos impactam SBBR em tempos diferentes. Como SBBR demora entre 1 a 2 horas para se recuperar dos atrasos, todos os aeroportos apresentados na Figura 12 englobam esse período. Portanto, todos eles intensificam a autopropagação de SBBR.

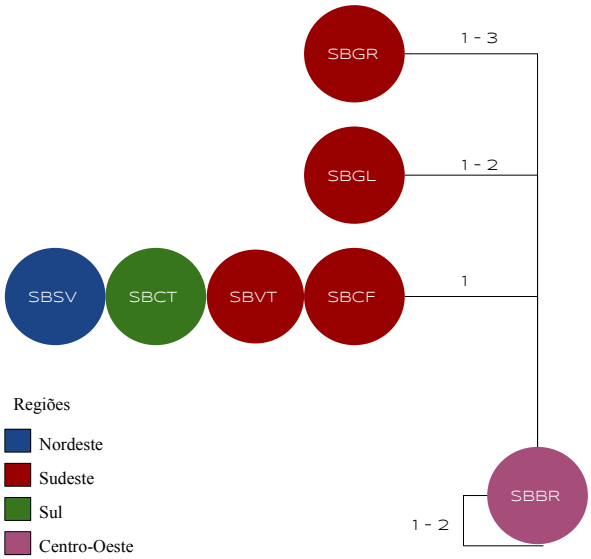

Figura 12: Propagação de atrasos que impactam o aeroporto SBBR

Na região Sul, Tabela 22, SBCT mostra ser o mais impactado, a concentração de aeroportos que o influência está presente nas duas primeiras horas, correspondendo a 7 aeroportos em cada hora. Pela Tabela 19, a autopropagação em SBCT é de 1 a 3 horas corridas do atraso. A mesma explicação para tal fato, pode ser semelhante à explicação em SBBR. Como nas primeiras duas horas o número de aeroportos que impactam SBCT é significante, ele demora para se recuperar até a terceira hora, que também sente reflexo de 2 aeroportos nesse momento. Nos outros aeroportos, poucos são impactados por outros e não possuem autopropagação, conforme a Tabela 19.

Tabela 22: Número de aeroportos que impactam, a cada hora corrida de seus atrasos, os aeroportos do Sul do Brasil

Aeroporto	1 Hora	2 Horas	3 Horas	4 Horas
SBCT	7	7	2	1
SBPA	1	1	0	0
SBFL	0	0	0	0

Por fim, no Nordeste, SBSV apresenta a maior duração da autopropagação, conforme Tabela 19. Sua duração se extende de 1 a 3 horas. Ou seja, SBSV se recupera de atrasos anteriores depois de 3 horas corridas. Essa duração extensa pode ser explicada pelo fato de que outros aeroportos propagam seus atrasos em SBSV principalmente entre 1 a 3 horas depois, conforme Tabela 23. Logo, SBSV, nesse período, tem mais dificuldade de se recuperar, com isso, sua taxa de atraso alta permanece durante esse tempo. SBRF mostra uma pecularidade, a duração da sua autopropagação demonstra permanecer por conta de atrasos internos, pois sua indicação de

duração entre 1 a 2 horas não corresponde a concentração de impactos externos, que está num período de 2 a 4 horas. E, SBFZ, não demostra autopropagação na Tabela 19, mas apresenta aeroportos propagando atrasos no mesmo, conforme 23. Essa situação pode ser explicada pelo fato de que SBFZ tem uma recuperação rápida nos atrasos, não os acumulando e assim, não propagando esses atrasos para horas seguintes.

Tabela 23: Número de aeroportos que impactam, a cada hora corrida dos atrasos, os aeroportos do Nordeste do Brasil

Aeroporto	1 Hora	2 Horas	3 Horas	4 Horas
SBSV	4	3	3	1
SBFZ	0	0	1	2
SBRF	1	3	3	3

Logo, podemos concluir com essa análise que grandes aeroportos como SBGR e SBGL demoram mais para se recuperarem de atrasos passados. Uma grande influência dessa auto-propagação pode ser explicada pelos reflexos dos atrasos sentidos de outros aeroportos. Essa mesma conclusão pode ser encontrada para a autoprapagação dos demais aeroportos, como SBBR e SBCT.

Capítulo 6

Conclusão

Analisar a propagação de atrasos no cenário dos voos brasileiros mostrou ser pouco explorada e importante de ser estudada. Voos atrasados podem prejudicar aeroportos, companhias e passageiros. Companhias recebem multas e passageiros devem reorganizar suas viagens por conta desses atrasos ocorridos. A propagação dos mesmos pode se extender para outros aeroportos, prejudicando toda a malha aérea. No ponto de vista do sistema aéreo brasileiro, os atrasos são bem recorrentes anualmente, tendo em média 30% dos voos atrasados entre 2009 e 2014. Além disso, na malha aérea brasileira, encontramos importantes aeroportos em escala mundial, como Guarulhos. Dito isso, estudar a propagação dos atrasos no Brasil se mostrou interessante. Portanto, essa pesquisa envolve uma análise sobre a propagação de atrasos nos aeroportos brasileiros, conduzida a responder as seguintes perguntas: (i) Os aeroportos estão interligados quando um atraso ocorre? (ii) Por quanto tempo um atraso se propaga entre os aeroportos? (iii) Como é o comportamento da propagação dentro do próprio aeroporto?

Nos trabalhos relacionados, o cenário da propagação de atrasos foi explorado significamente. Vários artigos utilizando diversos métodos foram encontrados para estudar as propagações dos atrasos. Contudo, a maioria deles sendo aplicada nas regiões da Europa e Estados Unidos. Além disso, nenhum artigo estrangeiro encontrado utilizou o método de mineração de dados de Padrões Frequentes para a realização de suas pesquisas. Apenas um artigo nacional englobou a proposta de estudar voos atrasados no cenário brasileiro com o uso de Padrões Frequentes [Sternberg et al., 2016], mas sem focar no estudo da propagação de atrasos. Portanto, há uma necessidade de explorar esse cenário nos aeroportos brasileiros.

Para ser possível a realização dessa pesquisa, coletamos os dados do *dataset* proveniente da Agência Nacional de Aviação Civil (ANAC), mais especificamente a tabela pública e atualizada mensalmente chamada VRA. Uma grande quantidade de dados foi utilizada, de 2009 ao início de 2015. Aproveitamos os dados pré-processados realizado no trabalho de [Sternberg et al., 2016], que são os mesmos dados necessitados para o desenvolvimento dessa pesquisa. Foram realizadas várias atividades de pré-processamento, para melhorar a qualidade dos resultados, que incluem: seleção dos principais atributos a serem utilizados na mineração de dados, identi-

ficação dos principais aeroportos, limpeza de dados, transformação dos dados e materialização da janela deslizante.

A materialização da janela deslizante foi um processo extenso e delicado da pesquisa. Foi a partir dessa janela que aplicamos o processo de mineração de dados, utilizando o método de Padrões Frequentes. No total, foram criadas 17 janelas, uma para cada aeroporto. Essas janelas contém as porcentagens de atrasos do aeroporto que está sendo analisado e de outros aeroportos da rede, em até 4 horas depois de um atraso ocorrido. Assim, se torna possível analisar a propagação em outros aeroportos e neles próprios. Essas porcentagens foram divididas, através da técnica *binning*, em intervalos qualitativos, representados por: *low, medium* e *high*.

O método de Padrões Frequentes foi aplicado em cima dessas janelas, a fim de encontrar regras de associação. O algoritmo Apriori foi usado e definimos o suporte e a confiança como 7% e 25%, respectivamente. Suporte corresponde à probabilidade (frequência) de um padrão ocorrer na base de dados e a confiança se refere à probabilidade condicional do padrão acontecer. Com a automatização do processo de mineração foi possível fazer uma análise de sensibilidade desses atributos a fim de garantir sua precisão. Regras de associação interessantes foram encontradas decorrentes do algoritmo Apriori, ainda assim, para consolidar o valor dos resultados, aplicamos uma filtragem que seleciona as regras com um *lift* maior do que um. Além disso, apenas regras apresentando taxas *high* foram consideradas. Ou seja, apenas consideramos regras contendo a informação de que um aeroporto com uma taxa alta de atrasos os propaga para outro aeroporto influenciando numa taxa alta de atrasos também.

Os resultados mostraram que os aeroportos brasileiros estão interligados no quesito da propagação do atraso. Pelo conhecimento prévio que temos sobre os aeroportos brasileiros, esperávamos que alguns, como o de Guarulhos (SBGR), tivesse um grande impacto na propagação de atrasos em outros aeroportos. Os resultados do processo de mineração confirmaram algumas expectativas. A análise da relação entre os aeroportos deixa claro a extrema influência negativa dos atrasos do aeroporto SBGR no restante do país. Enquanto aeroportos menores como SBFZ impactam quatro aeroportos da rede, SBGR chega a impactar nove.

A evidência dos impactos negativos não estão apenas ligados à quantidade de aeroportos impactados, mas também a quantidade de horas que o aeroporto de origem continua a influenciar os demais. Os aeroportos SBGR e SBGL, que são os mais influentes na malha aerea brasileira, mantém sua influência em até 4 horas depois. Já SBSP mantém até 3 horas e SBRJ apenas até 2 horas após ter uma grande quantidade de atrasos ocorridos.

Por fim, foi possível também analisar a autopropagação em cada região brasileira. Os resultados mostraram que o principal problema da permanência do atraso está nos próprios aeroportos. A dificuldade de cada aeroporto se recuperar de um atraso prévio é bastante impactante, o que acaba por aumentar ainda mais o número de atrasos oriundos desse aeroporto.

Logo, conseguimos responder as três perguntas formuladas para o desenvolvimento dessa pesquisa. Para uma investigação futura e mais detalhada dos atrasos, seria interessante considerar também a motivação desses atrasos. Incluir atributos de outras dimensões, como metereológica, pode ser bastante útil para evidenciar dependências no sistema aéreo brasileiro.

Referências Bibliográficas

Aeroporto de Brasília (2016). Sobre o aeroporto. Technical report, http://www.bsb.aero/br/o-aeroporto/sobre-o-aeroporto-de-brasilia/.

Aeroporto Internacional de São Paulo (2015). Sobre gru airport. Technical report, http://www.gru.com.br/pt-br/Institucional.

Agrawal, R., Srikant, R., and others (1994). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499.

AhmadBeygi, S., Cohn, A., Guan, Y., and Belobaba, P. (2008). Analysis of the potential for delay propagation in passenger airline networks. *Journal of Air Transport Management*, 14(5):221 – 236.

Ahmed, C. F., Tanbeer, S. K., Jeong, B.-S., Lee, Y.-K., and Choi, H.-J. (2012). Single-pass incremental and interactive mining for weighted frequent patterns. *Elsevier*.

ANAC (2015). Agência Nacional de Aviação Civil. Technical report, http://www.anac.gov.br/.

Baspinar, B., Ure, N., Koyuncu, E., and Inalhan, G. (2016). Analysis of delay characteristics of european air traffic through a data-driven airport-centric queuing network model. *IFAC-PapersOnLine*.

Britto, R., Dresner, M., and Voltes, A. (2012). The impact of flight delays on passenger demand and societal welfare. *Transportation Research Part E: Logistics and Transportation Review*, 48(2):460 – 469.

Burden, R., Faires, J., and Burden, A. (2015). *Numerical analysis*. Nelson Education.

Campanelli, B., Fleurquin, P., Arranz, A., Etxebarria, I., Ciruelos, C., Eguíluz, V., and Ramasco, J. (2016). Comparing the modeling of delay propagation in the us and european air traffic networks. *Journal of Air Transport Management*.

Deypir, M. and Sadreddini, M. H. (2011). A dynamic layout of sliding window for frequent itemset mining over data streams. *Elsevier*.

Deypir, M., Sadreddini, M. H., and Hashemi, S. (2012). Towards a variable size sliding window model for frequent itemset mining over data streams. *Elsevier*.

EUROCONTROL (2015). Annual Network Operations Report 2014. Technical report, https://www.eurocontrol.int/sites/default/files/publication/ performance/2014_annual/final/annual_network_operations_report_2014_main_report_final_edition.pdf.

Han, J., Kamber, M., and Pei, J. (2011). *Data Mining: Concepts and Techniques, Third Edition*. Morgan Kaufmann, Waltham, Mass., 3 edition edition.

Han, J., Pei, J., and Yin, Y. (2000). Mining frequent patterns without candidate generation. In *ACM SIGMOD Record*, volume 29, pages 1–12. ACM.

Hunter, G., Boisvert, B., and Ramamoorthy, K. (2007). Advanced national airspace traffic flow management simulation experiments and vlidation. In *Simulation Conference, 2007 Winter*, pages 1261–1267.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer.

Meng, H. and Peng, Y. (2015). Two-stage extraction method for flight delay pattern. *Harbin Gongye Daxue Xuebao/Journal of Harbin Institute of Technology*.

Pyrgiotis, N., Malone, K. M., and Odoni, A. (2013). Modelling delay propagation within an airport network. *Transportation Research Part C: Emerging Technologies*, 27(0):60 – 75.

Qiu, S., Wu, W., and Hou, M. (2015). Correlation analysis of flight delay based on copula function. *Wuhan Ligong Daxue Xuebao (Jiaotong Kexue Yu Gongcheng Ban)/Journal of Wuhan University of Technology (Transportation Science and Engineering)*.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reynolds-Feighan, A. J. and Button, K. J. (1999). An assessment of the capacity and congestion levels at European airports. *Journal of Air Transport Management*, 5(3):113 – 134.

Secretaria de Aviação Civil (2015). Os 65 aeroportos que movimentam o brasil. Technical report, http://www.aviacao.gov.br/noticias/2015/10/conheca-os-65-aeroportos-que-movimentam-o-brasil.

Silberschatz, A., Korth, H., and Sudarshan, S. (2010). *Database System Concepts*. McGraw-Hill Science/Engineering/Math, New York, 6 edition edition.

Sternberg, A., Carvalho, D., Murta, L., Soares, J., and Ogasawara, E. (2016). An analysis of brazilian flight delays based on frequent patterns. *Elsevier*.

Tiao, G. C. (1972). Asymptotic behaviour of temporal aggregates of time series. *Biometrika*, 59(3):525–531.

Tu, Y., Ball, M. O., and Jank, W. S. (2008). Estimating flight departure delay distributions—a statistical approach with long-term trend and short-term pattern. *Journal of the American Statistical Association*, 103(481):112–125.

Xu, N., Donohue, G., Laskey, K. B., and Chen, C.-H. (2005). Estimation of delay propagation in the national aviation system using Bayesian networks. In *6th USA/Europe Air Traffic Management Research and Development Seminar*. Citeseer.

Xu, X. and Li, S. (2015). Identification and prediction of air traffic congestion. *Hangkong Xuebao/Acta Aeronautica et Astronautica Sinica*.

Zaki, M. J., Parthasarathy, S., Ogihara, M., Li, W., and others (1997). New Algorithms for Fast Discovery of Association Rules. In *KDD*, volume 97, pages 283–286.

Zhao, X., Tang, J., Lu, F., and Han, B. (2016). Strategy analysis for delayed fights pushback and sensitivity analysis of the length of virtual queue. *Sichuan Daxue Xuebao (Gongcheng Kexue Ban)/Journal of Sichuan University (Engineering Science Edition)*.