CEFET/RJ
DIRETORIA DE PESQUISA E PÓS-GRADUAÇÃO

ANÁLISE DE PADRÕES FREQUENTES DE ATRASOS EM VOOS NACIONAIS

Alice Wagner Sternberg

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia de Produção e Sistemas, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ, como parte dos requisitos necessários à obtenção do título de mestre.

Orientadores:
Eduardo Soares Ogasawara
Jorge de Abreu Soares

Rio de Janeiro,
Abril 2016

Análise de Padrões Frequentes de Atrasos em Voos Nacionais

Dissertação de Mestrado em Engenharia de Produção e Sistemas, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ.

Alice Wagner Sternberg

Aprovada por:

Presidente, Prof. Eduardo Soares Ogasawara, D.Sc. (orientador)

Prof. Jorge de Abreu Soares, D.Sc. (co-orientador)

Prof. Diego Moreira de Araujo Carvalho, D.Sc.

Prof. Leonardo Gresta Paulino Murta, D.Sc.

Ficha catalográfica elaborada pela Biblioteca Central do CEFET/RJ

S839 Sternberg, Alice Wagner
Análise de padrões frequentes de atrasos em voos nacionais / Alice Wagner Sternberg.—2016.
x, 58f. : il.color. , grafs. , tabs. ; enc.

Dissertação (Mestrado) Centro Federal de Educação Tecnológica Celso Suckow da Fonseca , 2016.
Bibliografia : f. 54-58
Orientadores : Eduardo Soares Ogasawara
Jorge de Abreu Soares

1. Aeronáutica comercial – Brasil. 2. Aeronáutica – Voos – Brasil. 3. Mineração de dados (Computação). 4. Engenharia de produção. I. Ogasawara, Eduardo Soares (Orient.). II. Soares, Jorge de Abreu (Orient.). III. Título.

CDD 387.70981

RESUMO

Análise de Padrões Frequentes de Atrasos em Voos Nacionais

Alice Wagner Sternberg

Orientadores:
Eduardo Soares Ogasawara
Jorge de Abreu Soares

Resumo da Dissertação submetida ao Programa de Pós-graduação em Engenharia de Produção e Sistemas do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ como parte dos requisitos necessários à obtenção do título de mestre.

Os atrasos aéreos constituem um dos grandes desafios do setor de aviação comercial. Além da fidelidade de passageiros, os atrasos provocam custos adicionais às companhias aéreas, congestionamentos em aeroportos e no espaço aéreo e danos ambientais a partir do aumento do consumo de combustível. Desta forma, a compreensão das condições que conduzem a estes atrasos pode auxiliar o processo de tomada de decisão de companhias aéreas, aeroportos e investidores, direcionando seus esforços para as reais causas dos atrasos. Durante os últimos anos, muitas publicações estudaram atrasos aéreos e seus impactos por meio de análises de suas distribuições ou correlações, especialmente nos Estados Unidos, Europa e China. Neste contexto, a presente pesquisa apresenta uma aplicação original de regras de associação para a investigação das causas de atrasos aéreos, permitindo a descoberta de padrões ocultos e suas correlações com os atrasos. Além disso, apesar do importante mercado brasileiro de aviação comercial, esta é a primeira pesquisa a analisar os atrasos aéreos do país sob o ponto de vista de ciência de dados. Para isso, a metodologia aplicada inclui a construção de um *data warehouse* para a integração de múltiplas fontes de dados, etapas de pré-processamento, discussão de medidas de interesse para a geração das regras de associação e, finalmente, a análise das regras de maneira geral e comparativa. Os resultados apontam às principais causas de atrasos no Brasil, bem como comportamentos e tendências de aeroportos e companhias aéreas.

Palavras-chave:
Atrasos aéreos; Padrões frequentes; Aviação comercial.

Rio de Janeiro,

Abril 2016

ABSTRACT

Analysis of Frequent Patterns of Flight Delays in National Flights

Alice Wagner Sternberg

Advisors:
Eduardo Soares Ogasawara
Jorge de Abreu Soares

Abstract of dissertation submitted to Programa de Pós-graduação em Engenharia de Produção e Sistemas - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ as partial fulfillment of the requirements for the degree of master.

Flight delays are one of the major challenges of commercial aviation. Beyond passengers loyalty, delays cause additional costs to airlines, congestion at airports and in airspace, and environmental damages by increasing fuel consumption. Thus, understanding the conditions that lead to flight delays may improve the decision making process of airlines, airports and investors, directing their efforts towards the real reasons of delays. Over the last years, many journal papers evaluated flight delays impacts and their distributions or correlations, especially in the United States, Europe and China. In this context, this research presents the original usage of association rules to investigate the causes of flight delays, leading to the discovery of hidden patterns in data and their correlations with delays. Furthermore, despite the importance of commercial aviation industry in Brazil, this is the first study to analyze Brazilian flight delays using a data science approach. Therefore, the methodology of this research includes the construction of a data warehouse to integrate multiple sources of data, preprocessing stages, the discussion of interest measures for rules generation and, finaly, the analysis of the association rules in general and comparative ways. Results show the main causes of delays in Brazil and behaviors and trends of airports and airlines.

Key-words:
Flight delays; Frequent patterns; Commercial aviation.

Rio de Janeiro,

April 2016

Sumário

I Introdução **1**

II O Cenário de Atrasos Aéreos **5**

II.1 Problema 6

II.2 Escopo 7

II.3 Dados 7

II.4 Métodos 10

II.5 Considerações sobre análises de atrasos e cancelamentos aéreos 13

III Mineração de Padrões Frequentes **15**

III.1 Processo de Mineração de Dados 15

III.2 Integração 15

III.3 Seleção 17

III.4 Limpeza 18

III.5 Transformação 19

III.6 Padrões Frequentes 20

III.7 Métodos de avaliação 22

IV Metodologia **24**

IV.1 Integração 25

IV.2 Seleção 26

IV.3 Limpeza 29

IV.4 Transformação 30

IV.5 Geração das Regras de Associação 33

V Padrões Frequentes em Atrasos e Cancelamentos de Voos Nacionais **35**

V.1 Análise de regras de comprimento dois 36

V.2 Análise de regras de comprimento três 37

V.3 Análise de regras de comprimento quatro 39

V.4 Análises complementares 43

VI Trabalhos Relacionados **48**

VII Conclusão **51**

Referências Bibliográficas 53

Lista de Figuras

II.1 Operação típica de um voo comercial 5

II.2 Taxonomia da análise de atrasos e cancelamentos aéreos 6

II.3 Dimensões dos dados de voos 9

II.4 Linha do tempo das principais publicações relacionadas à análise de atrasos e cancelamentos aéreos considerando seu tipo de problema, região de aplicação e método 14

III.1 Mineração de dados como um processo de descoberta de conhecimento. Fonte: adaptado de Han, Kamber e Pei (2011) 16

III.2 Exemplo do processo de construção de um *data warehouse* de voos nacionais 17

III.3 Exemplo de *boxplot* para a identificação de *outliers* 19

IV.1 Processo de ciência de dados para analise de atrasos nos voos brasileiros 25

IV.2 Diagrama de classes para integração de múltiplas fontes de dados 27

IV.3 Redução do número de tuplas do *dataset* após a etapa de seleção 28

IV.4 Exemplo de verificação de *outliers* na rota Guarulhos (São Paulo) - Galeão (Rio de Janeiro) 30

IV.5 Redução do número de tuplas do *dataset* após a etapa de limpeza 31

V.1 Dimensões e atributos do *dataset* final de voos brasileiros 35

V.2 *Boxplot* dos *lifts* das regras de associação de *comprimento = 3* 38

V.3 *Boxplot* dos *lifts* das regras de associação de *comprimento = 4* 40

V.4 Análise do *lift* das regras de associação contendo somente o ano de partida no Lado esquerdo - Left Hand Side (LHS) e a ocorrência de um atraso no Lado direito - Right Hand Side (RHS) 44

V.5 Análise do *lift* das regras de associação contendo a companhia aérea e o ano de partida no LHS e a ocorrência de um atraso no RHS 45

V.6 Análise do *lift* das regras de associação contendo o aeroporto e o ano de partida no LHS e a ocorrência de um atraso no RHS 45

V.7 Análise do *lift* das regras de associação contendo somente o período do dia no LHS e a ocorrência de um atraso no RHS 46

V.8 Análise do *lift* das regras de associação contendo o aeroporto e o período do dia no LHS e a ocorrência de um atraso no RHS 47

Lista de Tabelas

III.1 Exemplo de detecção de discrepâncias e *outliers* 19

III.2 Exemplo de um *dataset* representando as operações do turno da manhã de um determinado aeroporto 22

III.3 Examplo de padrões frequentes para um suporte de 50% 22

IV.1 Aeroportos selecionados para a análise 28

IV.2 Número de registros resultantes após a etapa de seleção 28

IV.3 Número de registros resultantes após a etapa de limpeza 30

IV.4 Transformações 32

IV.5 Definição do comprimento máximo e do suporte 34

V.1 Número de regras interessantes para cada comprimento 36

V.2 Regras para atrasos aéreos de *comprimento = 2* com *lift* superior a 1,1 36

V.3 Dez primeiros *lifts* das regras interessantes para atrasos aéreos de *comprimento = 3* 38

V.4 Dez primeiros *lifts* das regras interessantes para atrasos aéreos de *comprimento = 4* 40

V.5 Principais regras de associação para atrasos aéreos gerados nos aeroportos do Rio de Janeiro e de São Paulo 42

V.6 Principais regras de associação para atrasos nos voos entre os aeroportos de Congonhas (São Paulo) e Santos Dumont (Rio de Janeiro) 43

VI.1 Classificação das publicações relacionadas a atrasos aéreos quanto ao tipo de problema 48

VI.2 Classificação das publicações relacionadas a atrasos aéreos quanto ao método aplicado 48

Lista de Abreviações

ANAC	Agência Nacional De Aviação Civil 25, 50
BAV	Boletim De Alteração De Voo 25
BDMEP	Banco De Dados Meteorológicos Para Ensino E Pesquisa 26
BKN	Nublado - Broken 32
DAC	Departamento De Aviação Civil 25
ETL	Extração, Transformação E Carga - Extract, Transform, Load 16
FEW	Poucas Nuvens - Few 32
HOTRAN	Horários De Transporte 25, 26
IAC	Instrução De Aviação Civil 25, 29
INMET	Instituto Nacional De Meteorologia 26, 50
LHS	Lado Esquerdo - Left Hand Side .. vii, viii, 22, 23, 33, 36, 38, 39, 40, 42, 43, 44, 45, 46, 47, 52
METAR	Informe Meteorológico Aeronáutico Regular - Meteorological Aerodrome Report 20, 31
OACI	Organização Da Aviação Civil Internacional 2, 28
OVC	Completamente Encoberto - Overcast 32, 36, 38, 40, 42, 43
RHS	Lado Direito - Right Hand Side ..... vii, viii, 22, 33, 36, 43, 44, 45, 46, 47
RPK	Passageiros-Quilômetros Pagos Transportados - Revenue Passenger-Kilometer 2
SCT	Nuvens Esparsas - Scattered 32
SKC	Céu Limpo - Sky Clear 32
VRA	Voo Regular Ativo 25, 26

Capítulo I Introdução

Os atrasos constituem importantes indicadores de desempenho em qualquer sistema de transporte. No setor de aviação comercial, o atraso pode ser definido como a diferença entre os horários realizados e previstos para partidas e chegadas de um voo (WIELAND, 1997). Apesar de, muitas vezes, utilizarem esta mesma definição, as autoridades reguladoras de diferentes países possuem tolerâncias diversas quanto aos limites de atraso. Em 2014, 16,6% dos voos atrasaram mais de 15 minutos na Europa (EUROCONTROL, 2015) e 24,7% nos Estados Unidos (THE UNITED STATES DEPARTMENT OF TRANSPORTATION, 2015). No Brasil, 19,1% foram cancelados ou sofreram atrasos superiores a 30 minutos (ANAC, 2015b).

Considerando o setor de aviação comercial, os atrasos e cancelamentos de voos provocam consequências negativas tanto para passageiros como para companhias aéreas e aeroportos. Dada à incerteza de sua ocorrência, usualmente, os passageiros planejam suas viagens com bastante antecedência em relação a seus compromissos, aumentando os custos, mas garantindo sua chegada a tempo. Por outro lado, as companhias aéreas sofrem com multas e custos extras, como realocação de tripulação e aviões parados em pátios de aeroportos (BRITTO; DRESNER; VOLTES, 2012). Além disso, atrasos podem causar danos ambientais por meio do aumento do consumo de combustível e de emissões de gases (PEJOVIC et al., 2009; RYERSON; HANSEN; BONN, 2014; SIMIć; BABIć, 2015).

Outro problema está ligado às estratégias de *marketing* de companhias aéreas, uma vez que a fidelidade de passageiros é afetada pela pontualidade dos voos (VLACHOS; LIN, 2014). Além disso, atrasos estão relacionados às tarifas praticadas pelas companhias aéreas, bem como ao tamanho das aeronaves, frequência dos voos e reclamações sobre o serviço aéreo (BHADRA, 2009; PAI, 2010; ZOU; HANSEN, 2014). Desta forma, a correta compreensão das condições que conduzem a atrasos e cancelamentos pode auxiliar na tomada de decisão de gestores de companhias aéreas e aeroportos ou de investidores públicos e privados, além de auxiliar passageiros a organizarem suas viagens (LV; WANG, 2009).

Com a finalidade de compreender o sistema de transporte aéreo, um grande volume de dados sobre as operações de aviação comercial tem sido coletado e armazenado em bancos de dados de instituições públicas e privadas. Submersos nesta grande quantidade de dados, analistas e cientistas de dados têm intensificado suas pesquisas de modo a obter conhecimento novo e

útil a partir destes dados, como, por exemplo, a compreensão de causas e previsões de atrasos, análises da propagação destes atrasos ou do relacionamento entre aeroportos do sistema. Neste contexto, as análises de ciência de dados, que envolvem a compreensão do domínio, a gestão de dados e a aplicação de um modelo, têm sido uma tendência na resolução de problemas de *big data* relacionados a atrasos aéreos (DHAR, 2013; JAGADISH et al., 2014; MATSUDAIRA, 2015).

Nos últimos 20 anos, muitas publicações têm discutido atrasos e cancelamentos aéreos, especialmente nos Estados Unidos, Europa e Ásia (REYNOLDS-FEIGHAN; BUTTON, 1999; SCHAEFER; MILLNER, 2001; HANSEN, 2002; LU; WANG; XU, 2009). Além de dados de aeroportos e companhias aéreas, a existência de bases de dados públicas de instituições governamentais, como *The Bureau of Transportation Statistics* (BTS, 2015) e *The Federal Aviation Administration* (FAA, 2015), ligadas ao *The United States Department of Transportation* (DOT, 2015) favorecem abordagens de ciência de dados a partir de diferentes dimensões e métodos. Nas regiões citadas, estes trabalhos têm contribuído para encontrar padrões (ABDEL-ATY et al., 2007) ou distribuições de atrasos (MUELLER; CHATTERJI, 2002; TU; BALL; JANK, 2008), para prever a sua ocorrência (REBOLLO; BALAKRISHNAN, 2014) ou para comparar aeroportos (PATHOMSIRI et al., 2008) e companhias aéreas (AHMADBEYGI et al., 2008).

No Brasil, apesar da importância do setor de aviação comercial, das características geográficas e de infraestrutura e da existência de bases de dados públicas e de regulamentações quanto à divulgação de percentuais de atrasos e cancelamentos, não se tem conhecimento de trabalhos que analisam, sob o ponto de vista de ciência de dados, as condições que levam a atrasos e cancelamentos. Cabe destacar que, em 2014, o país foi o primeiro colocado entre os latinos americanos e décimo quinto mundialmente no *ranking* de Passageiros-Quilômetros Pagos Transportados - Revenue Passenger-Kilometer (RPK) da Organização da Aviação Civil Internacional (OACI) (ICAO, 2015). Este indicador é calculado a partir da multiplicação da distância de um voo em quilômetros pelo seu número de passageiros pagantes, demonstrando que, além de grandes distâncias, o sistema brasileiro de aviação comercial também transporta uma quantidade considerável de passageiros.

A lacuna existente na avaliação de atrasos e cancelamentos aéreos no Brasil faz com que o conhecimento sobre este assunto permaneça empírico, de modo que as decisões de companhias aéreas, aeroportos e investidores podem não estar considerando todos os fatores associados a atrasos e cancelamentos. Diante do cenário atual, este trabalho tem como objetivo a construção de um panorama geral das principais condições que conduzem a atrasos e cancelamentos nos voos domésticos e comerciais no Brasil. Sendo assim, de caráter exploratório, esta pesquisa pretende descobrir e analisar estas causas, além de comparar o desempenho de aeroportos e companhias aéreas em relação à ocorrência de atrasos.

Por se tratar de um problema com grande quantidade de dados e com a finalidade de descobrir conhecimento novo e útil para a tomada de decisão de todos os atores envolvidos nas operações de voos domésticos e comerciais, este trabalho foi conduzido sob uma perspectiva de ciência de dados. Sendo assim, um *data warehouse* foi construído em PostgreSQL (PostgreSQL GLOBAL DEVELOPMENT GROUP, 2015), de modo a integrar as bases de dados contendo informações sobre as operações dos voos (ANAC, 2015a) e condições meteorológicas (INMET, 2015). Após a execução de etapas de seleção, limpeza e transformação, o método de mineração de padrões frequentes foi aplicado em R (R CORE TEAM, 2014) para a obtenção de regras de associação relacionadas aos atrasos aéreos. Esta técnica, aplicada de maneira original a problemas de aviação comercial, permite a compreensão dos dados e de seus relacionamentos a partir da visualização de associações e correlações, mostrando-se adequada para a construção do panorama geral das causas de atrasos e cancelamentos no Brasil.

Este estudo pretende começar a cobrir o espaço existente no Brasil em relação à análise de atrasos e cancelamentos aéreos sob o ponto de vista de ciência de dados. Espera-se que o conhecimento gerado, além de auxiliar passageiros, companhias aéreas, aeroportos e investidores, possa motivar trabalhos futuros com novas abordagens sobre este tema, assim como tem sido realizado em outros países. Sendo assim, dentre as principais contribuições desta pesquisa destacam-se: (i) elaboração de uma taxonomia e uma linha do tempo para análise da literatura mundial relacionada a atrasos e cancelamentos aéreos, (ii) construção de um panorama geral das causas de atrasos e cancelamentos aéreos no Brasil e (iii) comparação do desempenho de companhias aéreas e aeroportos brasileiros em relação a atrasos e cancelamentos.

A elaboração de um mapa sistemático para a avaliação da literatura mundial relacionada a atrasos e cancelamentos aéreos permitiu a identificação dos principais trabalhos da área e das abordagens utilizadas para a condução dos estudos. Desta forma, foi construída uma taxonomia da análise de atrasos e cancelamentos aéreos, dividida em questões ligadas ao domínio, como tipo de problema e escopo e à ciência de dados, como fontes, dimensões e gestão de dados e métodos. Além disso, a linha do tempo das principais publicações possibilitou a visualização clara do aumento do número de trabalhos nos últimos anos, bem como os principais tipos de problemas e métodos abordados nos estudos.

A condução da análise de atrasos e cancelamentos aéreos no Brasil a partir da construção de um *data warehouse*, da execução de etapas de pré-processamento e da aplicação original de regras de associação permitiram a construção de um panorama geral das principais causas de atrasos e cancelamentos no país. Dentre os principais resultados, ou seja, condições que conduzem a atrasos e cancelamentos, destacam-se a identificação da forte propagação de atrasos e dificuldades na recuperação de alguns aeroportos e de uma companhia aérea em cenários

de atrasos preexistentes. Além disso, foram verificadas influências temporais, como dia da semana e período do dia, bem como meteorológicas, como a nebulosidade. Estes casos foram especialmente discutidos para os aeroportos de São Paulo e Rio de Janeiro.

A identificação das principais causas de atrasos e cancelamentos aéreos do Brasil foi utilizada como base para a comparação do desempenho de companhias aéreas e aeroportos. Verificou-se que a companhia aérea mais relacionada a atrasos no período da análise tem apresentado significativas melhoras nos últimos anos. Para os aeroportos, foram avaliadas as performances ao longo dos anos e durante os diferentes períodos do dia. Os resultados mostraram comportamentos e tendências diversas, possibilitando a visualização de aeroportos frequentemente atrasados, aqueles com alguns picos de atrasos e, finalmente, aqueles com desempenhos mais regulares e pouco associados a atrasos.

Além da presente introdução, esta pesquisa está organizada em outros seis capítulos. O Capítulo II introduz o cenário de atrasos e cancelamentos aéreos, descrevendo as operações típicas da aviação comercial e apresentando uma taxonomia dos principais conceitos e uma linha do tempo das publicações da área. O Capítulo III apresenta o processo de mineração de dados e a mineração de padrões frequentes, discutindo os conceitos de regras de associação e de suas medidas de interesse. O Capítulo IV apresenta a metodologia para a geração de regras de associação para a análise de atrasos e cancelamentos aéreos no Brasil, destacando as etapas de pré-processamento, a definição das medidas de interesse e o processo de geração das regras. A avaliação das regras de associação, ou seja, a análise das condições que conduzem a atrasos e cancelamentos é apresentada no Capítulo V, destacando as condições, os aeroportos e as companhias aéreas com maiores chances de produzir novos atrasos. O Capítulo VI destaca os principais trabalhos publicados relacionados à análise de atrasos aéreos, além de uma comparação com os resultados obtidos para o caso brasileiro. Finalmente, a Capítulo VII discute as principais contribuições deste trabalho e indica possíveis soluções para minimizar a ocorrência de atrasos no sistema aéreo brasileiro, apontando também algumas possibilidades para estudos futuros.

Capítulo II O Cenário de Atrasos Aéreos

A aviação comercial é um sistema de transporte bastante complexo. Além de lidar com recursos caros e flutuação de demanda, seus diversos pontos de origem e de destino podem exigir diferentes condições para a garantia da segurança e da confiabilidade das operações. Seus usuários, ou seja, os passageiros seguem seus itinerários ao longo do dia, enquanto as companhias aéreas provedoras do serviço planejam diferentes operações para suas aeronaves, pilotos e comissários (BALL et al., 2007). A Figura II.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, tais como problemas mecânicos, condições meteorológicas, problemas nas operações de solo, filas para decolagem ou pouso e restrições de capacidade (REYNOLDS-FEIGHAN; BUTTON, 1999; HUNTER; BOISVERT; RAMAMOORTHY, 2007; AHMADBEYGI et al., 2008).

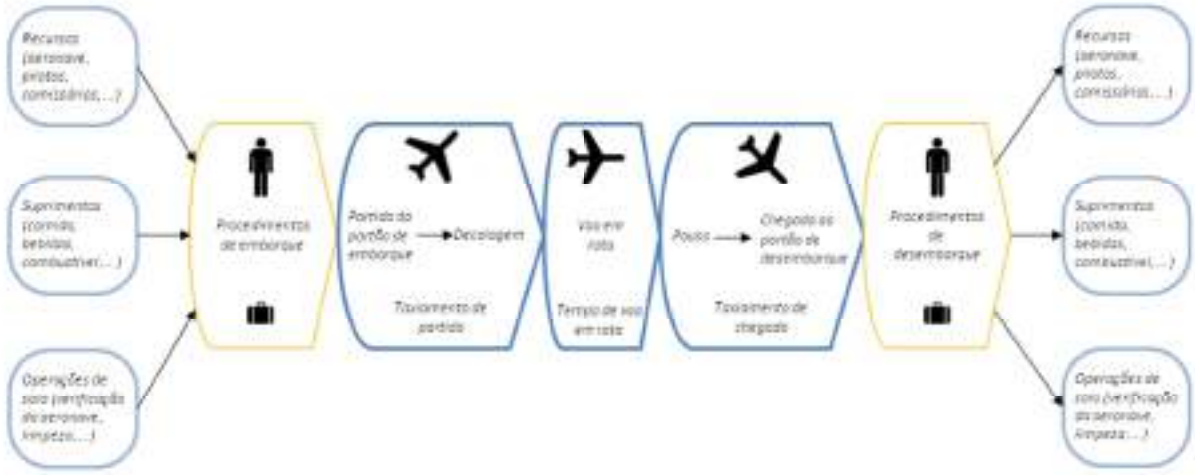


Figura II.1: Operação típica de um voo comercial

Este esquema se repete várias vezes ao longo do dia para cada voo do sistema. Os pilotos, comissários e aeronaves podem ter programações diferentes, devido à regulamentação para descansos e folgas e ao plano de manutenção das aeronaves. Desta forma, qualquer atraso ou cancelamento pode afetar os voos subsequentes de uma determinada companhia aérea (ABDELGHANY et al., 2004). Além disso, estes atrasos ou cancelamentos podem provocar congestionamento no espaço aéreo ou em aeroportos, afetando voos de outras companhias aéreas do sistema (SCHAEFER; MILLNER, 2001; XU et al., 2005). Sendo assim, a compreensão das

causas e do comportamento de atrasos e cancelamentos pode auxiliar na tomada de decisão de companhias aéreas quanto ao planejamento de seus recursos críticos, além de direcionar investimentos públicos e privados e auxiliar passageiros na organização de suas viagens.

O crescimento das bases de dados de aviação comercial nos últimos anos tem impulsionado as publicações sobre atrasos e cancelamentos aéreos. Desta forma, a construção destas análises pode ser desdobrada na taxonomia apresentada na Figura II.2. O domínio refere-se à escolha do problema e à delimitação do escopo, enquanto as atividades de ciência de dados incluem a coleta, seleção e gestão de dados e a escolha de um método para a análise propriamente dita.



Figura II.2: Taxonomia da análise de atrasos e cancelamentos aéreos

II.1 Problema

O problema é o ponto central do domínio de análises de atrasos e cancelamentos aéreos. Apesar do setor de aviação comercial constituir um sistema dinâmico em que novos atrasos e cancelamentos podem ocorrer a qualquer momento e se propagar para o restante da rede, muitos trabalhos isolam partes do sistema e, desta forma, focam em apenas uma situação específica. Sendo assim, mesmo que ocorrendo simultaneamente em sistemas reais, para fins de pesquisa, o problema de atrasos e cancelamentos aéreos pode ser visto de três maneiras distintas:

(i) Propagação de atrasos: pressupõe que um atraso já ocorreu em algum ponto do sistema e analisa sua propagação para os voos de uma determinada companhia aérea (BOSWELL; EVANS, 1997; BEATTY et al., 1998; ABDELGHANY et al., 2004; WU, 2005; WONG; TSAI, 2012), para a rede de aeroportos (XU et al., 2005; PYRGIOTIS; MALONE; ODONI, 2013; HAO et al., 2014) ou em ambos os casos (HANSEN, 2002);

(ii) Novos atrasos: considerando que novas ocorrências podem acontecer a qualquer momento, este tipo de problema procura compreender as condições que conduzem a novos atra-

sos (ABDEL-ATY et al., 2007; HUNTER; RAMAMOORTHY, 2007; BALAKRISHNA; GANESAN; SHERRY, 2010; ZOU; HANSEN, 2012) e, eventualmente, prever quando e onde acontecerão e as suas durações (LU; WANG; ZHENG, 2008; TU; BALL; JANK, 2008; REBOLLO; BALAKRISHNAN, 2014);

(iii) Análise de cancelamentos: em situações específicas, atrasos podem se transformar em cancelamentos. Sendo assim, este tipo de problema analisa as causas de cancelamentos e o processo de tomada de decisão de companhias aéreas para a escolha dos voos a serem cancelados (BOSWELL; EVANS, 1997; HANSEN, 2002).

II.2 Escopo

A definição do escopo independe do tipo de problema e refere-se ao ambiente de análise de atrasos e cancelamentos. Em um sistema real de transporte aéreo, atrasos e cancelamentos podem ser motivados por diversas causas e ter diferentes fontes, afetando tanto companhias aéreas, como aeroportos e espaço aéreo. Entretanto, dependendo do estudo, uma simplificação do sistema pode ser interessante, restringindo o ambiente de análise a somente companhias aéreas, aeroportos ou espaço aéreo, ou ainda uma combinação destes.

Pesquisas para a previsão de atrasos em decolagens de um determinado aeroporto podem considerar apenas o próprio aeroporto, tratando de maneira indiferente todas as companhias aéreas e não avaliando o espaço aéreo (SCHAEFER; MILLNER, 2001; REBOLLO; BALAKRISHNAN, 2014). Por outro lado, estudos com a finalidade de comparar o desempenho de companhias aéreas considerando as mesmas condições operacionais de aeroportos e espaço aéreo restringem seus escopos às companhias aéreas (AHMADBEYGI et al., 2008). Existe ainda a possibilidade de um escopo combinado, como por exemplo, a avaliação de impactos de investimentos nos níveis de atraso considerando o congestionamento de aeroportos e do espaço aéreo (MORRISON; WINSTON, 2008) ou ainda a combinação de companhias aéreas, aeroportos e espaço aéreo por meio de um modelo para a estimativa das distribuições de atrasos em rota, nas partidas e nas chegadas com base em características das companhias aéreas e das aeronaves, condições meteorológicas e capacidade e demanda de aeroportos e espaço aéreo (MUELLER; CHATTERJI, 2002).

II.3 Dados

As atividades relacionadas aos dados envolvem coleta, definição das dimensões e atributos e gestão dos dados de modo a facilitar e possibilitar a implementação de um modelo, além de melhorar seus resultados. As bases de dados relacionadas aos sistemas aéreos de transporte são,

em geral, disponibilizadas por órgãos governamentais, como por exemplo, *The Bureau of Transportation Statistics* (BTS, 2015) e *The Federal Aviation Administration* (FAA, 2015) ligados ao *The United States Department of Transportation* (BOSWELL; EVANS, 1997; MUELLER; CHATTERJI, 2002; XU et al., 2005; TU; BALL; JANK, 2008; PYRGIOTIS; MALONE; ODONI, 2013; REBOLLO; BALAKRISHNAN, 2014; DOT, 2015), companhias aéreas (BEATTY et al., 1998; ABDELGHANY et al., 2004; WU, 2005; AHMADBEYGI et al., 2008; WONG; TSAI, 2012), aeroportos (LU; WANG; ZHENG, 2008) e provedores de serviços privados (WIELAND, 1997; SCHAEFER; MILLNER, 2001; HANSEN, 2002). A escolha de uma ou mais bases de dados depende, frequentemente, da região em que o estudo está sendo realizado, do relacionamento entre o pesquisador e as fontes de dados e das dimensões a serem consideradas.

Desta forma, múltiplas bases de dados podem ser necessárias para a aplicação de um modelo. Por exemplo, dados meteorológicos são utilizados com frequência nos estudos de atrasos e cancelamentos aéreos. Assim, bases de dados como a do *National Oceanic and Atmospheric Administration* (NOAA, 2015) ou do *Weather Company* (TWC, 2015) podem ser associadas aos registros de voos (ABDEL-ATY et al., 2007; HUNTER; BOISVERT; RAMAMOORTHY, 2007; HUNTER; RAMAMOORTHY, 2007; HAO et al., 2014) de modo a incorporar novas dimensões ao problema.

Por sua vez, as dimensões correspondem às classes dos atributos que irão compor o *dataset* do modelo de análise de atrasos e cancelamentos aéreos. Além dos horários previstos e realizados de partida e chegada, diferentes atributos podem ser considerados dependendo do foco da pesquisa. Assim, a Figura II.3 apresenta as sete classes de dimensões verificadas na literatura e alguns exemplos de seus atributos.

A dimensão espacial refere-se às posições assumidas pela aeronave, como, por exemplo, aeroportos de origem e destino ou suas cidades e regiões (HAO et al., 2014; REBOLLO; BALAKRISHNAN, 2014). Já a dimensão temporal inclui características associadas às datas, como estação do ano, mês e dia da semana e aos horários, como período do dia (MUELLER; CHATTERJI, 2002; ABDEL-ATY et al., 2007; TU; BALL; JANK, 2008). Frequentemente, os atributos desta dimensão necessitam de transformações para possibilitar a aplicação de modelos que procuram capturar sazonalidade e padrões periódicos nos dados. Para expressar condições externas e ambientais em um dado momento, usa-se a dimensão meteorológica, que pode incluir diferentes parâmetros, como teto e visibilidade (REYNOLDS-FEIGHAN; BUTTON, 1999), importantes para as operações de decolagem e pouso ou retratar as condições climáticas de um determinado aeroporto ou rota (ABDEL-ATY et al., 2007; HUNTER; BOISVERT; RAMAMOORTHY, 2007; HUNTER; RAMAMOORTHY, 2007).

A dimensão de características expressa particularidades de companhias aéreas, aeroportos

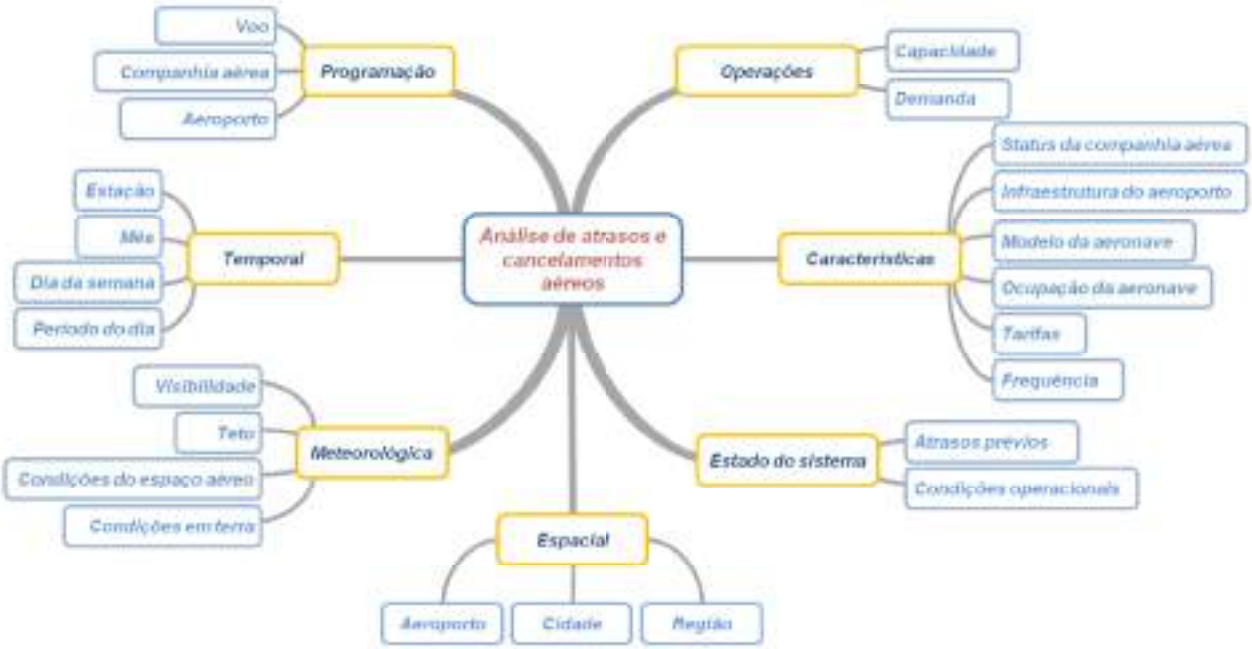


Figura II.3: Dimensões dos dados de voos

ou aeronaves, como o status de uma companhia aérea em relação à sua presença nacional e ao seu modo de operação, o tamanho, número de assentos e ocupação de aeronaves e a infraestrutura de aeroportos, como número de pistas, portões e disponibilidade de serviços de solo (PATHOMSIRI et al., 2008; SOOMER; FRANX, 2008; XIONG; HANSEN, 2013). Por outro lado, a programação descreve o que companhias aéreas, aeroportos e controladores de tráfego aéreo pretendem fazer com os recursos críticos das operações. Assim, companhias aéreas definem seus pontos de origem e destino, frequência e a alocação de aeronaves e tripulantes (BOSWELL; EVANS, 1997; BEATTY et al., 1998; WU, 2005; AHMADBEYGI et al., 2008; DüCK et al., 2012). As programações dos aeroportos indicam os horários de pouso e decolagem, enquanto os planos de voo aprovados pelo controle aéreo informam os parâmetros de rota, como distância, rota, velocidade e altura (HANSEN, 2002).

Finalmente, a dimensão operacional indica a capacidade e demanda gerais de aeroportos e espaço aéreo (MORRISON; WINSTON, 2008), pois quando a demanda excede a capacidade um cenário de congestionamento é formado, permitindo a ocorrência de atrasos. Diferentemente, a dimensão do estado do sistema reflete em que condições as companhias aéreas, os aeroportos ou o espaço aéreo estão operando em um momento específico. Um exemplo é o nível de atrasos já registrado ao longo do dia (LU; WANG; XU, 2009), que pode ser útil para a previsão do comportamento futuro do sistema.

A gestão dos dados pode ser vista como uma fase de pré-processamento ou de preparação dos dados para a implementação de um modelo. Frequentemente, esta etapa está associada às

atividades de integração de múltiplas fontes de dados, limpeza, seleção e transformação dos dados. A limpeza pode incluir a remoção de *outliers* (TU; BALL; JANK, 2008) e de atributos correlacionados (WONG; TSAI, 2012). A seleção pode se referir tanto à escolha de atributos específicos como à escolha de determinados aeroportos ou companhias aéreas (MUELLER; CHATTERJI, 2002; ABDEL-ATY et al., 2007; PATHOMSIRI et al., 2008; PYRGIOTIS; MALONE; ODONI, 2013; HAO et al., 2014; REBOLLO; BALAKRISHNAN, 2014). Finalmente, a transformação procura adaptar os dados de modo a permitir uma aplicação mais eficiente do modelo escolhido. Entre as possíveis abordagens para transformação, destacam-se a normalização (LU; WANG; XU, 2009) e a discretização de atributos, especialmente por meio da utilização da técnica de *binning* (BOSWELL; EVANS, 1997; BEATTY et al., 1998; XU et al., 2005; BALAKRISHNA; GANESAN; SHERRY, 2010; KIM; HANSEN, 2013).

II.4 Métodos

Dependendo do objetivo da pesquisa, a análise de atrasos e cancelamentos aéreos pode ser modelada por meio de inúmeros métodos. Sendo assim, a escolha adequada de um método e seu alinhamento com as demais atividades da taxonomia são fundamentais para a obtenção de bons resultados. De forma geral, é possível agrupar os métodos já aplicados às análises de atrasos e cancelamentos aéreos em cinco grandes categorias:

(i) Análise estatística: inclui modelos econométricos, testes não paramétricos, modelos de regressão e análises de correlação. Seus principais objetivos são a avaliação de novos atrasos no sistema como um todo ou em aeroportos específicos, a análise dos efeitos da propagação de atraso em companhias aéreas ou entre os aeroportos e o entendimento do processo de tomada de decisão de companhias aéreas para o cancelamento de voos. Assim, considerando a ocorrência de novos atrasos, Morrison e Winston (2008) construíram um modelo econométrico para avaliar a eficácia de investimentos de uma agência governamental norte-americana. Reynolds-Feighan e Button (1999) utilizaram estatística descritiva e correlações para compreender a relação entre atrasos em aeroportos europeus e suas capacidades, enquanto Pathomsiri et al. (2008) aplicaram testes não paramétricos para avaliar a eficiência de aeroportos dos Estados Unidos. O acontecimento de novos atrasos também pode ser analisado em um aeroporto específico, como realizado por Abdel-Aty et al. (2007), que verificaram as médias diárias de atrasos nas chegadas de um aeroporto dos Estados Unidos para obter padrões periódicos e, então, analisá-los a partir de regressões e correlações.

Beatty et al. (1998) propuseram um multiplicador de atrasos e um modelo de regressão linear para expressar os efeitos da propagação de atrasos nos voos subsequentes de uma determinada companhia aérea. Wang, Schaefer e Wojcik (2003) também utilizaram modelos de regressão,

mas para avaliar os efeitos da propagação de atrasos entre alguns aeroportos dos Estados Unidos, enquanto Hao et al. (2014) construíram um modelo econométrico para quantificar como os atrasos originados nos aeroportos de Nova Iorque se propagam para o restante da rede. Xiong e Hansen (2013) também elaboraram um modelo econométrico, porém com o objetivo de entender a decisão de companhias aéreas sobre a escolha dos voos a serem cancelados.

(ii) Modelos probabilísticos: incluem análise e estimativa de distribuições para a ocorrência de novos atrasos, funções de probabilidade e modelos de sobrevivência para a propagação de atrasos e probabilidades condicionais para a análise de cancelamentos. Mueller e Chatterji (2002) trabalharam com funções de densidade e concluíram que os atrasos nas partidas são bem representados pela distribuição normal, enquanto os atrasos na chegada ou em rota podem ser modelados pela distribuição de Poisson. Tu, Ball e Jank (2008) identificaram os principais fatores causadores de atrasos nas partidas para estimar sua distribuição em um aeroporto dos Estados Unidos.

Boswell e Evans (1997) expressaram os atrasos por meio de uma função de probabilidade e utilizaram uma matriz de transição para verificar a propagação de atrasos aos voos subsequentes de uma companhia aérea. Além disso, os autores verificaram a probabilidade condicional de um voo ser cancelado dado que seu voo precedente atrasou. Wong e Tsai (2012) construíram um modelo de sobrevivência para analisar a duração total de um atraso considerando sua propagação para a rede de voos de uma companhia aérea taiwanesa.

(iii) Representação de redes: envolve aplicações de teoria dos grafos e redes bayesianas. Seus objetivos, em geral, estão relacionados à análise da propagação de atrasos em diferentes companhias aéreas ou entre aeroportos. Abdelghany et al. (2004) construíram um grafo acíclico dirigido para modelar a programação de uma companhia aérea considerando os horários dos seus voos e a alocação de recursos críticos para detectar interrupções e seus impactos na rede. Os efeitos da propagação foram avaliados por meio do algoritmo clássico do caminho mínimo. AhmadBeygi et al. (2008) desenharam duas árvores para comparar a propagação de atrasos em duas companhias aéreas distintas, sendo uma delas tradicional operando em um esquema *hub-and-spoke* e a outra uma *low cost* operando em um sistema ponto a ponto. Finalmente, Xu et al. (2005) desenvolveram uma rede bayesiana para modelar a propagação de atrasos entre três grandes aeroportos dos Estados Unidos, restringindo dois deles a somente partidas e o outro a apenas chegadas.

(iv) Aprendizado de máquina: corresponde a métodos de classificação, predição e sistemas de recomendação. Entre os principais objetivos destacam-se classificações ou predições de novos atrasos e a análise de propagação de atrasos baseada em situações similares observadas no passado. Rebollo e Balakrishnan (2014) utilizaram o método *Random Forests* para classificar

as partidas dos aeroportos mais movimentados dos Estados Unidos quanto à ocorrência de novos atrasos. Lu, Wang e Zheng (2008) compararam os desempenhos de Naïve Bayes, árvores de decisão e redes neurais quanto à classificação para novos atrasos na China. Os melhores resultados foram obtidos para as árvores de decisão, com confiança de pelo menos 80%.

Balakrishna, Ganesan e Sherry (2010) aplicaram um algoritmo de aprendizagem por reforço para prever atrasos nos taxiamentos de partidas. O problema foi modelado por meio de um processo de Markov e atingiu bons resultados quando executado 15 minutos antes dos horários programados para partida no aeroporto de Tampa Bay, Estados Unidos. Finalmente, a propagação de atrasos entre aeroportos foi analisada por Lu, Wang e Xu (2009) a partir de um sistema de recomendação. Assim, com base em dados históricos, o sistema reconhece situações semelhantes no passado, indicando os atrasos que devem se propagar para os aeroportos relacionados.

(v) Pesquisa operacional: inclui basicamente modelos de simulação e de teoria de filas. Destacam-se aplicações relacionadas à ocorrência de novos atrasos em companhias aéreas ou sua análise em relação ao congestionamento de aeroportos ou do espaço aéreo e a avaliação da propagação de atrasos em companhias aéreas, aeroportos ou em ambos os casos. Wieland (1997) descreveu o DPAT, um modelo de simulação baseado em teoria das filas para prever novos atrasos relacionadas a problemas de congestionamento. Kim e Hansen (2013) também construíram um modelo de filas para simular os efeitos de capacidade e demanda nos níveis de atraso dos aeroportos da região de Nova Iorque. Ainda estudando congestionamentos, mas especialmente aqueles relacionados às condições meteorológicas, Hunter, Boisvert e Ramamoorthy (2007) e Hunter e Ramamoorthy (2007) desenvolveram uma plataforma para simular o *trade-off* envolvendo demanda e capacidade e indicar atrasos ou desvios para alguns voos. A relação entre congestionamento e atrasos foi verificada por meio de uma análise de sensibilidade para diferentes níveis de congestionamento.

Com o objetivo de minimizar os custos da ocorrência de novos atrasos em programações de companhias aéreas, Soomer e Franx (2008) formularam um modelo heurístico de busca local baseado no custo do atraso de cada voo. Ainda analisando as programações de companhias aéreas, mas sob o ponto de vista de propagação de atrasos, Wu (2005) e Dück et al. (2012) aplicaram modelos de simulação para verificação de sua confiabilidade e estabilidade. A propagação de atrasos em ou entre aeroportos foi estudada por Schaefer e Millner (2001) a partir de um modelo de simulação para capturar filas para decolagem ou pouso considerando diferentes condições de visibilidade de três aeroportos e por Pyrgiotis, Malone e Odoni (2013) para detectar o relacionamento entre aeroportos por meio de um modelo de filas atualizado constantemente com suas programações e níveis de demanda e capacidade. Hansen (2002) também desenvolveu um modelo de filas, mas avaliou tanto os efeitos de propagação de atrasos e cancelamentos sofridos

pelas companhias aéreas como pelo aeroporto internacional de Los Angeles nos Estados Unidos.

II.5 Considerações sobre análises de atrasos e cancelamentos aéreos

Devido à sua grande importância para companhias aéreas, aeroportos, investidores públicos e privados e passageiros, a análise de atrasos e cancelamentos aéreos tem sido um tema bastante discutido nos últimos anos. Considerando os principais trabalhos desta área, a Figura II.4 apresenta uma linha do tempo destas publicações destacando, conforme a classificação da taxonomia, o tipo de problema tratado, a região de aplicação e o grupo de métodos utilizado. Percebe-se um aumento no número de publicações a partir do final dos anos 2000, uma vez que 63% dos trabalhos foram publicados entre 2007 e 2014.

Analisando a linha do tempo quanto aos aspectos destacados na taxonomia, verifica-se um equilíbrio entre estudos de propagação e novos atrasos, enquanto poucos autores abordam a análise de cancelamentos. As múltiplas possibilidades de combinações para escopo e dimensões demonstram que há espaço para a busca de resultados diferentes e, eventualmente, melhores do que aqueles já discutidos. Observa-se ainda que os Estados Unidos aparecem como a região de aplicação da grande maioria das pesquisas, seguidos por Europa e Ásia. Apesar de ser um mercado importante de aviação comercial devido às suas características territoriais e de infraestrutura, o caso brasileiro não tem sido discutido em análises de atrasos e cancelamentos aéreos. Assim, o conhecimento sobre este assunto no Brasil ainda é, de certa forma, empírico, pouco se sabendo de fato sobre as condições que conduzem a atrasos ou cancelamentos no país. Finalmente, em relação à gestão de dados e aos métodos, percebe-se uma valorização de técnicas de pré-processamento e a predominância de análises estatísticas e de modelos de pesquisa operacional. No entanto, a introdução de modelagens de aprendizado de máquina nos últimos anos é um fator que merece destaque.

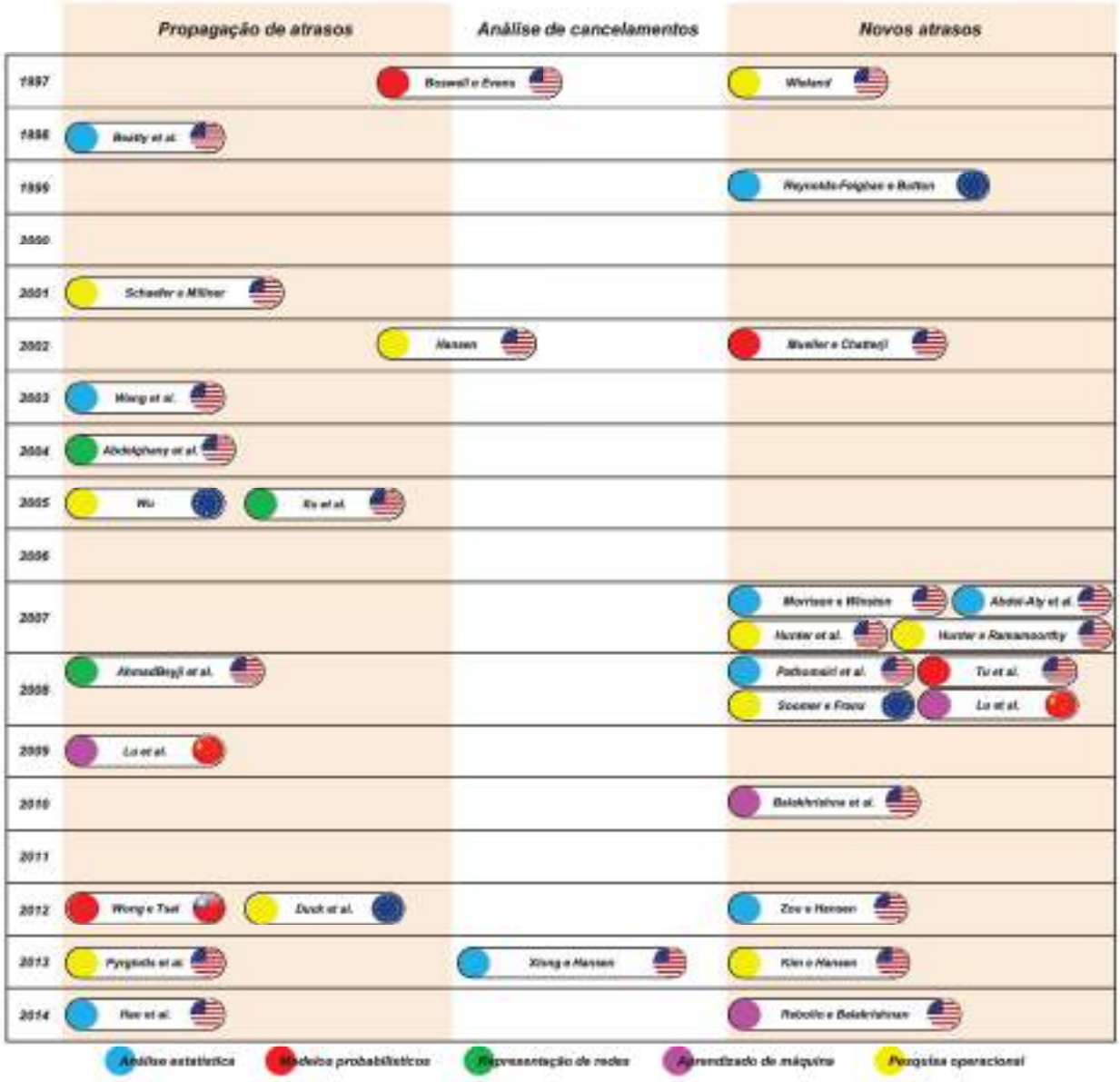


Figura II.4: Linha do tempo das principais publicações relacionadas à análise de atrasos e cancelamentos aéreos considerando seu tipo de problema, região de aplicação e método

Capítulo III Mineração de Padrões Frequentes

III.1 Processo de Mineração de Dados

A facilidade de captura e baixo custo de armazenagem das informações impulsionaram o crescimento das bases de dados em diversos setores. Assim, as análises para compreensão dos dados e suporte à decisão, tradicionalmente realizadas por especialistas, passaram a necessitar do auxílio de computadores. Surge, então, o processo de mineração de dados, uma sequência de etapas que culmina na descoberta de conhecimento válido, novo, potencialmente útil e que pode ser compreendido por seres humanos (MITRA; PAL; MITRA, 2002).

A Figura III.1 ilustra o processo de descoberta de conhecimento, que é comumente chamado apenas de mineração de dados. De uma maneira geral, inicia-se com o entendimento do domínio de aplicação e coleta dos dados, que podem estar disponíveis em diferentes bases de dados e em diferentes formatos. Os estágios seguintes de seleção, limpeza e transformação têm como objetivo a garantia da qualidade dos dados e podem ser agrupados em uma etapa de pré-processamento. O método de extração do conhecimento depende da análise a ser realizada, envolvendo a descoberta de padrões a partir de uma grande quantidade de dados. Estes padrões são avaliados por meio de medidas de interesse para depois serem apresentados sob a forma de conhecimento (HAN; KAMBER; PEI, 2011).

III.2 Integração

Em aplicações reais, muitas vezes, os dados estão presentes em diferentes fontes e em diferentes formatos. Isso ocorre, por exemplo, em problemas de aviação comercial, em que os dados operacionais dos voos costumam estar em bases de dados distintas daquelas contendo as condições meteorológicas. Sendo assim, para construir uma análise mais interessante e completa, torna-se necessário integrar estas diversas fontes, o que pode ser realizado a partir de um *data warehouse*, um ambiente não volátil, variante no tempo, orientado ao assunto a ser explorado e capaz de apresentar os dados de maneira integrada.

A construção de um *data warehouse* começa com a extração de dados das diferentes fontes identificadas para o problema. De modo a evitar redundâncias e inconsistências, é interessante tratar os dados extraídos antes de carregá-los no *data warehouse*. Assim, por exemplo, datas

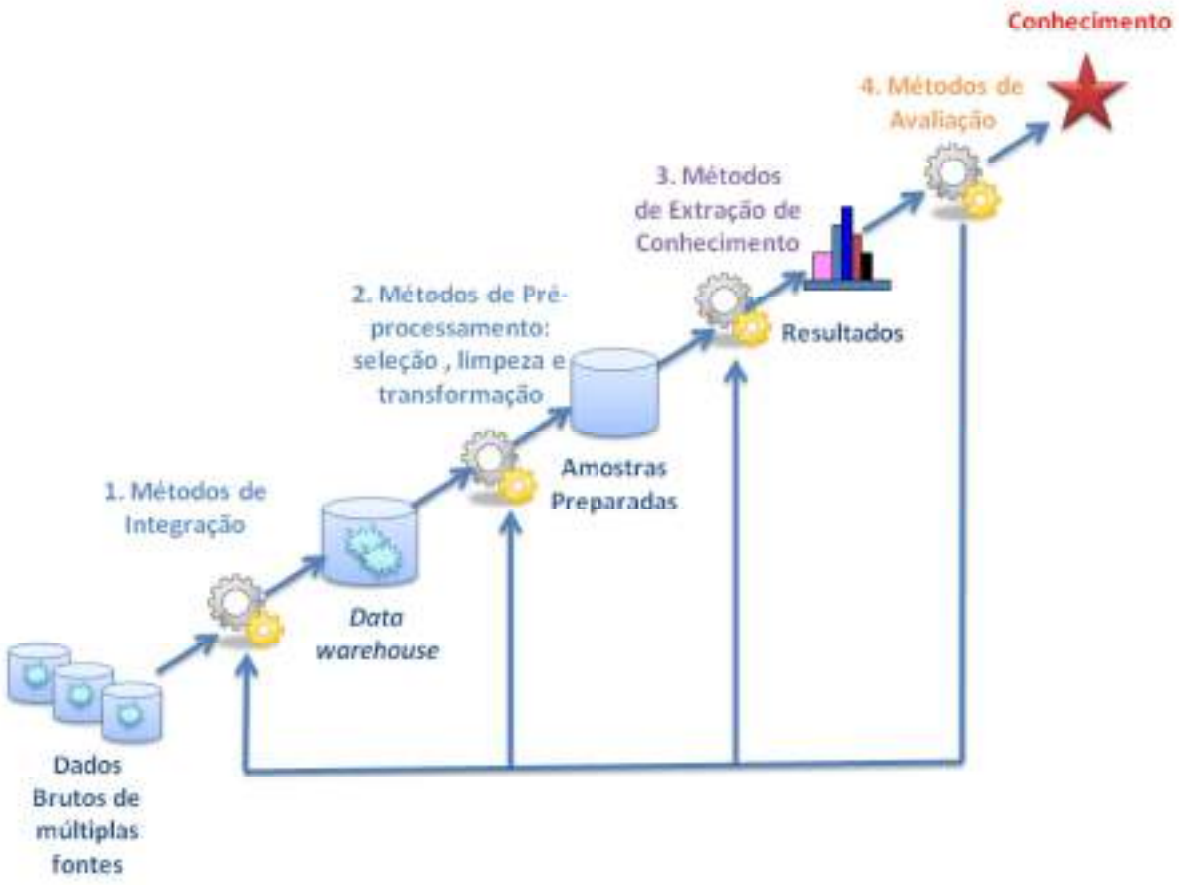


Figura III.1: Mineração de dados como um processo de descoberta de conhecimento. Fonte: adaptado de Han, Kamber e Pei (2011)

podem ser colocadas em um mesmo formato, atributos podem ser agrupados, derivados ou ignorados, ou ainda serem criadas algumas regras de validação para então ocorrer o carregamento dos dados no *data warehouse*. Este processo é chamado de Extração, Transformação e Carga - Extract, Transform, Load (ETL) e é comumente utilizado para a integração de múltiplas fontes de dados em um *data warehouse* (SILBERSCHATZ; KORTH; SUDARSHAN, 2010; HAN; KAMBER; PEI, 2011).

A Figura III.2 ilustra o processo de ETL considerando um exemplo para a construção de um *data warehouse* de voos nacionais a partir da extração de dados de duas bases, uma de voos e outra de condições meteorológicas. Após a extração, os voos internacionais da base Voo são ignorados, enquanto os dados de data e hora da base Meteorológica são agrupados em um único atributo e têm seu formato alterado para coincidir com o formato de data e horário da base Voo. Finalmente os dados são carregados no *data warehouse* podendo ser integrados a partir das informações de aeroporto e data e horário.

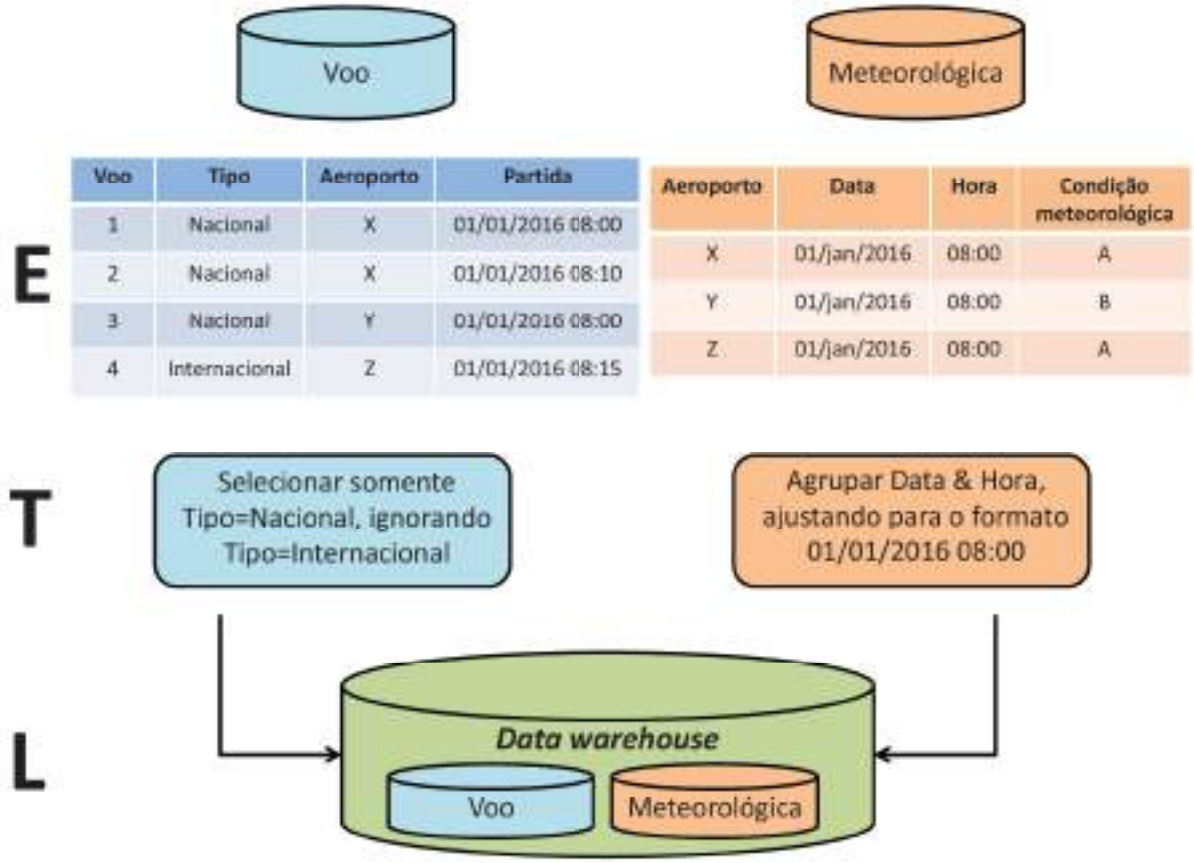


Figura III.2: Exemplo do processo de construção de um *data warehouse* de voos nacionais

III.3 Seleção

Um *data warehouse* pode ter uma grande quantidade de atributos e dados, alguns até mesmo redundantes ou com pouca relevância para determinadas análises. Sendo assim, pode ser incluída uma etapa de seleção de modo a considerar apenas parte destes atributos ou dados para a aplicação de um método de extração de conhecimento. O objetivo desta redução é deixar o modelo mais eficiente, sem que haja perdas significativas nos resultados (HAN; KAMBER; PEI, 2011).

Considerando a redução de tuplas, podem ser utilizadas técnicas de amostragem, por exemplo. Entretanto, em problemas de aviação comercial, frequentemente os *datasets* têm seu número de tuplas reduzido a partir da seleção dos principais aeroportos ou companhias aéreas (MUELLER; CHATTERJI, 2002; XIONG; HANSEN, 2013; REBOLLO; BALAKRISHNAN, 2014). Dentre outros métodos, a redução de tuplas pode ocorrer por amostragens aleatórias, estratificadas, quartis, histogramas e Pareto.

III.4 Limpeza

Frequentemente, as bases de dados reais sofrem com a presença de ruídos devido a erros de digitação ou na transmissão dos dados, por exemplo. Uma vez que se deseja extrair conhecimento a partir dos dados, a garantia de sua qualidade é essencial para a obtenção de bons resultados. Cabe destacar que o processo de entendimento do domínio de aplicação realizado previamente é fundamental para o sucesso desta etapa, pois a compreensão prévia auxilia na correta identificação dos atributos e dados a serem analisados e o tipo de tratamento necessário para cada um deles (HAN; KAMBER; PEI, 2011).

A primeira etapa de um processo de limpeza é a detecção de discrepâncias, ou seja, valores não permitidos pelo domínio (HAN; KAMBER; PEI, 2011). Isso pode incluir tanto regras universais como específicas do problema. Um exemplo de verificação de uma regra universal seria a validação de um atributo que representa o percentual de atrasos aéreos e que, portanto, só pode assumir valores entre 0 e 1. Por outro lado, uma regra específica envolve o contexto em que o problema está sendo aplicado. Por exemplo, a duração máxima de um atraso pode ter limites diferentes em países diferentes.

Diferentemente das discrepâncias, os *outliers* são valores permitidos pelo domínio, mas muito distantes do valor esperado. Uma técnica de estatística descritiva comumente utilizada para a identificação de *outliers* é a avaliação por meio de *boxplot*. Neste caso, devem ser calculados o primeiro e o terceiro quartis, *Q1* e *Q3*, respectivamente. Em uma análise conservadora, são considerados *outliers* os valores inferiores a $Q1 - 3 * IQR$ ou superiores a $Q3 + 3 * IQR$, sendo *IQR* a diferença entre *Q3* e *Q1*, ou a amplitude inter-quartil.

A diferença entre discrepâncias e *outliers* pode ser ilustrada pelo exemplo fictício apresentado a seguir. Supondo que, para a realização de uma avaliação geral, o gerente de um aeroporto tenha solicitado a medição dos tempos de todos os atrasos ocorridos durante um dia. O relatório produzido está representado na Tabela III.1. Se o tempo máximo de atraso permitido pela autoridade reguladora é de 24 horas, o voo 5 possui uma discrepância, uma vez que seu atraso é de 1500 minutos. Sendo assim, este voo foi removido da análise.

Todos os outros voos possuem atrasos abaixo do limite estipulado pela autoridade reguladora, entretanto, conforme pode ser visto no *boxplot* ilustrado na Figura III.3, os voos 10 e 16 possuem tempos muito superiores aos demais. Estes pontos estão, inclusive, fora do intervalo de $Q1 - 3 * IQR$ a $Q3 + 3 * IQR$, uma vez que *Q1* e *Q3* são iguais a 5,5 e 26, respectivamente, de modo que $Q1 - 3 * IQR = -56$ e $Q3 + 3 * IQR = 87,5$. Desta forma, estes voos foram considerados *outliers* oriundos, na avaliação do gerente, de erros durante as atividades de anotação ou de digitação ou de situações muito específicas e sem interesse para a sua avaliação geral.

Tabela III.1: Exemplo de detecção de discrepâncias e *outliers*

Voo	Tempo de atraso (minutos)
1	5
2	20
3	17
4	3
5	1500
6	8
7	14
8	4
9	6
10	500
11	12
12	39
13	2
14	11
15	5
16	160
17	9
18	61
19	15
20	32

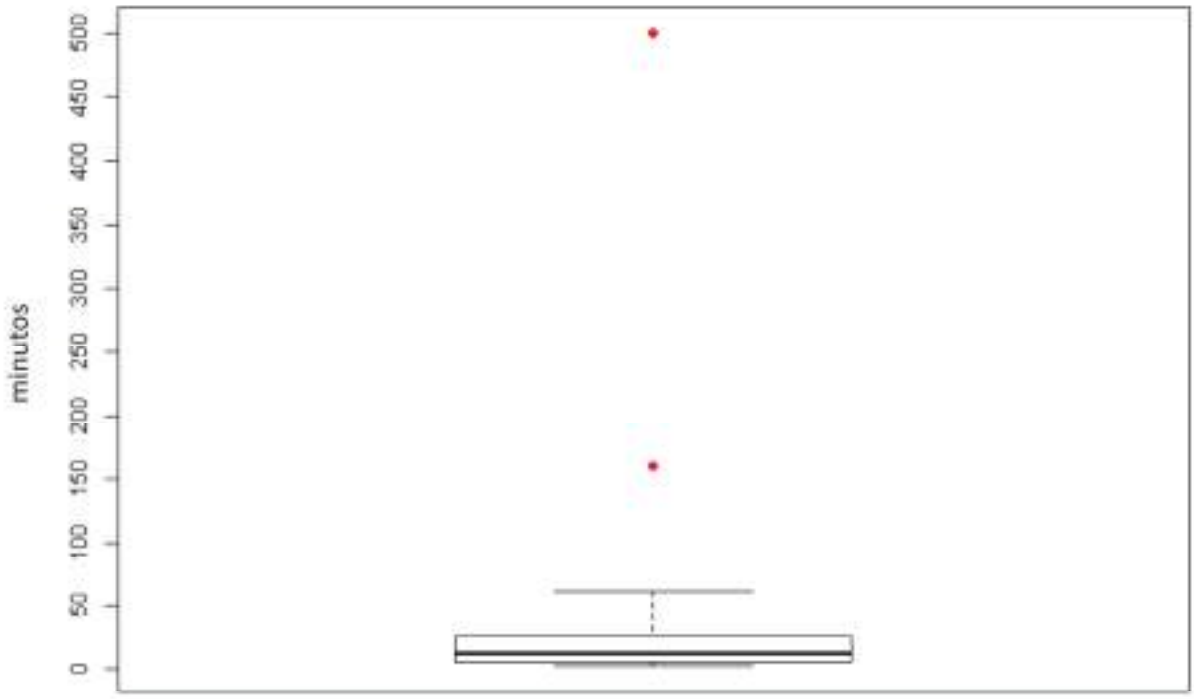


Figura III.3: Exemplo de *boxplot* para a identificação de *outliers*

III.5 Transformação

A preparação dos dados para a aplicação de um método de extração de conhecimento inclui ainda uma etapa de transformação, na qual os dados são convertidos para um formato apropri-

ado. Além de ser fundamental para possibilitar a implementação de um modelo, a transformação também o torna mais eficiente e com melhores resultados. Algumas técnicas de transformação incluem discretizações por *binning* ou hierarquia conceitual e agregações (HAN; KAMBER; PEI, 2011).

Na discretização por *binning*, o domínio é dividido em intervalos chamados de *bins*, de modo que cada valor pertença a apenas um *bin* e, desta forma, possa ser representado pelo rótulo de seu *bin*. Neste caso, a qualidade do *binning* é sensível ao número de *bins*, que pode ser previamente definido ou avaliado a partir da distribuição dos dados (HAN; KAMBER; PEI, 2011). Uma maneira de determinar este número é a avaliação dos erros em relação às médias a partir da utilização da máxima curvatura, o que será discutido na Seção III.7.

Por outro lado, a discretização por hierarquia conceitual substitui dados de níveis detalhados por conceitos de níveis superiores (HAN; KAMBER; PEI, 2011). Por exemplo, de um atributo contendo uma data podem ser derivados o mês e o ano, ou seja, 01/01/2016 poderia ser representado por janeiro ou 2016 dependendo do foco da análise. Particularmente em problemas de aviação, também é possível utilizar escalas amplamente conhecidas para a discretização de atributos numéricos. Um exemplo são os cinco códigos utilizados no Informe Meteorológico Aeronáutico Regular - Meteorological Aerodrome Report (METAR) para a indicação da nebulosidade, que é usualmente medida em oitavos ou décimos do céu.

Em análises envolvendo séries temporais, a transformação para níveis superiores pode ser realizada por meio de agregação temporal (TIAO, 1972). Por exemplo, uma série temporal com observações a cada minuto pode ser agregada para uma série com observações a cada hora. Assim, sendo x_t uma série dos voos atrasados em um aeroporto a cada minuto, é possível agregá-la em uma série y_t, representando o percentual de voos atrasados neste aeroporto a cada hora.

III.6 Padrões Frequentes

Após a conclusão da etapa de pré-processamento, o *dataset* está preparado para a aplicação de um método de extração de conhecimento. Dentre estes métodos, destacam-se a mineração de padrões frequentes, agrupamento, classificação e predição. É importante que a escolha do método leve em consideração os objetivos da análise (HAN; KAMBER; PEI, 2011).

A mineração de padrões frequentes é um método de extração de conhecimento que, como o próprio nome indica, busca encontrar os padrões que se repetem nos dados. Estes padrões podem ser, por exemplo, conjuntos de dados ou sequências e a sua descoberta pode resultar em conhecimento novo e útil sobre associações ou correlações existentes entre os dados. Existem diferentes técnicas para a mineração de padrões frequentes, sendo as mais difundias Apriori, FP-growth e Eclat (HAN; KAMBER; PEI, 2011).

Proposto por Agrawal e Srikant (1994), o algoritmo Apriori procura conjuntos de itens frequentes baseado no princípio de que um conjunto de itens somente será frequente se todos os seus subconjuntos também forem frequentes. Desta forma, ao pesquisar o *dataset* em busca dos conjuntos frequentes de tamanho 1, geram-se os candidatos de tamanho 2 e assim sucessivamente. É importante destacar que a geração de novos conjuntos candidatos implica em uma nova pesquisa de itens frequentes no *dataset*.

Para evitar a geração de um número muito grande de conjuntos candidatos e de pesquisas no *dataset*, Han, Pei e Yin (2000) apresentaram o algoritmo FP-growth, que comprime o dataset em uma árvore de padrões frequentes. A árvore cresce com base em fragmentos oriundos de uma lista decrescente de padrões, não necessitando, portanto, da geração de candidatos. Ao invés de realizar a pesquisa de maneira horizontal como os algoritmos Apriori e FP-Growth, o Eclat introduzido por Zaki et al. (1997) transforma o *dataset* em um formato vertical de dados e aplica o mesmo princípio utilizado pelo Apriori.

De modo a compreender o relacionamento entre os diferentes itens frequentes do *dataset*, podem ser geradas regras de associação no formato $X \rightarrow Y$, ou seja, a ocorrência de um conjunto de itens frequentes $X = \{x_1, x_2, ..., x_n\}$ implica na ocorrência de um outro conjunto de itens frequentes $Y = \{y_1, y_2, ..., y_n\}$. Para que uma regra seja considerada interessante, é necessário que algumas condições sejam satisfeitas, principalmente suporte e confiança mínimos. O suporte de uma regra é definido como o percentual de registros do *dataset* que contem todos os elementos de X e de Y. Considerando apenas os registros do *dataset* que contem todos os elementos de X, a confiança mede o percentual destes registros que também contem todos os elementos de Y, ou seja, representa a probabilidade da ocorrência de Y dado que X já ocorreu (ELMASRI; NAVATHE, 2015).

Mesmo com filtros de suporte e confiança, o número de regras geradas ainda pode ser muito grande, especialmente em problemas em que o suporte e a confiança são baixos. A geração de um conjunto compacto de regras úteis continua sendo um grande desafio na área de mineração de padrões frequentes. Neste contexto, algumas técnicas foram desenvolvidas com o objetivo de facilitar esta análise, como a utilização de uma medida de correlação chamada *lift* (HAN et al., 2007). Utilizando a regra de associação $X \rightarrow Y$, o *lift* pode ser calculado a partir da Equação III.1 e representa a correlação entre X e Y, indicando se X e Y são negativamente correlacionados, independentes ou positivamente correlacionados quando seu resultado é menor, igual ou maior que 1, respectivamente (HAN; KAMBER; PEI, 2011).

$$Lift = \frac{Confianca(X \rightarrow Y)}{Suporte(Y)} \tag{III.1}$$

.

Para ilustrar os conceitos de conjuntos frequentes, regras de associação, suporte, confiança e *lift*, apresenta-se o exemplo fictício de um gerente de aeroporto que está estudando os motivos dos atrasos de voos. Para isso, os dados de uma manhã de operações foram coletados, conforme representados na Tabela III.2. Considerando um suporte mínimo de 50%, encontram-se os itens frequentes apresentados na Tabela III.3. Neste caso, um exemplo de regra de associação é $\{Horario\ de\ partida = Inicio\ da\ manha\} \rightarrow \{Atraso = Verdadeiro\}$, ou seja, partidas no início da manhã conduzem a atrasos com um suporte de 50%, confiança de 66,6% e *lift* de 1,33, o que significa que partidas no início da manhã e atrasos estão positivamente correlacionados. O valor do *lift* também indica que quando o voo parte no início da manhã, há 33% mais chances de ocorrer um atraso.

Tabela III.2: Exemplo de um *dataset* representando as operações do turno da manhã de um determinado aeroporto

Voo	Companhia aérea	Horário da partida	Atraso
1	A	Início da manhã	Verdadeiro
2	B	Início da manhã	Verdadeiro
3	C	Início da manhã	Falso
4	A	Final da manhã	Falso

Tabela III.3: Examplo de padrões frequentes para um suporte de 50%

Padrão frequente	Suporte
{Horário da partida=Início da manhã}	75%
{Companhia aérea=A}	50%
{Atraso=Verdadeiro}	50%
{Atraso=Falso}	50%
{Horário da partida=Início da manhã, Atraso=Verdadeiro}	50%

III.7 Métodos de avaliação

Em problemas envolvendo a utilização de regras de associação, primeiramente é importante definir os valores mínimos para suporte e confiança. No entanto, conforme visto na Seção III.6, mesmo satisfazendo estas condições, uma regra pode não ser interessante. Sendo assim, é importante considerar também uma medida de correlação, como o *lift*. A avaliação do *lift* permite diferenciar atributos negativamente correlacionados, independentes ou positivamente correlacionados (HAN; KAMBER; PEI, 2011). Quando o atraso está no RHS da regra e há interesse em descobrir suas causas, se o *lift* é maior do que 1, maiores são as chances das condições do LHS conduzirem a novos atrasos. Isto é, nesta abordagem, quanto maior o *lift* mais interessante é a regra. A análise oposta, ou seja, para *lifts* menores do que 1, pode indicar as condições que tendem a produzir voos pontuais.

Por exemplo, supondo que o aeroporto apresentado no exemplo da Seção III.6 tenha realizado uma análise mais completa incluindo outros atributos. Uma vez que o gerente está interessado nas condições que conduzem a novos atrasos, ao obter uma regra de associação com *lift* superior a 1,33, ele pode concluir que os atributos do LHS têm mais chances de produzir atrasos do que as partidas no início da manhã. Da mesma maneira, atributos de regras com *lifts* entre 1 e 1,33 podem ser consideradas menos interessantes do que as partidas no início da manhã.

Adicionalmente, para auxiliar na definição dos parâmetros envolvidos na geração de regras de associação ou na etapa de pré-processamento dos dados, podem ser utilizados alguns métodos de análise numérica. Uma abordagem possível é a geração de uma *spline* a partir da interpolação de pontos conhecidos para a verificação da máxima curvatura. Assim, o ponto em que a curvatura é máxima indica o valor ótimo para a aplicação (BURDEN; FAIRES; BURDEN, 2015).

Assim como em um ciclo de melhoria contínua, os resultados devem ser constantemente avaliados para garantir a produção de conhecimento novo e útil. É importante definir e revisar periodicamente as medidas de interesse para analisar os resultados e, eventualmente, filtrá-los. Cabe destacar que o conhecimento somente será gerado se os resultados forem apresentados de forma compreensível aos seres humanos e se puderem ser aplicados em processos estratégicos e decisórios (HAN; KAMBER; PEI, 2011).

Capítulo IV Metodologia

A análise de padrões frequentes nos dados de voos brasileiros possui como objetivo principal a construção de um panorama geral para a obtenção das principais condições que conduzem a novos atrasos nas partidas ou a cancelamentos. Desta forma, avaliando este estudo sob a ótica de ciência de dados e considerando a taxonomia de atrasos e cancelamentos aéreos apresentada no Capítulo II, é possível classificá-lo como um problema de novos atrasos e análise de cancelamentos. Como os atrasos e cancelamentos são avaliados no âmbito de companhias aéreas e aeroportos e retratam apenas ocorrências registradas nas partidas dos voos, o espaço aéreo não é considerado neste trabalho, ficando o escopo limitado a apenas companhias aéreas e aeroportos.

Os dados para a modelagem do problema foram obtidos por meio de bases de dados públicas providas por instituições governamentais e possuem dimensões espacial, temporal, meteorológica, características e do estado do sistema. Com a finalidade de possibilitar a aplicação de um modelo e de melhorar seus resultados, torna-se necessária a execução de etapas de gestão de dados, como integração de múltiplas fontes, limpeza, seleção e transformação. Finalmente, o problema é tratado como um processo de mineração de dados, uma vez que se deseja extrair conhecimento novo e útil, possível de ser usado para a tomada de decisões de passageiros, companhias aéreas, aeroportos e investidores, sendo a mineração de padrões frequentes por meio de regras de associação o método de aprendizado de máquina escolhido para a realização desta análise.

A metodologia para análise de atrasos e cancelamentos nos voos brasileiros é composta por cinco etapas (Integração, Seleção, Limpeza, Transformação e Geração de Regras de Associação), conforme apresentado na Figura IV.1. As próximas seções detalham as atividades executadas em cada um destes estágios, seguindo o fluxo de um processo tradicional de ciência de dados. É importante destacar que atrasos nas partidas e cancelamentos são tratados igualmente no restante deste estudo e, portanto, ambos serão chamados apenas de atrasos nas partidas, ou simplesmente atrasos.

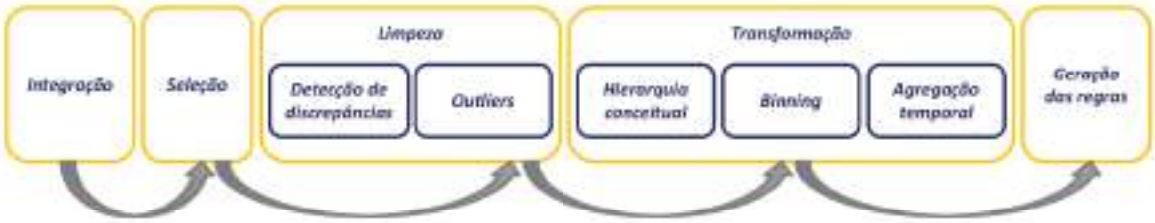


Figura IV.1: Processo de ciência de dados para analise de atrasos nos voos brasileiros

IV.1 Integração

No Brasil, a Agência Nacional de Aviação Civil (ANAC) é o órgão responsável pela regulação e fiscalização das atividades de aviação civil, bem como da infraestrutura aeronáutica e aeroportuária. Sua criação ocorreu em 2005 em substituição ao Departamento de Aviação Civil (DAC) dentro de um contexto nacional de reformulação político-econômica, que buscava retirar dos Ministérios as funções de fiscalização e regulação de serviços públicos exercidos por instituições privadas. por meio de Instrução de Aviação Civil (IAC), o DAC e, posteriormente, a ANAC, define regras e procedimentos que devem ser seguidos pelas empresas de transporte aéreo atuantes no Brasil (ANAC, 2015a).

A IAC 1223 (DAC, 2000a) estabelece os procedimentos para concessão de Horários de Transporte (HOTRAN), ou seja, a aprovação recebida por uma companhia aérea para explorar regularmente uma determinada linha, em horários e frequências específicos, com números de voos, tipos de aeronave e oferta de assentos previamente definidos. Conforme IAC 1504 (DAC, 2000b), qualquer diferença entre a operação efetivamente realizada e aquela prevista no HOTRAN deve ser informada pela companhia aérea à ANAC por meio de um Boletim de Alteração de Voo (BAV). Desta forma, a partir do cruzamento de HOTRAN e BAV foi criada pela ANAC a base de dados Voo Regular Ativo (VRA) contendo a relação de voos previstos em um determinado período e informações sobre sua execução (ANAC, 2015a).

A ANAC disponibiliza mensalmente a toda a comunidade os dados do VRA correspondentes aos voos com partidas previstas durante o mês em questão (ANAC, 2015a). Para cada operação são apresentados a companhia aérea, o número do voo, o tipo de linha (cargueiro, cargueiro internacional, especial, internacional, nacional, rede postal, regional ou sub-regional), o dígito identificador (regular, extra com HOTRAN, extra sem HOTRAN, retorno, inclusão de etapa, cargueiro, serviço, fretamento, *charter*, instrução ou experiência), aeroportos de origem e destino e horários previstos e reais de partida e de chegada. Como os dados são divulgados mensalmente, uma análise mais abrangente ao longo do tempo exige a consolidação destas publicações.

Com o objetivo de compreender as principais causas gerais de atrasos de voos brasileiros, neste estudo serão considerados somente os voos regulares e nacionais ou especiais, denomina-

dos a partir de agora apenas como voos, previstos para partirem entre janeiro de 2009 e fevereiro de 2015. Voos regulares são aqueles que ligam dois ou mais locais por meio de um serviço regular de acordo com horários, linhas, aeronaves e frequências previstos no HOTRAN, enquanto identificadores nacionais e especiais correspondem aos voos que ligam cidades com alto tráfego aéreo ou que atendam pelo menos um aeroporto coordenado, respectivamente (ANAC, 2015a). Durante o período considerado na análise, 4.476.660 voos estão classificados como regulares e nacionais ou especiais.

A escolha do período de análise deve-se às significativas mudanças ocorridas nos últimos anos no setor da aviação comercial brasileira. A abertura do mercado e o regime de livre concorrência instaurados em 2005 permitiram a entrada de novas companhias aéreas, resultando na saída de empresas tradicionais do setor. Sendo assim, o ano de 2009 marca o momento em que as quatro principais companhias aéreas atuais (Tam, Gol, Avianca e Azul) já operavam voos no país. O término em fevereiro de 2015 está relacionado apenas à disponibilidade de dados no momento do início deste trabalho.

Além da consolidação dos dados mensais do VRA, a integração com outras fontes de dados proporciona uma análise mais interessante e completa. Diante deste cenário, um *data warehouse* foi construído em PostgreSQL (PostgreSQL GLOBAL DEVELOPMENT GROUP, 2015), conforme apresentado no modelo conceitual da Figura IV.2. Os dados meteorológicos foram obtidos no Banco de Dados Meteorológicos para Ensino e Pesquisa (BDMEP) do Instituto Nacional de Meteorologia (INMET). Esta base de dados pública apresenta as séries históricas das observações meteorológicas de cada uma das estações de medição. Estas observações são realizadas três vezes por dia e medem seis atributos (temperatura, pressão, umidade relativa, direção do vento, velocidade do vento e nebulosidade) (INMET, 2015). Como as estações estão localizadas nas cidades e não nos aeroportos, foi necessária a construção de uma tabela para relacionar os aeroportos a suas cidades. Além disso, utilizou-se a observação meteorológica realizada no horário mais próximo à previsão de partida ou de chegada do voo para a integração dos dados meteorológicos aos dados do VRA. Para o acréscimo do atributo temporal relacionado à identificação de dias úteis nas partidas, foi construída uma tabela de feriados municipais, estaduais e nacionais, que também foi relacionada aos dados do VRA por meio da data prevista para partida e da cidade em que o aeroporto de origem está localizado.

IV.2 Seleção

A etapa de seleção foi dividida em dois estágios: (i) seleção de aeroportos e (ii) seleção de companhias aéreas. Devido às suas características geográficas e de infraestrutura, os aeroportos brasileiros possuem tamanhos e importâncias muito diferentes. Como este trabalho pretende

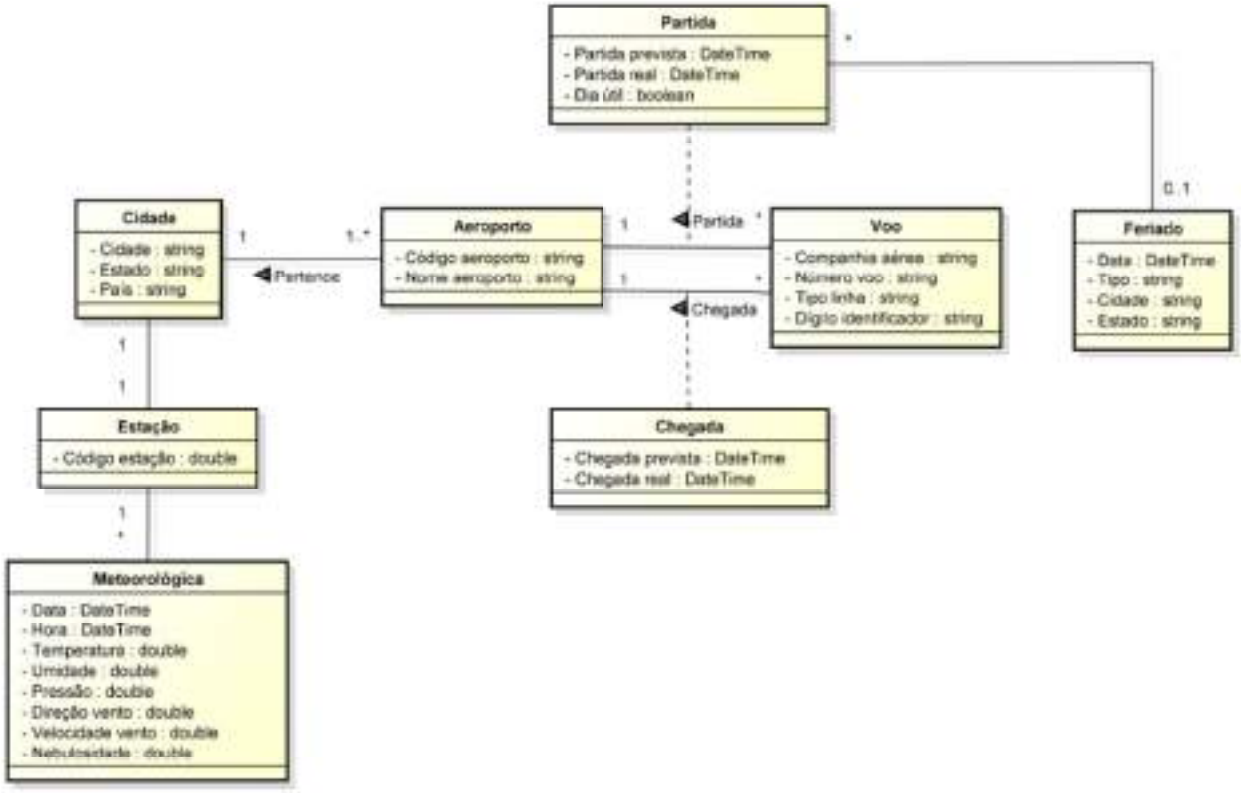


Figura IV.2: Diagrama de classes para integração de múltiplas fontes de dados

obter uma visão geral das principais causas de atrasos aéreos no Brasil, os principais aeroportos foram definidos de acordo com o número de partidas previstas entre janeiro de 2009 a fevereiro de 2015. Desta forma, conforme demonstrado na Tabela IV.1, com base no princípio de Pareto, foram selecionados os 17 principais aeroportos da rede de aviação comercial brasileira, responsáveis por 80% dos voos do período.

A seleção de companhias aéreas também é fundamental para a obtenção de regras de associação significativas e interessantes. No Brasil, existem apenas quatro companhias com aeronaves de grande porte e com importantes malhas aéreas e fatias de mercado. Considerando os voos entre janeiro de 2009 e fevereiro de 2015, Tam, Gol, Avianca e Azul foram responsáveis por 93% das partidas previstas. Sendo assim, apenas estas quatro companhias aéreas foram selecionadas para a análise de atrasos.

A Tabela IV.2 resume o número de registros resultantes após a execução de cada um dos estágios da etapa de seleção, enquanto que a Figura IV.3 ilustra a redução do número de tuplas do *dataset*. É importante destacar que a identificação dos 17 principais aeroportos foi realizada com base nos percentuais de partidas previstas considerando todos os voos do *dataset*, enquanto que

Tabela IV.1: Aeroportos selecionados para a análise

Código OACI	Aeroporto	Cidadde	Frequência acumulada
SBSP	Congonhas	São Paulo	10%
SBGR	Guarulhos - Governador Andre Franco 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	Deputado Luis Eduardo 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	Hercílio Luz	Florianópolis	76%
SBBE	Val De Cans	Belém	78%
SBGO	Santa Genoveva	Goiânia	79%
SBEG	Eduardo Gomes	Manaus	80%

para a seleção foram mantidos somente os voos com partidas e chegadas nestes 17 aeroportos. Assim, apesar dos 17 aeroportos representarem 80% do total de partidas previstas, como nem todos os voos ocorrem entre eles, o número de registros foi reduzido a 69% ao final do primeiro estágio de seleção.

Tabela IV.2: Número de registros resultantes após a etapa de seleção

Número inicial de registros	Número de registros após seleção de aeroportos	Número de registros após seleção de companhias aéreas
4.476.660	3.067.966	2.852.441

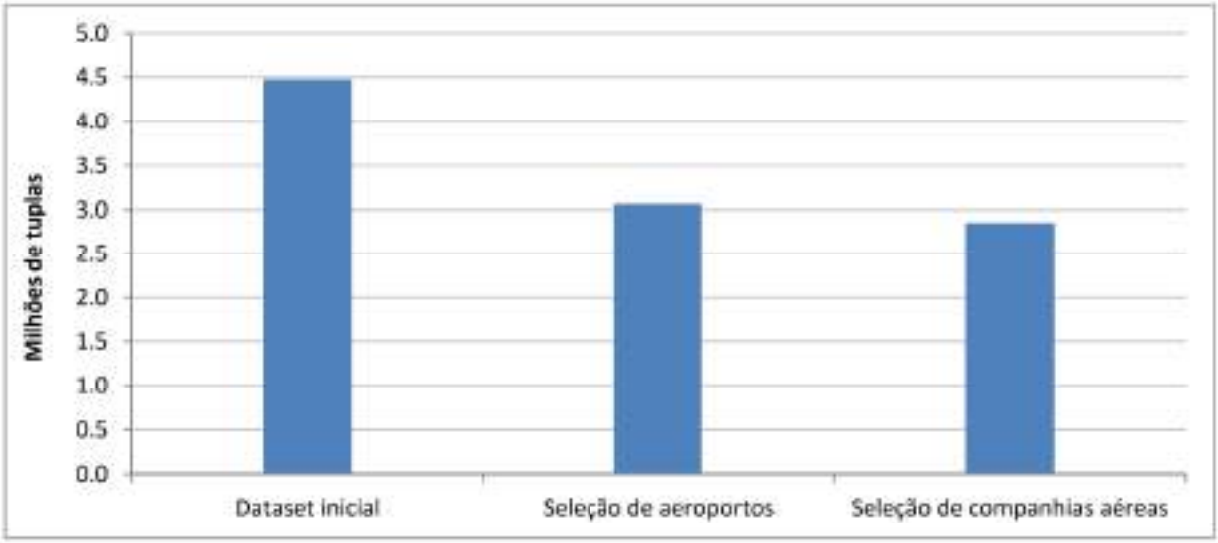


Figura IV.3: Redução do número de tuplas do *dataset* após a etapa de seleção

IV.3 Limpeza

A etapa de limpeza foi dividida em duas partes: (i) detecção de discrepâncias e (ii) tratamento de *outliers*. A IAC 1504 define a duração máxima de um atraso como 24 horas, momento a partir do qual o voo passa a ser considerado cancelado. Desta forma, voos realizados com atrasos superiores a 24 horas possuem um valor de atraso discrepante e, portanto, erros em seus horários previstos ou reais. Sendo assim, como existem erros nos registros, não é possível determinar se estes atrasos de fato ocorreram. De modo a não prejudicar a qualidade da análise de atrasos de voos nacionais, estas tuplas foram excluídas do *dataset*.

Diferentemente das discrepâncias, os *outliers* são valores permitidos pelo domínio, mas muito distantes das expectativas. No caso de voos comerciais, os *outliers* podem ter a mesma origem das discrepâncias, ou seja, erros nas informações dos horários previstos ou reais, ou ainda serem causados por situações extremamente específicas e que não se enquadram nas causas principais e gerais de atrasos. De modo a contemplar estas duas possibilidades, tanto as durações dos voos como as diferenças entre horários previstos e reais foram analisadas por meio de *boxplot*. Foram considerados como *outliers* os valores menores que $Q1 - 3 * IQR$ ou maiores que $Q3 + 3 * IQR$, sendo *Q1* o primeiro quartil, *Q3* o terceiro quartil e *IQR* a diferença entre eles, ou amplitude inter-quartil.

A análise das durações dos voos deve ser executada para cada rota individualmente e pode indicar possíveis erros nas informações de horários previstos ou reais, como por exemplo, equívocos devido à alteração para o horário de verão ou diferenças de fusos horários entre aeroportos. Assim como para as discrepâncias, estes erros podem ocultar ou indicar indevidamente o acontecimento de um atraso e, por isso, foram removidos do *dataset*.

O exemplo da Figura IV.4 ilustra a verificação de *outliers* na duração dos voos entre Guarulhos (São Paulo) e Galeão (Rio de Janeiro). Para esta rota, *Q1* e *Q3* são iguais a 1 hora e 1 hora e 6 minutos, respectivamente. Assim, os voos com durações inferiores a 42 minutos ou superiores a 1 hora e 24 minutos são considerados *outliers*. Este é o caso de 941 voos de um total de 22.422, que possuem durações previstas de 1 hora e 25 minutos, 1 hora e 27 minutos e 1 hora e 35 minutos. A título de comparação, a duração média prevista para os voos desta rota é de 1 hora e 4 minutos.

Por outro lado, as diferenças entre horários previstos e reais podem ser avaliadas de maneira global para todos os voos. Os *outliers* encontrados indicam atrasos ou adiantamentos excessivos em relação ao esperado para os voos brasileiros. Uma vez que estes valores podem representar erros nos horários previstos ou reais ou situações extremas não interessantes para a construção de um panorama geral de atrasos, estes *outliers* também foram removidos do *dataset*. Isso

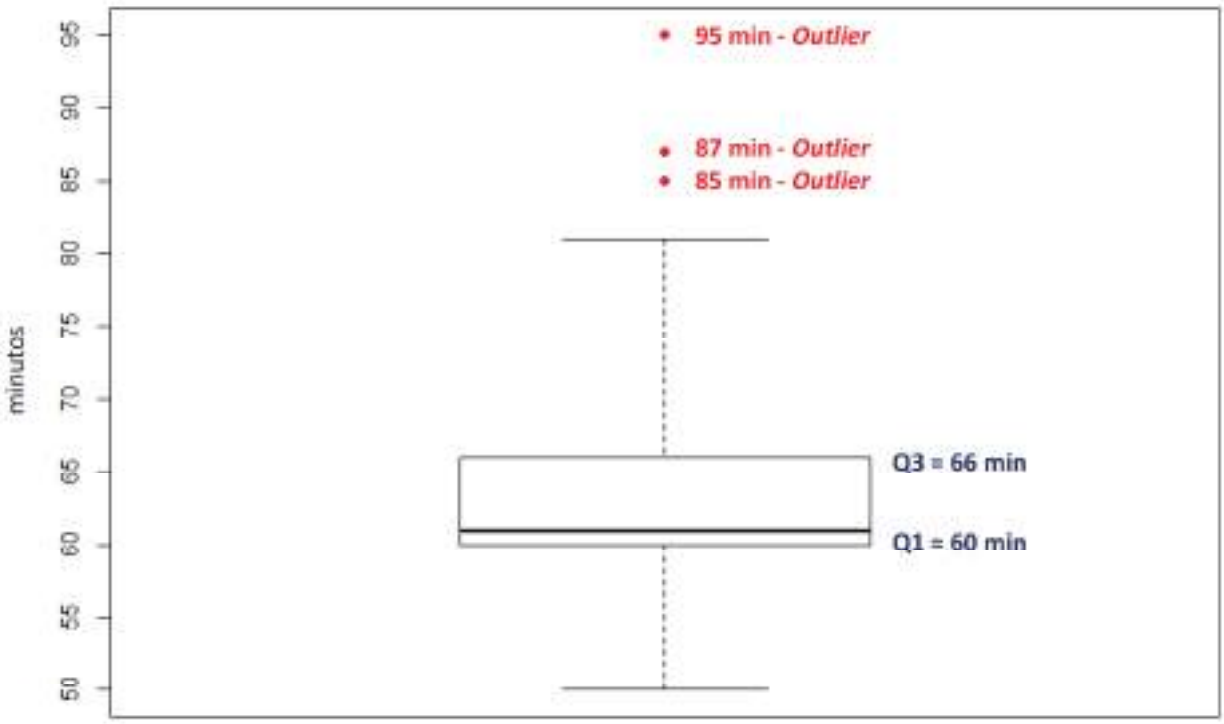


Figura IV.4: Exemplo de verificação de *outliers* na rota Guarulhos (São Paulo) - Galeão (Rio de Janeiro)

significou a remoção de voos com atrasos superiores a 2 horas e 45 minutos ou adiantamentos superiores a 37 minutos.

A Tabela IV.3 apresenta o resumo da etapa de limpeza indicando o número de registros resultantes após a remoção de discrepâncias e *outliers*, enquanto que a Figura IV.5 ilustra a redução do número de tuplas do *dataset*. Verifica-se que foram removidos aproximadamente 2% dos registros do *dataset*. Cabe destacar que o número de registros obtidos ao final da etapa de limpeza é o número final de voos considerados para a geração das regras de associação.

Tabela IV.3: Número de registros resultantes após a etapa de limpeza

Número inicial de registros	Número de registros após etapa de seleção	Número de registros após remoção de discrepâncias	Número de registros após remoção de *outliers* de duração dos voos	Número de registros após remoção de *outliers* de atrasos e adiantamentos
4.476.660	2.852.441	2.851.864	2.825.325	2.798.002

IV.4 Transformação

A etapa de transformação foi executada para discretizar atributos contínuos ou para substituir dados muito específicos por níveis mais gerais. Sendo assim, foram derivados 11 novos atributos a partir da utilização das técnicas de hierarquia conceitual, *binning* e agregação temporal, con-

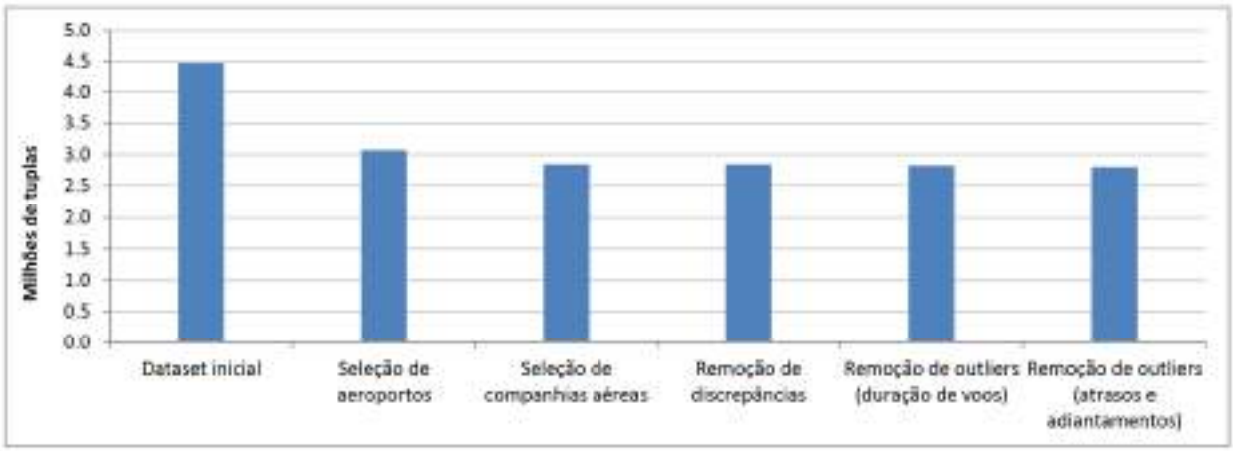


Figura IV.5: Redução do número de tuplas do *dataset* após a etapa de limpeza

forme apresentado na Tabela IV.4. Além disso, foi criado um indicador de atraso, definido como verdadeiro sempre que a diferença entre os horários reais e previstos de partida é de pelo menos 1 minuto ou quando o voo é cancelado.

Considerando que datas e horários de partida podem ser únicos, dependendo do suporte mínimo, poucas regras de associação envolvendo atributos temporais seriam geradas. Sendo assim, ao aplicar a técnica de hierarquia conceitual para discretizar a data e o horário previstos de partida, é possível derivar cinco novos atributos de níveis superiores: ano, mês, dia da semana, um indicador de dia útil e período do dia. Além destes, dois atributos meteorológicos que possuem escalas bastante difundidas também foram transformados a partir de hierarquias conceituais. A velocidade do vento foi substituída pelos rótulos das categorias da Escala de Beaufort (STORM PREDICTION CENTER, 2015), enquanto os décimos de nebulosidade foram representados pelos códigos dos relatórios METAR (THE UNITED STATES DEPARTMENT OF TRANSPORTATION, 1999).

A técnica de *binning* foi aplicada por meio do pacote SM (BOWMAN; AZZALINI, 2015) do *software* R (R CORE TEAM, 2014) para a discretização de atributos meteorológicos, como temperatura, pressão e umidade relativa. Por serem contínuos, assim como as datas e horários previstos de partida, os valores destes atributos dificilmente se repetem no *dataset*. No entanto, diferentemente da velocidade do vento e da nebulosidade, não existem escalas amplamente divulgadas para a substituição por meio de hierarquia conceitual. Sendo assim, considerando a distribuição dos dados, o domínio dos valores originais foi dividido em intervalos ou *bins* e cada valor passou a ser representado pela média de seu *bin*. O número de *bins* foi selecionado com base na máxima curvatura da *spline* considerando os valores residuais da divisão em q-quantis, com q variando de dois a dez. Uma vez que todas as análises de máxima curvatura indicaram

Tabela IV.4: Transformações

Dimensão	Atributo original	Domínio dos valores originais	Atributo transformado	Domínio dos valores transformados	Técnica
Temporal	Partida prevista (data/hora)	de 01/01/2009 0:00 a 28/02/2015 23:59	Ano	2009 a 2015	Hierarquia conceitual
			Mês	1 a 12	Hierarquia conceitual
			Dia da semana	Domingo a Sábado	Hierarquia conceitual
			Dia útil	Verdadeiro ou Falso	Hierarquia conceitual
			Período do dia	Início da manhã: 5:00 a 8:59 Meio da manhã: 9:00 a 10:59 Fim da manhã: 11:00 a 12:59 Tarde: 13:00 a 16:59 Início da noite: 17:00 a 19:59 Fim da noite: 20:00 a 22:59 Madrugada: 23:00 a 4:59	Hierarquia conceitual
Meteorológica	Temperatura(C)	1.6 a 41.4	Temperatura	Baixa:1.6 to 21.7 Média: 21.8 a 26.0 Alta: 26.1 a 41.4	*Binning*
Meteorológica	Pressão (mbar)	865.3 a 1032.7	Pressão	Baixa:865.3 a 924.4 Média: 924.5 a 1008.8 Alta:1008.9 a 1032.7	*Binning*
Meteorológica	Umidade relativa(%)	10 a 100	Umidade relativa	Baixa: 10 a 64 Média: 65 a 79 Alta: 80 a 100	*Binning*
Meteorológica	Velocidade do vento (m/s)	0 a 14	Velocidade do vento	Calma: 0 a 0.514 Aragem: 0.515 a 1.543 Brisa leve: 1.544 a 3.087 Brisa fraca: 3.088 a 5.144 Brisa moderada: 5.145 a 8.231 Brisa forte: 8.232 a 10.803 Vento fresco: 10.804 a 13.890 Vento forte: 13.890 a 14.000	Hierarquia conceitual
Meteorológica	Nebulosidade (décimos)	0 a 10	Nebulosidade	Céu limpo - Sky clear (SKC): 0 Poucas nuvens - Few (FEW): 1 a 3 Nuvens esparsas - Scattered (SCT): 4 a 5 Nublado - Broken (BKN): 6 a 9 Completamente encoberto - Overcast (OVC): 10	Hierarquia conceitual
Estado do sistema	Percentual de atrasos	0 a 100%	Nível de atraso	Baixo: 0 a 15.37% Médio: 15.38% a 33.32% Alto: 33.33% a 100%	Agregação temporal e *binning*

de voos em cada *bin*.

A agregação temporal foi utilizada para estimar o nível de atraso de um determinado aeroporto

no momento da partida prevista de um voo. Assim, para cada voo, este atributo foi definido como o percentual de voos atrasados considerando todas as partidas e chegadas previstas no aeroporto até uma hora antes da partida prevista do voo. Uma vez que os percentuais são valores contínuos e sem escalas conhecidas, a discretização foi realizada a partir da aplicação da técnica de *binning* utilizando a máxima curvatura da *spline* para a determinação do número de *bins*, o que novamente resultou em três *bins* rotulados como baixo, médio e alto.

IV.5 Geração das Regras de Associação

Para a geração das regras de associação, o algoritmo Apriori foi aplicado por meio do pacote Arules (HAHSLER et al., 2015) do *software* R (R CORE TEAM, 2014). Este pacote permite a definição de valores mínimos para suporte e confiança, número de atributos por regra (comprimento), bem como a fixação de atributos e valores como antecedente ou consequente das regras. O consequente representa o RHS da regra e foi definido como *Atraso = Verdadeiro*, o que significa que todas as regras estão associadas a atrasos nas partidas. O antecedente corresponde ao LHS da regra e indica as condições que levam a um atraso, podendo conter todos os outros atributos e valores do *dataset*.

A confiança foi determinada com base no percentual total de atrasos do *dataset*. Este percentual, chamado de *p* neste trabalho, representa a probabilidade de um voo sofrer um atraso. Considerando os voos entre os 17 aeroportos selecionados durante o período de janeiro de 2009 a fevereiro de 2015, *p* é igual a 27,14%. Conforme visto no Capítulo III, Seção III.6, a confiança pode ser interpretada como uma probabilidade condicional. Sendo assim, nesta análise, a confiança mínima foi definida como *p*=0,2714, o que significa que as condições do LHS precisam produzir um percentual de atrasos maior ou igual a probabilidade simples *p* de um voo sofrer um atraso, ou seja, $P(atraso \mid atributos\ LHS) \geq P(atraso)$.

O comprimento mínimo das regras de associação representa o número mínimo de atributos necessários para a obtenção de uma regra significativa. No caso desta análise, este número é igual a 2, pois as regras devem ter ao menos um atributo no LHS e o indicador de atrasos no RHS para fazerem sentido. Em outras palavras, sem nenhum atributo no LHS, não é possível obter as condições que conduzem a atrasos aéreos.

O comprimento máximo e o suporte mínimo foram definidos de maneira conjunta com base no tempo de execução e no número de regras geradas. Como deseja-se obter conhecimento a partir destas regras, é importante preservar sua interpretabilidade e sentido, o que dificilmente é alcançado com a permissão de um grande número de atributos no LHS. Além disso, suportes muito baixos podem indicar padrões muito específicos, aumentando o número de regras e inviabilizando a análise e interpretação por seres humanos. Sendo assim, uma vez que o *dataset* possui

aproximadamente seis anos de operações e que cada ano possui 365 dias, o número de regras geradas foi verificado variando o suporte considerando o equivalente de 1 a 16 vezes por dia e aumentando o comprimento em um atributo a cada rodada, conforme demonstrado na Tabela IV.5. Para o *comprimento = 4*, já havia um número importante de regras e um tempo considerável de execução, de modo que o comprimento máximo das regras desta análise foi definido como quatro.

Tabela IV.5: Definição do comprimento máximo e do suporte

		Comprimento=2		Comprimento=3		Comprimento=4	
Suporte	Equivalente em vezes por dia	Número de regras	Tempo de execução(s)	Número de regras	Tempo de execução(s)	Número de regras	Tempo de execução(s)
0,000783	1	64	49,368	3.372	61,520	60.142	129,773
0,001563	2	59	49,660	2.928	61,529	36.669	126,561
0,002348	3	59	50,226	2.579	62,015	24.824	121,461
0,003131	4	59	49,757	2.308	62,721	18.017	117,824
0,003913	5	59	49,718	2.058	61,473	13.908	114,569
0,004696	6	59	49,649	1.839	61,080	11.017	110,566
0,005478	7	57	49,190	1.639	61,799	8.634	107,795
0,006262	8	56	49,490	1.494	60,674	6.867	106,148
0,007044	9	54	49,620	1.354	60,796	5.636	103,294
0,007827	10	54	50,372	1.244	60,734	4.719	101,389
0,008610	11	54	49,557	1.128	60,544	4.064	99,285
0,009392	12	54	49,677	1.047	60,352	3.540	97,238
0,010175	13	54	49,571	967	60,898	3.073	95,650
0,010958	14	51	49,335	894	60,786	2.681	93,916
0,011741	15	51	49,506	846	59,503	2.364	92,417
0,012523	16	51	49,745	810	59,595	2.104	90,730

Como a análise culmina em uma avaliação de regras com quatro atributos, o suporte foi determinando considerando o número de regras de associação geradas para *comprimento=4*. Sendo assim, a máxima curvatura da *spline* considerando o número de regras geradas para cada suporte resultou em um suporte equivalente de quatro vezes por dia, ou $3.131 \cdot 10^{-3}$. Isso significa que, para ser considerada frequente, a regra deve ter ocorrido em média quatro vezes por dia em um período de seis anos. Esta escolha é razoável, uma vez que o foco desta análise é um panorama geral do sistema brasileiro de aviação comercial e a avaliação de muitas regras poderia dificultar ou ocultar as principais causas dos atrasos sofridos por aeroportos ou companhias aéreas. Adicionalmente, este número não é tão baixo a ponto de desconsiderar as regras mais importantes, uma vez que o menor aeroporto e a menor companhia aérea possuem em média 20 e 86 voos por dia, respectivamente.

Capítulo V Padrões Frequentes em Atrasos e Cancelamentos de Voos Nacionais

O algoritmo Apriori foi aplicado ao *dataset* de voos brasileiros, que, conforme ilustrado na Figura V.1, possui os onze atributos transformados e outros quatro atributos originais classificados em cinco dimensões: características, estado do sistema, espacial, temporal e meteorológica. Este *dataset* de 2.798.002 tuplas representa os voos nacionais das quatro principais companhias aéreas previstos para partirem entre janeiro de 2009 e fevereiro de 2015 entre os 17 maiores aeroportos do país.

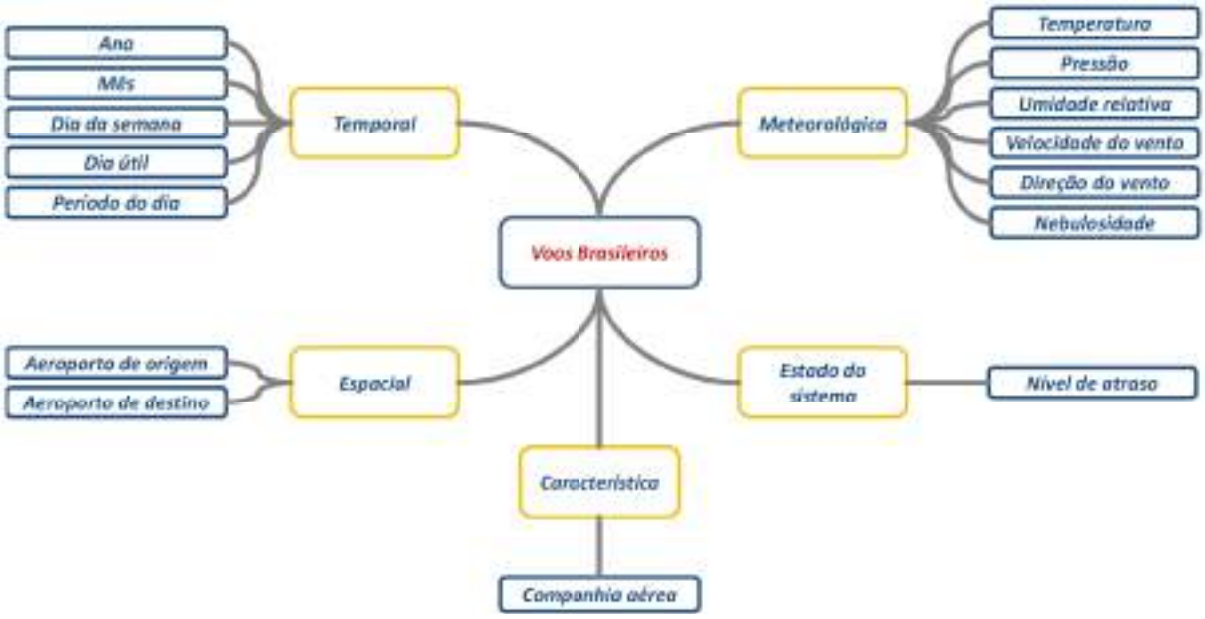


Figura V.1: Dimensões e atributos do *dataset* final de voos brasileiros

Como o número de regras geradas ainda é muito grande para a análise e intrpretação de seres humanas, estabeleceu-se um critério de avaliação baseado no *lift* a partir das regras de associação de comprimento mínimo (*comprimento mín*) até as regras de comprimento máximo (*comprimento máx*). Primeiramente, as regras significativas de *comprimento mín* foram selecionadas e avaliadas. Depois, foram pesquisadas regras de *comprimento = comprimento mín + 1* contendo os mesmos atributos das regras de *comprimento mín*, mas com *lifts* maiores para cada um dos atributos envolvidos. Este procedimento foi repetido até *comprimento = comprimento máx*.

Adicionalmente, como as regras estão sendo filtradas de acordo com o seu *lift*, um segundo critério de análise foi estabelecido de modo a garantir que todas as regras com *lifts* altos sejam incluídas na avaliação. Desta forma, com base no *lift*, todas as regras de associação do primeiro quartil cujos atributos não foram analisados para comprimentos menores também serão consideradas como regras de interesse. Sendo assim, a Tabela V.1 destaca para cada comprimento o número total de regras e o número que satisfaz pelo menos uma das condições de interesse.

Tabela V.1: Número de regras interessantes para cada comprimento

Comprimento	Número de regras	Número de regras interessantes
2	59	59
3	2.308	1.114
4	18.017	6.604

V.1 Análise de regras de comprimento dois

Quando o *comprimento = 2*, as regras possuem apenas um atributo no LHS e o indicador de atraso no RHS. Neste caso, foram obtidas 59 regras, sendo que as 18 regras contendo *lift* superior a 1,1 estão representadas na Tabela V.2. O conjunto completo de regras pode ser encontrado em Sternberg et al. (2016). As regras foram analisadas de acordo com suas dimensões, entretanto regras envolvendo alguns atributos meteorológicos, como direção do vento, temperatura, pressão e umidade relativa não foram avaliadas neste estudo porque sua interpretação dependeria de conhecimentos específicos sobre meteorologia e, eventualmente, segurança de voo.

Tabela V.2: Regras para atrasos aéreos de *comprimento = 2* com *lift* superior a 1,1

LHS	Suporte	Confiança	*Lift*
Nível de atraso=Alto	0,1507	0,4238	1,5618
Mês=Dezembro	0,0306	0,3451	1,2720
Horário=Fim da noite	0,0454	0,3288	1,2116
Aeroporto de destino=Manaus	0,0051	0,3260	1,2013
Aeroporto de origem=Guarulhos (São Paulo)	0,0353	0,3236	1,1926
Ano=2011	0,0523	0,3104	1,1438
Mês=Junho	0,0206	0,3096	1,1409
Aeroporto de origem=Brasília	0,0331	0,3087	1,1377
Aeroporto de destino=Belém	0,0055	0,3087	1,1376
Nebulosidade na origem=OVC	0,0851	0,3087	1,1375
Dia da semana=Sexta-feira	0,0462	0,3071	1,1316
Horário=Madrugada	0,0137	0,3062	1,1285
Companhia aérea=Gol	0,1289	0,3055	1,1258
Aeroporto de origem=Galeão (Rio de Janeiro)	0,0247	0,3046	1,1225
Direção do vento=Sudoeste	0,0156	0,3041	1,1207
Mês=Julho	0,0211	0,3019	1,1127
Aeroporto de origem=Manaus	0,0047	0,3009	1,1090
Mês=Janeiro	0,0297	0,2993	1,1032

O estado do sistema, representado pelo atributo *nível de atraso=alto* apresentou o maior

lift do conjunto de regras. O *lift* é de aproximadamente 1,56, o que significa que quando um aeroporto está enfrentando altos níveis de atraso, um voo tem 56% mais chances de atrasar. Isso sugere que os aeroportos brasileiros sofrem com cenários de congestionamento e propagação de atrasos. Além disso, o suporte da regra aponta que os atrasos sistêmicos representados pelos altos níveis de atraso são a causa de novos atrasos em 15,07% dos casos.

As regras de dimensão temporal sugerem que o sistema brasileiro de aviação comercial tem melhorado seu desempenho, uma vez que apenas os anos de 2010, 2011 e 2012 aparecem no conjunto de regras geradas. Os meses associados aos atrasos são, principalmente, meses de férias de verão e inverno. Além disso, os atrasos são mais comuns em dias úteis, especialmente nas quartas, quintas e sextas-feiras. Neste caso, é importante destacar que conforme mencionado no Capítulo IV , o indicador de atrasos também sinaliza cancelamentos. Assim, estes atrasos podem apontar tanto picos de demanda como cancelamentos devido aos baixos níveis de demanda. Adicionalmente, os períodos do início e meio da manhã são os únicos que não aparecem no conjunto de regras, sugerindo, novamente, que os atrasos se propagam ao longo do dia.

Dentre os atributos meteorológicos remanescentes, a nebulosidade foi o único que formou regras significativas. Quando o céu está totalmente encoberto, os voos têm 14% mais chances de sofrerem um atraso. As regras contendo a velocidade do vento foram descartadas, pois não fizeram sentido em um contexto de atrasos aéreos. No conjunto de regras, aparecem apenas velocidades classificadas como calma, aragem e brisas fraca e moderada, que são muito comuns, mas não fortes o suficiente para influenciar a performance de um voo.

As regras relacionadas à dimensão espacial representam os aeroportos de origem com mais chances de produzir novos atrasos e os aeroportos de destino que podem influenciar as operações dos aeroportos de origem. Os maiores *lifts* foram verificados para as partidas de Guarulhos (São Paulo), Brasília, e Galeão (Rio de Janeiro) e chegadas em Manaus, Belém e Recife. Finalmente, quando consideradas regras de *comprimento=2*, apenas a Gol entre as quatro principais companhias aéreas aparece relacionada a atrasos. Os motivos para os atrasos associados a aeroportos e companhias aéreas somente poderão ser investigados profundamente nas regras de comprimentos maiores.

V.2 Análise de regras de comprimento três

Inicialmente, foram investigadas as regras contendo ao menos um atributo avaliado na análise de *comprimento = 2*, porém com *lifts* superiores àqueles encontrados para *comprimento = 2*. Esta condição foi satisfeita por 1.113 regras de *comprimento = 3*, enquanto o critério do primeiro quartil foi atendido por mais uma regra. A Figura V.2 apresenta o *boxplot* dos *lifts* em ordem decrescente,

destacando o valor máximo, o mínimo, a mediana e os *lifts* correspondentes ao primeiro e terceiro quartis. Assim, todas as regras com *lifts* iguais ou superiores a 1,163349 foram verificadas.



Figura V.2: *Boxplot* dos *lifts* das regras de associação de *comprimento = 3*

Devido à grande quantidade, a apresentação e avaliação do conjunto completo de regras interessantes não é possível, de modo que apenas os padrões mais significativos serão discutidos. A Tabela V.3 resume as 10 primeiras regras de associação encontradas para *comprimento = 3*, ordenadas de maneira decrescente de acordo com o valor do *lift*. O conjunto completo de regras pode ser encontrado em Sternberg et al. (2016).

Tabela V.3: Dez primeiros *lifts* das regras interessantes para atrasos aéreos de *comprimento = 3*

LHS	Suporte	Confiança	*Lift*
Aeroporto de origem=Santos Dumont (Rio de Janeiro) Nível de atraso=Alto	0,0130	0,4800	1,7691
Aeroporto de destino=Congonhas (São Paulo) Nível de atraso=Alto	0,0210	0,4796	1,7673
Mês=Dezembro Nível de atraso=Alto	0,0208	0,4742	1,7476
Companhia aérea=Gol Nível de atraso=Alto	0,0722	0,4732	1,7437
Horário=Fim da noite Nível de atraso=Alto	0,0301	0,4661	1,7178
Nebulosidade na origem=OVC Nível de atraso=Alto	0,0529	0,4643	1,7112
Aeroporto de destino=Goiânia Nível de atraso=Alto	0,0036	0,4634	1,7077
Dia da semana=Sexta-feira Nível de atraso=Alto	0,0287	0,4579	1,6875
Mês=Junho Nível de atraso=Alto	0,0132	0,4569	1,6838
Horário=Meio da manhã Nível de atraso=Alto	0,0159	0,4564	1,6821

Os maiores *lifts* foram verificados para as regras que combinam altos níveis de atraso com partidas no aeroporto Santos Dumont no Rio de Janeiro ou chegadas no aeroporto de Congo-

nhas em São Paulo. Estes dois aeroportos possuem características comuns, como localizações próximas ao centro da cidade, pistas curtas e altas demandas. Além disso, estes aeroportos estão fortemente conectados por voos que partem aproximadamente a cada 8 minutos em horários de pico. Outros aeroportos de origem, como Galeão no Rio de Janeiro, Guarulhos e novamente Congonhas em São Paulo também apareceram associados a novos atrasos quando combinados com altos níveis de atraso. Todos estão entre os cinco maiores do Brasil. Assim, os resultados sugerem que uma ou mais etapas das operações dos voos nestes aeroportos estão sendo executadas em níveis próximos a suas capacidades, dificultando a recuperação quanto à propagação de atrasos.

Os altos níveis de atraso também apareceram combinados aos meses de férias e períodos de pico, como manhã e noite. Isso sugere que os aeroportos têm dificuldades em lidar com picos de demanda, levando a ocorrência de novos atrasos nestes períodos e propagando-os aos próximos voos. Além de atributos temporais, os altos níveis de atraso somados a momentos em que o céu está completamente encoberto também conduz a novos atrasos. Adicionalmente, algumas regras apresentaram esta condição de nebulosidade combinada com alguns aeroportos ou com a companhia aérea Gol. Sabe-se que no Brasil, alguns aeroportos e aeronaves não possuem instrumentos apropriados ou possuem limitações geográficas para operar com segurança quando o céu está totalmente encoberto. Espera-se identificar estes aeroportos e suas características comuns na análise de *comprimento = 4*.

Considerando o critério de *lift* utilizado para selecionar as regras de *comprimento=3*, novamente a Gol foi a única companhia aérea presente no conjunto de regras interessantes. Estas regras apresentaram no LHS, além da companhia aérea, altos níveis de atraso ou alguns aeroportos de origem, como Guarulhos em São Paulo, Galeão e Santos Dumont no Rio de Janeiro e Brasília. Isso pode sugerir que a Gol opera com menores tempos de espera entre seus voos, especialmente nos aeroportos mencionados.

V.3 Análise de regras de comprimento quatro

Novamente, devido ao grande número de regras geradas, serão detalhados somente os padrões mais significativos. Para ilustrar a distribuição dos *lifts* das regras, a Figura V.3 apresenta seu *boxplot*, destacando o valor máximo, o mínimo, a mediana e os *lifts* correspondentes ao primeiro e terceiro quartis. Sendo assim, além da investigação das regras contendo os atributos avaliados na análise de *comprimento = 3*, todas as regras com *lifts* iguais ou superiores a 1,242746 foram verificadas.

Com base nos resultados anteriores, há interesse em descobrir porque algumas companhias aéreas e aeroportos tendem a produzir mais atrasos do que outros. Além disso, pretende-se veri-



Figura V.3: *Boxplot* dos *lifts* das regras de associação de *comprimento = 4*

ficar se existem condições temporais ou meteorológicas que aumentam as chances de ocorrência de um novo atraso. Neste sentido, a Tabela V.4 resume as 10 primeiras regras de associação encontradas para *comprimento = 4*, ordenadas de maneira decrescente de acordo com o valor do *lift*. O conjunto completo de regras pode ser encontrado em Sternberg et al. (2016).

Tabela V.4: Dez primeiros *lifts* das regras interessantes para atrasos aéreos de *comprimento = 4*

LHS	Suporte	Confiança	*Lift*
Companhia aérea=Gol Aeroporto de origem=Santos Dumont (Rio de Janeiro) Nível de Atraso=Alto	0,0062	0,5715	2,1061
Aeroporto de origem=Santos Dumont (Rio de Janeiro) Aeroporto de destino=Congonhas (São Paulo) Nível de Atraso=Alto	0,0084	0,5562	2,0498
Companhia aérea=Gol Mês=Dezembro Nível de Atraso=Alto	0,0102	0,5471	2,0162
Aeroporto de destino=Congonhas (São Paulo) Horário=Início da noite Nível de Atraso=Alto	0,0058	0,5457	2,0112
Aeroporto de origem=Santos Dumont (Rio de Janeiro) Nebulosidade na origem=OVC Nível de Atraso=Alto	0,0048	0,5416	1,9961
Horário=Fim da noite Mês=Dezembro Nível de Atraso=Alto	0,0042	0,5399	1,9898
Aeroporto de origem=Santos Dumont (Rio de Janeiro) Nebulosidade no destino=OVC Nível de Atraso=Alto	0,0062	0,5364	1,9766
Aeroporto de destino=Congonhas (São Paulo) Dia da semana=Sexta-feira Nível de Atraso=Alto	0,0044	0,5354	1,9732
Aeroporto de origem=Galeão (Rio de Janeiro) Nebulosidade na origem=OVC Nível de Atraso=Alto	0,0052	0,5344	1,9695
Aeroporto de origem=Santos Dumont (Rio de Janeiro) Umidade relativa na origem=Alta Nível de Atraso=Alto	0,0055	0,5321	1,9608

Conforme verificado nas Seções V.1 e V.2, os resultados sugerem que a Gol opera com menores tempos de espera entre seus voos, uma vez que esta companhia aérea tem mais chances de produzir novos atrasos quando os aeroportos estão enfrentando altos níveis de atraso. Analisando as regras de *comprimento = 4*, este cenário ocorre, principalmente, nos aeroportos do Rio de Janeiro, São Paulo, Vitória e Brasília. Em especial, no Santos Dumont no Rio de Janeiro, o *lift* alcança o patamar de 2,10, o que significa que quando este aeroporto registra altos níveis de atraso, um voo da Gol possui 110% mais chances de atrasar. Entretanto, é importante destacar que os voos da Gol apresentaram *lifts* maiores durante o período de 2009 a 2012, claramente sugerindo que a Gol tem melhorado suas operações nos últimos três anos.

Considerando as regras do primeiro quartil, Azul e Tam também produziram regras significativas. Os maiores *lifts* foram de aproximadamente 1,5 e foram encontrados para altos níveis de atraso combinados com atributos espaciais ou temporais, respectivamente. Os voos da Azul têm mais chances de atrasarem quando partem do aeroporto de Campinas. Neste caso, é importante destacar que Campinas é a base e principal aeroporto para as operações da Azul, ou seja, quando este aeroporto enfrenta altos níveis de atraso, a Azul é a companhia aérea mais impactada. No caso da Tam, o maior *lift* foi verificado para partidas em 2014, mas ainda foi menor que o *lift* verificado para a Gol sob as mesmas condições.

Comparando as duas maiores cidades do Brasil, os aeroportos do Rio de Janeiro parecem ser muito influenciados pelas condições de nebulosidade, especialmente quando o céu está completamente encoberto, enquanto em São Paulo os atributos temporais, como dia da semana, são mais relevantes, conforme demonstrado na Tabela V.5. A nebulosidade pode afetar mais os aeroportos do Rio de Janeiro por diversos motivos, como condições geográficas ou instrumentação de aeroportos e, desta forma, seria necessário um estudo especifico para determinar a raiz deste problema. Por outro lado, sabe-se que sextas-feiras representam picos de demanda nos aeroportos de São Paulo, uma vez que esta é a cidade brasileira mais importante para os negócios. Assim, muitas pessoas deixam São Paulo e retornam para suas casas nas sextas-feiras, sugerindo a formação de um cenário de congestionamento acima das possibilidades dos aeroportos.

Com base nos resultados de *comprimento=3*, conforme esperado, voos partindo do Santos Dumont (Rio de Janeiro) para Congonhas (São Paulo) quando o aeroporto de origem está enfrentando altos níveis de atraso têm *lift* de 2,05, ou 105% mais chances de atrasarem. A operação inversa, ou seja, de Congonhas para Santos Dumont produz um *lift* de apenas 1,77. Novamente, é importante destacar que estes aeroportos são muito demandados e importantes para suas cidades, especialmente devido às suas localizações próximas às regiões centrais. Além disso, São Paulo e Rio de Janeiro são as maiores cidades brasileiras e as mais importantes para os negócios. O serviço de voos entre estes dois aeroportos leva menos de 1 hora e é o mais fre-

Tabela V.5: Principais regras de associação para atrasos aéreos gerados nos aeroportos do Rio de Janeiro e de São Paulo

LHS	Suporte	Confiança	*Lift*
Origem=Santos Dumont (Rio de Janeiro) Nível de atraso=Alto Nebulosidade na origem=OVC	0,005	0,542	1,9961
Origem=Galeão (Rio de Janeiro) Nível de atraso=Alto Nebulosidade na origem=OVC	0,005	0,534	1,9695
Origem=Congonhas (São Paulo) Nível de atraso=Alto Dia da semana=Sexta-feira	0,005	0,503	1,8549
Origem=Guarulhos (São Paulo) Nível de atraso=Alto Dia da semana=Sexta-feira	0,004	0,485	1,7889

quente do Brasil. Apesar de suas pistas curtas e restrições quanto ao tamanho das aeronaves, a rota entre estes dois aeroportos é extremamente demandada e possui voos partindo a cada 8 minutos de cada aeroporto em horários de pico.

Considerando este par de aeroportos, além dos altos níveis de atrasos, *lifts* importantes foram verificados para condições de nebulosidade em que o céu está completamente encoberto no Rio de Janeiro, mesmo quando o voo decola em Congonhas, e para o horário de partida em ambos os aeroportos. A nebulosidade em São Paulo também é considerável, porém em menor escala quando comparada ao Rio de Janeiro. A Tabela V.6 apresenta as principais regras de associação para estes dois aeroportos, corroborando a ideia de que o Rio de Janeiro sofre mais com questões relacionadas à nebulosidade e, assim, tende a produzir mais atrasos. Adicionalmente, melhores programações no início da noite ou investimentos em capacidade em ambos os aeroportos também poderiam ser avaliados para evitar novos atrasos.

Os atributos temporais continuam indicando que novos atrasos são produzidos a partir de propagação. Mesmo analisando as regras contendo os períodos do início e meio da manhã, percebe-se que os maiores *lifts* ocorrem para o período do meio da manhã combinado com altos níveis de atraso e outros atributos, como companhia aérea e nebulosidade. Adicionalmente, os atributos temporais também sugerem atrasos em demandas de pico, como meses de férias. Outro resultado importante está relacionado ao ano do voo. Os resultados apontam que o Brasil tem melhorado em termos de atrasos aéreos, uma vez que somente 65 das 889 regras contendo o ano de partida ocorreram em ou após 2013.

Considerando a dimensão meteorológica, a condição de nebulosidade em que o céu está completamente encoberto continua sendo o atributo mais importante associado a atrasos. Entretanto, também foi verificado que quando o aeroporto de origem apresenta temperatura média,

Tabela V.6: Principais regras de associação para atrasos nos voos entre os aeroportos de Congonhas (São Paulo) e Santos Dumont (Rio de Janeiro)

LHS	Suporte	Confiança	*Lift*
Origem=Santos Dumont (Rio de Janeiro) Destino=Congonhas (São Paulo) Nível de atraso=Alto	0,008	0,556	2,050
Origem=Congonhas (São Paulo) Destino=Santos Dumont (Rio de Janeiro) Nível de atraso=Alto	0,008	0,480	1,7674
Origem=Santos Dumont (Rio de Janeiro) Destino=Congonhas (São Paulo) Nebulosidade na origem=OVC	0,004	0,392	1,4448
Origem=Santos Dumont (Rio de Janeiro) Destino=Congonhas (São Paulo) Período do dia=Início da noite	0,004	0,370	1,3618
Origem=Congonhas (São Paulo) Destino=Santos Dumont (Rio de Janeiro) Nebulosidade no destino=OVC	0,004	0,368	1,3545
Origem=Santos Dumont (Rio de Janeiro) Destino=Congonhas (São Paulo) Nebulosidade no destino=OVC	0,008	0,3426	1,2625
Origem=Congonhas (São Paulo) Destino=Santos Dumont (Rio de Janeiro) Período do dia=Início da noite	0,003	0,3227	1,1892
Origem=Congonhas (São Paulo) Destino=Santos Dumont (Rio de Janeiro) Nebulosidade na origem=OVC	0,007	0,3155	1,1627

umidade relativa alta e pressão baixa, atrasos têm 25% mais chances de ocorrer. Neste caso, é importante destacar que o Brasil é um país tropical e, desta forma, temperaturas médias já representam momentos quentes. Estes três atributos, ou seja, temperatura, umidade relativa e pressão, estão relacionados à precipitação. Assim, este resultado pode indicar que os aeroportos ou companhias aéreas brasileiras encontram algumas dificuldades em prover operações seguras quando está chovendo.

V.4 Análises complementares

Com base nos resultados verificados para as regras de associação gerais e com o objetivo de melhor compreender o sistema de transporte aéreo brasileiro, os principais atributos encontrados nas análises anteriores foram explorados de forma mais detalhada nesta etapa. Assim, novos conjuntos de regras foram gerados para suportes e confianças mais baixos (ambos iguais a to $1 \cdot 10^{-9}$). O RHS continuou sendo filtrado por {Atraso=Verdadeiro}, porém o LHS passou a ser fixado com os atributos de interesse, como ano, período do dia, companhia aérea, aeroporto ou

uma combinação destes. A exemplo das análises anteriores, as regras de associação geradas foram avaliadas de acordo com seus *lifts*. Neste caso, como os valores mínimos de suporte e confiança são praticamente nulos, o *lift* mostra quais valores de cada atributo tendem a produzir novos atrasos, mesmo em condições não muito frequentes durante o período de janeiro de 2009 a fevereiro de 2015.

Inicialmente, foram avaliadas as regras de associação contendo o *ano de partida*, que conduzem aos atrasos apresentados na Figura V.4. Neste caso, quanto maior o *lift*, maior o percentual de atrasos ocorridos ao longo do ano. Conforme verificado, desde 2013, os *lifts* têm se mantido abaixo de 1, indicando que o Brasil tem reduzido seus atrasos aéreos, ou seja, o desempenho do sistema de transporte aéreo do país melhorou.

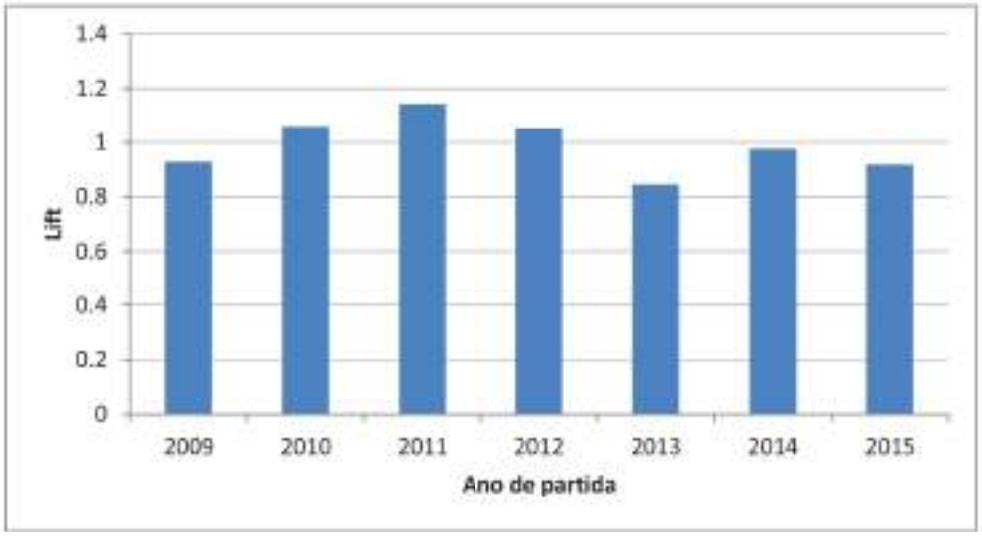


Figura V.4: Análise do *lift* das regras de associação contendo somente o ano de partida no LHS e a ocorrência de um atraso no RHS

Ao analisar a ocorrência de atrasos a partir da combinação de *companhia aérea* e *ano de partida*, foi observado que a Gol é a companhia que apresenta mais chances de produzir um novo atraso, conforme representado na Figura V.5. No entanto, apesar da Gol ainda ser a companhia com o maior *lift* em relação à ocorrência de atrasos, é importante destacar que houve considerável progresso nos últimos três anos. Por outro lado, a Azul aumentou seus *lifts* ao longo do tempo, ficando próximo a 1 em 2015, provavelmente devido ao aumento de suas operações nos últimos anos. Este aumento pode ser percebido como um alerta aos gestores desta companhia aérea, que devem buscar soluções para evitar que os atrasos continuem crescendo. Por outro lado, a Avianca e a Tam apresentaram comportamentos e valores de *lift* semelhantes e, apesar do aumento nos anos centrais da análise, seus *lifts* em 2015 voltaram a cair para patamares próximos aos de 2009.

De maneira semelhante, a Figura V.6 ilustra a análise da combinação de *aeroporto* e *ano de*

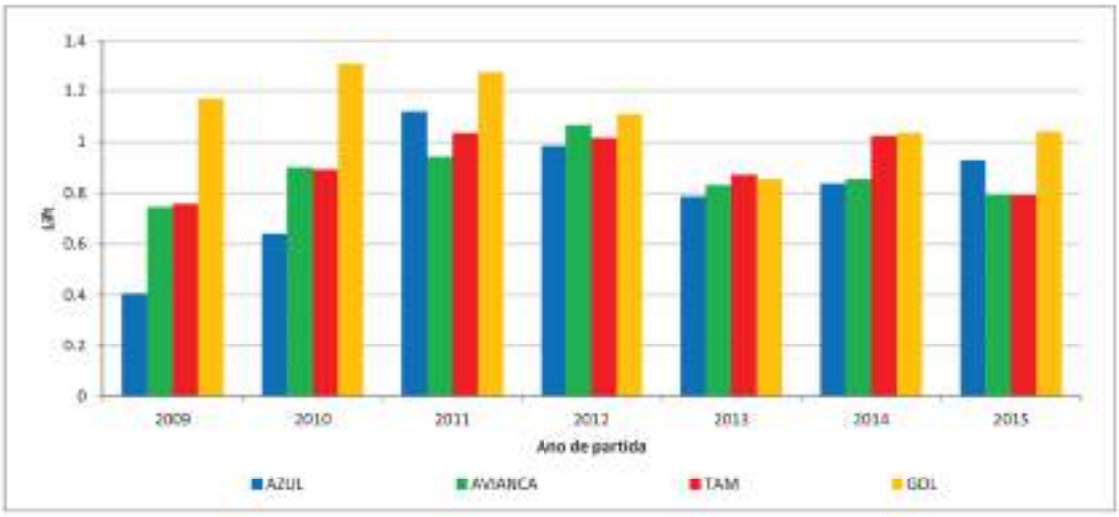


Figura V.5: Análise do *lift* das regras de associação contendo a companhia aérea e o ano de partida no LHS e a ocorrência de um atraso no RHS

partida em relação à ocorrência de novos atrasos. Foram observados diferentes comportamentos nos últimos seis anos. Como tendência geral, é possível apontar que a maioria dos aeroportos apresentou seu menor *lift* em 2013. Os picos gerais de atrasos entre 2010 e 2012 já verificados anteriormente podem ser atribuídos a picos de atrasos em aeroportos centrais ou *hubs*, como Guarulhos em São Paulo, Galeão no Rio de Janeiro, Campinas e Salvador. Apesar da tendência de redução de atrasos nos últimos três anos, Brasília e Manaus tiveram *lifts* superiores a 1 e com pouca variação na maior parte do tempo. Ao mesmo tempo, Belo Horizonte parece definitivamente não seguir esta tendência, uma vez que entre seus *lifts*, o de 2015 somente não foi maior que o de 2011. Finalmente, apesar de serem aeroportos pequenos ou médios, Florianópolis, Fortaleza, Vitória e Goiânia apresentaram *lifts* inferiores a 1 em grande parte do tempo, demostrando bons desempenhos em relação à ocorrência de atrasos.

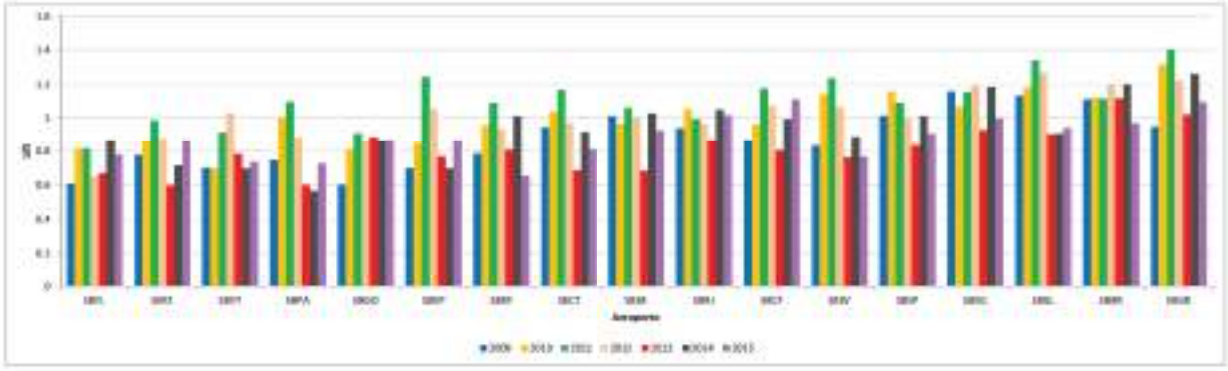


Figura V.6: Análise do *lift* das regras de associação contendo o aeroporto e o ano de partida no LHS e a ocorrência de um atraso no RHS

A Figura V.7 ilustra a análise das regras de associação contendo o *período do dia* que con-

duz a novos atrasos. Pode ser verificado que o início e o meio da manhã estão negativamente correlacionados à ocorrência de novos atrasos, novamente indicando que o Brasil sofre com a propagação de atrasos ao longo do dia. Como a maioria dos aeroportos brasileiros fecha ou tem suas operações reduzidas durante a madrugada, este período é utilizado para a recuperação dos atrasos sofridos durante o dia. Assim, na maioria dos casos, o início da manhã representa a abertura do aeroporto, sendo, portanto, menos suscetível a ocorrência de atrasos apesar do alto nível de demanda. Este efeito não pode ser observado no início e no fim da noite, outros momentos de grande demanda. Nestes períodos, o sistema não parece ser capaz de se recuperar de atrasos prévios ocorridos durante a manhã e a tarde, sendo, portanto, os momentos do dia com maiores chances de produzir novos atrasos. Finalmente, conforme mencionado anteriormente, durante a madrugada os aeroportos conseguem regularizar suas operações e, desta forma, os níveis de atraso diminuem.



Figura V.7: Análise do *lift* das regras de associação contendo somente o período do dia no LHS e a ocorrência de um atraso no RHS

Ao combinar os *aeroportos* com os *períodos do dia* para uma análise mais profunda, percebe-se que a maioria dos aeroportos segue a mesma tendência observada na análise geral, ou seja, menos atrasos nos dois primeiros períodos do dia e maior numero de ocorrências no início e no fim da noite. No entanto, alguns comportamentos inesperados foram descobertos, conforme demostrado na Figura V.8. Belo Horizonte, Curitiba, Campinas e Santos Dumont (Rio de Janeiro) apresentaram suas maiores correlações com atrasos na madrugada. Isso pode ser atribuído ao pequeno número de voos neste período, mas que frequentemente atrasam. Belém e Manaus possuem *lifts* altos durante a tarde e no fim da manhã. É importante destacar que estes aeroportos estão localizados na região norte do Brasil, em diferentes fusos horários, distante dos outros aeroportos principais e com condições particulares de clima.

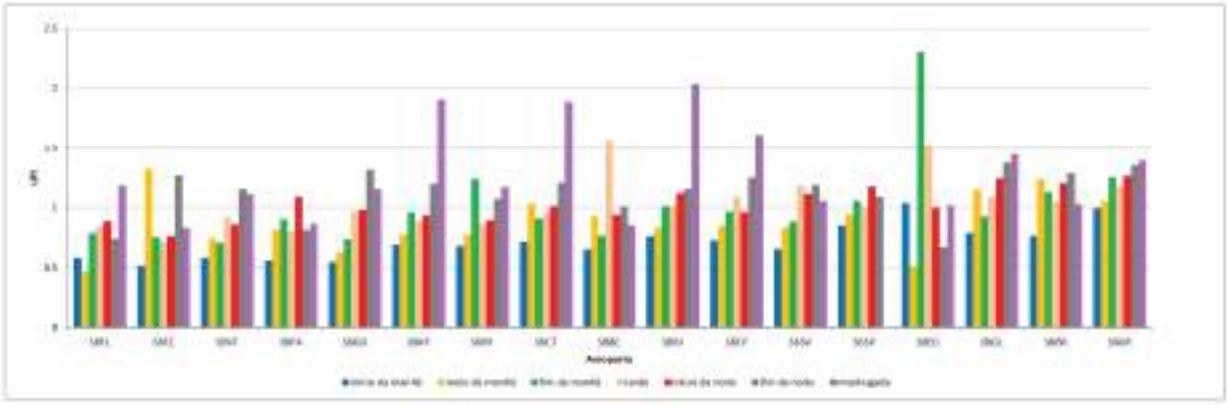


Figura V.8: Análise do *lift* das regras de associação contendo o aeroporto e o período do dia no LHS e a ocorrência de um atraso no RHS

Capítulo VI Trabalhos Relacionados

Conforme introduzido no Capitulo II, devido a sua importância, os atrasos aéreos têm sido um assunto frequente entre artigos relacionados à aviação comercial. De modo a estabelecer um mapa sistemático para a avaliação da literatura , a *string* ("airport delay"or "flight delay") and ("predict"or "forecast"or "propagate") foi utilizada para buscar publicações nas bases Science Direct e IEEE em Janeiro de 2015. Foram obtidas 448 referências, das quais 61 foram selecionadas para uma análise preliminar devido a sua relação com o tema de atrasos aéreos. Adicionalmente, outros cinco trabalhos foram encontrados com base nas referências bibliográficas das 61 publicações analisadas.

Das 66 publicações selecionadas, verificou-se que 9 não abordavam o problema de atrasos aéreos sob o ponto de vista de ciência de dados. Sendo assim, somente os 57 trabalhos restantes foram utilizados como base para a construção da taxonomia do Capitulo II. Além de considerarem diversas dimensões e atributos e serem aplicadas a diferentes escopos, as análises já publicadas podem ser classificadas quanto ao tipo de problema e ao método utilizado, conforme os resumos apresentados nas Tabelas VI.1 e VI.2. É importante destacar que cinco trabalhos tratam de mais de um tipo de problema e, por isso, a soma das publicações constantes na Tabela VI.1 é igual a 62 ao invés de 57.

Tabela VI.1: Classificação das publicações relacionadas a atrasos aéreos quanto ao tipo de problema

Novos atrasos	Propagação de atrasos	Análise de Cancelamentos
30	26	6

Tabela VI.2: Classificação das publicações relacionadas a atrasos aéreos quanto ao método aplicado

Análise estatística	Modelos probabilísticos	Representação de redes	Aprendizado de máquina	Pesquisa operacional
16	7	4	7	23

Considerando as 57 publicações analisadas, este Capítulo destaca aquelas relacionadas a compreensão de causas ou padrões de atrasos aéreos. Por exemplo, os impactos das condições meteorológicas nos voos foram investigados por meio de um modelo de simulação considerando

a capacidade de aeroportos sob condições visuais ou por instrumentos. Considerando diferentes durações destas condições meteorológicas, os autores puderam entender atrasos nas partidas e nas chegadas devido à formação de filas (SCHAEFER; MILLNER, 2001).

Com o objetivo de estimar a distribuição de atrasos nas partidas de uma determinada companhia aérea no Aeroporto Internacional de Denver, os padrões sazonais, como condições meteorológicas e demanda, a propagação de atrasos e os resíduos randômicos, como problemas mecânicos ou com bagagens foram modelados e analisados (TU; BALL; JANK, 2008). Com um objetivo semelhante, mas considerando os principais aeroportos dos Estados Unidos, as possíveis causas de atrasos, como volume de tráfego e condições meteorológicas foram discutidas para cada fase do voo de modo a obter as distribuições dos atrasos de partidas, em rota e de chegadas (MUELLER; CHATTERJI, 2002).

Os padrões de atrasos nas chegadas foram estudados no Aeroporto Internacional de Orlando por meio de uma análise de frequências das médias diárias de atrasos e uma análise estatística baseada em regressões e correlações. O objetivo era identificar quais atributos estavam relacionados aos atrasos de chegada, tanto para um voo específico como diariamente para o aeroporto. Os resultados demonstraram que os atrasos de chegada estão ligados às dimensões temporal, meteorológica e de programação, incluindo atributos como estação do ano, dia da semana, horário, precipitação e o tempo entre duas chegadas (ABDEL-ATY et al., 2007).

Na Europa, as causas dos atrasos sofridos pelos principais aeroportos foram avaliadas a partir de uma análise de correlações. Além dos atrasos, foram consideradas as capacidades dos aeroportos em em termos de infraestrutura, condições operacionais, demanda, *mix* de aeronaves e controle de tráfego. Adicionalmente, após discussão dos resultados, foram propostas medidas para evitar novos atrasos, bem como algumas oportunidades de investimentos. (REYNOLDS-FEIGHAN; BUTTON, 1999).

Na China, uma previsão de atrasos foi desenvolvida a partir de um sistema de recomendação. Como premissa, foi assumido que a rede de aeroportos apresentaria o mesmo comportamento quando enfrentasse uma situação já vivenciada no histórico de dados. Assim, computando os níveis atuais e anteriores de atrasos nos aeroportos e considerando os tempos de voo entre eles, os autores puderam estimar os novos níveis de atraso a cada hora (LU; WANG; XU, 2009).

Por outro lado, para compreender os impactos dos atrasos nas companhias aéreas, uma comparação entre uma companhia tradicional e uma *low cost* foi conduzida para investigar as causas de atrasos e os efeitos de sua propagação nas redes de voos (AHMADBEYGI et al., 2008). Com base em um modelo de filas, o congestionamento no Aeroporto Internacional de Los Angeles foi investigado. Neste cenário, foram discutidos os impactos sofridos por diferentes companhias aéreas e aeronaves que operavam neste aeroporto (HANSEN, 2002).

Considerando o estudo do sistema aéreo brasileiro, este trabalho é o primeiro a analisar atrasos sob uma abordagem de ciência de dados. A integração de dados da ANAC e do INMET possibilitaram o entendimento das principais razões e tendências em relação à ocorrência de novos atrasos em voos comerciais no Brasil. Os resultados alcançados são interessantes, uma vez que podem ser verificadas muitas semelhanças com resultados de estudos desenvolvidos em outros países. Alguns exemplos são a influência das dimensões temporal e meteorológica (ABDEL-ATY et al., 2007), dificuldades de recuperação em cenários de atrasos pré-existentes e desempenhos diferentes de companhias aéreas operando sob as mesmas condições (HANSEN, 2002).

Finalmente, destaca-se a aplicação original de regras de associação para a investigação de atrasos aéreos. Esta abordagem, tradicionalmente empregada para a análise da cesta de compras (HAN et al., 2007), já foi utilizada em problemas de transporte nos últimos anos, especificamente na análise de atrasos ferroviários (FLIER et al., 2009; CULE et al., 2011). Sua primeira aplicação para a análise de atrasos aéreos permitiu, além da identificação de suas principais causas gerais, a quantificação das chances de cada uma destas razões (regras) produzir um atraso no sistema brasileiro de aviação comercial.

Capítulo VII Conclusão

Neste trabalho foi realizada uma análise de atrasos e cancelamentos de voos comerciais brasileiros. A pesquisa foi iniciada a partir da construção de um mapa sistemático para avaliação da literatura mundial relacionada a este assunto. Com base nas principais publicações, foi construída a taxonomia das análises de atrasos e cancelamentos aéreos, dividida basicamente em questões relacionadas ao domínio, tais como tipo de problema e escopo, e ciência de dados, envolvendo fontes, dimensões e gestão de dados, bem como os métodos aplicados. Adicionalmente, foi gerada uma linha do tempo, indicando os principais tipos de problema, métodos e áreas de aplicação.

Após a identificação de uma lacuna no estudo de atrasos aéreos no Brasil, definiu-se como objetivo a análise das principais condições que conduzem a atrasos no país por meio da construção de um panorama geral. Para isso, optou-se pela utilização de uma metodologia de ciência de dados com a aplicação de regras de associação. Esta técnica é bastante útil para descobrir padrões ocultos e relacionamentos entre os dados, o que neste caso, representam as condições que conduzem a atrasos e cancelamentos aéreos no Brasil.

De modo a produzir uma análise mais completa, um *data warehouse* foi construído para integrar múltiplas fontes de dados, incluindo, além das informações dos voos, dados meteorológicos e uma relação de feriados. Além disso, com o objetivo de melhorar a qualidade dos resultados, foram executadas etapas de pré-processamento para selecionar os principais aeroportos e companhias aéreas, tratar discrepâncias e *outliers* e transformar atributos para um formato mais apropriado para o modelo. Neste estágio também foi criado um atributo para expressar o nível prévio de atrasos do aeroporto no momento da partida prevista de um voo.

As regras de associação foram geradas para comprimentos de dois a quatro atributos por regra. A confiança foi estabelecida como 0,2714, representando o percentual total de atrasos do *dataset*, enquanto o suporte foi definido como $3.131 \cdot 10^{-3}$, indicando uma ocorrência de aproximadamente quatro vezes por dia em um período de seis anos. De modo a avaliar a correlação e a chance de produção de novos atrasos, a discussão foi realizada de acordo com o *lift*, partindo das regras de comprimentos menores para aquelas com comprimentos maiores. Posteriormente, com o objetivo de aprofundar a compreensão do sistema de aviação comercial brasileiro, foram geradas análises complementares com base nos resultados encontrados anteriormente. Assim,

foram gerados novos conjuntos de regras de associação com suporte e confiança mínimos praticamente nulos, além de atributos específicos no LHS, de modo a permitir comparações entre anos, períodos do dia, companhias aéreas e aeroportos envolvidos no estudo.

De acordo com as análises realizadas, percebe-se que o desempenho do sistema brasileiro de aviação comercial tem melhorado nos últimos anos em relação à ocorrência de atrasos. Além disso, por representarem o momento de abertura dos aeroportos, os voos têm menos chances de partirem com atraso no início ou no meio da manhã. Adicionalmente, os resultados indicam que o Brasil sofre com a propagação de atrasos ao longo do dia, uma vez que, além dos horários mais propensos para a geração de novos atrasos serem o início e o fim da noite, os altos níveis de atraso já existentes tendem a produzir novos atrasos. Este cenário é encontrado para alguns aeroportos e, em especial, para uma companhia aérea, indicando que, possivelmente, suas operações possuem menos tempo de espera entre os voos. Neste caso, a reprogramação de alguns voos e aumento dos tempos de espera poderiam ser algumas alternativas. No entanto, é importante destacar que apesar desta tendência de propagação e geração de novos atrasos, seu desempenho melhorou significativamente nos últimos três anos.

Os picos de demanda também parecem trazer dificuldades a alguns aeroportos, especialmente em São Paulo, em que o dia da semana e o período do dia aparecem correlacionados à ocorrência de novos atrasos. Novamente, a reprogramação de alguns voos poderia ser uma alternativa. Entretanto, é importante manter a atratividade dos novos horários às companhias aéreas e passageiros, de modo a melhorar o *trade-off* entre capacidade e demanda nestes aeroportos. Por outro lado, os aeroportos do Rio de Janeiro parecem sofrer mais com momentos em que as condições de nebulosidade indicam que o céu está completamente encoberto. Neste caso, é necessária uma investigação mais profunda do problema, uma vez que este cenário pode estar sendo causado por limitações geográficas ou por questões relacionadas aos instrumentos e infraestrutura dos aeroportos.

Uma vez verificadas as principais causas de atrasos aéreos no Brasil, surgem algumas possibilidades para trabalhos futuros. Primeiramente, é possível aprofundar a análise, aplicando a mesma metodologia de regras de associação a níveis cada vez mais específicos e fixando atributos no LHS. Quando disponíveis, também é possível acrescentar novas dimensões e atributos, enriquecendo o escopo da análise. Além disso, as etapas de pré-processamento também podem ser aprimoradas, a partir da utilização de lógica *fuzzy* para o tratamento de atributos quantitativos, obtendo-se, assim, regras de associação *fuzzy*, conforme proposto por Kuok, Fu e Wong (1998).

Outra possibilidade é a utilização dos resultados das regras de associação em problemas de previsão de atrasos. Neste caso, os atributos mais relacionados aos atrasos devem ser incorporados ao modelo. Ao mesmo tempo pode ser avaliada a retirada dos atributos pouco relacionados,

o que poderia aumentar a eficiência do modelo e a qualidade dos resultados.

Adicionalmente, como a propagação de atrasos foi identificada como um dos principais motivos da ocorrência de novos atrasos, sua exploração também seria interessante. Neste caso, as relações entre os aeroportos em termos de propagação de atrasos poderiam ser verificadas. Isso significa entender quais aeroportos costumam ser os geradores dos atrasos propagados pela rede. Para isso, poderiam ser utilizadas regras de associação com base em janelas deslizantes de tempo ou uma metodologia de mineração sequencial.

Como conclusão desta pesquisa, é possível apontar que as regras de associação permitiram traçar um panorama geral das principais causas e tendências de atrasos aéreos no Brasil. Com base nos resultados encontrados, algumas medidas podem ser avaliadas de modo a reduzir a ocorrência de novos atrasos. Além disso, considerando a importância do setor de aviação comercial no país, há espaço para trabalhos futuros, que podem investigar causas em níveis mais específicos ou o relacionamento entre aeroportos em termos de propagação de atrasos. Finalmente, a metodologia aplicada pode ser estendida a outros sistemas aéreos, bem como utilizada para uma avaliação a partir do tempo para indicar melhorias e novas fraquezas do sistema.

Referências Bibliográficas

ABDEL-ATY, M.; LEE, C.; BAI, Y.; LI, X.; MICHALAK, M. Detecting periodic patterns of arrival delay. *Journal of Air Transport Management*, v. 13, n. 6, p. 355 – 361, 2007.

ABDELGHANY, K. F.; SHAH, S. S.; RAINA, S.; ABDELGHANY, A. F. A model for projecting flight delays during irregular operation conditions. *Journal of Air Transport Management*, v. 10, n. 6, p. 385 – 394, 2004.

AGRAWAL, R.; SRIKANT, R. Fast algorithms for mining association rules. In: *Proc. 20th int. conf. very large data bases, VLDB*. [S.l.: s.n.], 1994. v. 1215, p. 487–499.

AHMADBEYGI, S.; COHN, A.; GUAN, Y.; BELOBABA, P. Analysis of the potential for delay propagation in passenger airline networks. *Journal of Air Transport Management*, v. 14, n. 5, p. 221 – 236, 2008.

ANAC. *Agência Nacional de Aviação Civil*. [S.l.], 2015. Disponível em: <http://www.anac.gov.br-/>. Acesso em: 28 jun. 2015.

ANAC. *Anuário Estatístico do Transporte 2014*. [S.l.], 2015. Disponível em: <http://www2.anac-.gov.br/estatistica/anuarios.asp>. Acesso em: 02 fev. 2016.

BALAKRISHNA, P.; GANESAN, R.; SHERRY, L. Accuracy of reinforcement learning algorithms for predicting aircraft taxi-out times: A case-study of Tampa Bay departures. *Transportation Research Part C: Emerging Technologies*, v. 18, n. 6, p. 950 – 962, 2010.

BALL, M.; BARNHART, C.; NEMHAUSER, G.; ODONI, A. Chapter 1 Air Transportation: Irregular Operations and Control. In: BARNHART, C.; LAPORTE, G. (Ed.). *Transportation*. [S.l.]: Elsevier, 2007, (Handbooks in Operations Research and Management Science, v. 14). p. 1 – 67.

BEATTY, R.; HSU, R.; BERRY, L.; ROME, J. Preliminary evaluation of flight delay propagation through an airline schedule. *2nd USA/EUROPE AIR TRAFFIC MANAGEMENT R&D SEMINAR*, v. 7, n. 4, p. 259–270, 1998.

BHADRA, D. You (expect to) get what you pay for: A system approach to delay, fare, and complaints. *Transportation Research Part A: Policy and Practice*, v. 43, n. 9, p. 829 – 843, 2009.

BOSWELL, S. B.; EVANS, J. E. *Analysis of downstream impacts of air traffic delay*. [S.l.]: Lincoln Laboratory, Massachusetts Institute of Technology, 1997.

BOWMAN, A. W.; AZZALINI, A. *R package sm: Smoothing methods for nonparametric regression and density estimation (version 2.2-5.4)*. [S.l.], 2015. Disponível em: <https://cran.r-project.org-/web/packages/sm/sm.pdf>. Acesso em: 01 nov. 2015.

BRITTO, R.; DRESNER, M.; VOLTES, A. The impact of flight delays on passenger demand and societal welfare. *Transportation Research Part E: Logistics and Transportation Review*, v. 48, n. 2, p. 460 – 469, 2012.

BTS. *The Bureau of Transportation Statistics databases*. [S.l.], 2015. Disponível em: <http://www.rita.dot.gov/bts/home>. Acesso em: 31 mar. 2015.

BURDEN, R.; FAIRES, J.; BURDEN, A. *Numerical analysis*. [S.l.]: Nelson Education, 2015.

CULE, B.; GOETHALS, B.; TASSENOY, S.; VERBOVEN, S. Mining train delays. In: *Advances in Intelligent Data Analysis X*. [S.l.]: Springer, 2011. p. 113–124.

DAC. *Instrução de Aviação Civil 1223 - Normas para Confecção e Aprovação de Horario de Transporte - HOTRAN*. [S.l.], 2000.

DAC. *Instrução de Aviação Civil 1504 - Procedimentos para o Registro de Alterações em Voos de Empresas de Transporte Aéreo Regular*. [S.l.], 2000.

DüCK, V.; IONESCUB, L.; KLIEWERB, N.; SUHLA, L. Increasing stability of crew and aircraft schedules. *Transportation Research Part C: Emerging Technologies*, v. 20, n. 1, p. 47 – 61, 2012.

DHAR, V. Data science and prediction. *Communications of the ACM*, v. 56, n. 12, p. 64–73, 2013.

DOT. *The United States Department of Transportation*. [S.l.], 2015. Disponivel em: <http://www-.dot.gov/>. Acesso em: 31 mar. 2015.

ELMASRI, R.; NAVATHE, S. B. *Fundamentals of Database Systems*. 7 edition. ed. [S.l.]: Addison-Wesley, 2015.

EUROCONTROL. *Annual Network Operations Report 2014*. [S.l.], 2015. Disponível em: <https://www.eurocontrol.int/sites/default/files/publication/ performance/2014_annual/final-/annual_network_operations_report_2014_main_report_final_edition.pdf>. Acesso em: 02 fev. 2016.

FAA. *Federal Aviation Administration*. [S.l.], 2015. Disponível em: <http://www.faa.gov/>. Acesso em: 31 mar. 2015.

FLIER, H.; GELASHVILI, R.; GRAFFAGNINO, T.; NUNKESSER, M. Mining railway delay dependencies in large-scale real-world delay data. In: *Robust and Online Large-Scale Optimization*. [S.l.]: Springer, 2009. p. 354–368.

HAHSLER, M.; BUCHTA, C.; GRUEN, B.; HORNIK, K. *arules: Mining Association Rules and Frequent Itemsets*. [s.n.], 2015. Disponível em: <http://CRAN.R-project.org/package=arules>.

HAN, J.; CHENG, H.; XIN, D.; YAN, X. Frequent pattern mining: current status and future directions. *Data Mining and Knowledge Discovery*, v. 15, n. 1, p. 55–86, 2007.

HAN, J.; KAMBER, M.; PEI, J. *Data Mining: Concepts and Techniques, Third Edition*. 3 edition. ed. Waltham, Mass.: Morgan Kaufmann, 2011.

HAN, J.; PEI, J.; YIN, Y. Mining frequent patterns without candidate generation. In: *ACM SIGMOD Record*. [S.l.]: ACM, 2000. v. 29, p. 1–12.

HANSEN, M. Micro-level analysis of airport delay externalities using deterministic queuing models: a case study. *Journal of Air Transport Management*, v. 8, n. 2, p. 73 – 87, 2002.

HAO, L.; HANSEN, M.; ZHANG, Y.; POST, J. New York, New York: Two ways of estimating the delay impact of New York airports. *Transportation Research Part E: Logistics and Transportation Review*, v. 70, n. 0, p. 245 – 260, 2014.

HUNTER, G.; BOISVERT, B.; RAMAMOORTHY, K. Advanced national airspace traffic flow management simulation experiments and vlidation. In: *Simulation Conference, 2007 Winter*. [S.l.: s.n.], 2007. p. 1261–1267.

HUNTER, G.; RAMAMOORTHY, K. Evaluation of the national airspace system aggregate performance sensitivity. In: *Digital Avionics Systems Conference, 2007. DASC '07. IEEE/AIAA 26th*. [S.l.: s.n.], 2007. p. 1.E.1–1–1.E.1–13.

ICAO. *Annual Report of the Council 2014*. [S.l.], 2015. Disponível em: <http://www.icao.int-/annual-report-2014/Pages/default.aspx>. Acesso em: 16 fev. 2016.

INMET. *Instituto Nacional de Meteorologia*. [S.l.], 2015. Disponível em: <http://www.inmet.gov-.br/portal/>. Acesso em: 28 jun. 2015.

JAGADISH, H.; GEHRKE, J.; LABRINIDIS, A.; PAPAKONSTANTINOU, Y.; PATEL, J. M.; RAMAKRISHNAN, R.; SHAHABI, C. Big data and its technical challenges. *Communications of the ACM*, v. 57, n. 7, p. 86–94, 2014.

KIM, A.; HANSEN, M. Deconstructing delay: A non-parametric approach to analyzing delay changes in single server queuing systems. *Transportation Research Part B: Methodological*, v. 58, n. 0, p. 119 – 133, 2013.

KUOK, C. M.; FU, A.; WONG, M. H. Mining fuzzy association rules in databases. *ACM Sigmod Record*, ACM, v. 27, n. 1, p. 41–46, 1998.

LU, Z.; WANG, J.; XU, T. A new method for flight delays forecast based on the recommendation system. In: *Computing, Communication, Control, and Management, 2009. CCCM 2009. ISECS International Colloquium on*. [S.l.: s.n.], 2009. v. 1, p. 46–49.

LU, Z.; WANG, J.; ZHENG, G. A New Method to Alarm Large Scale of Flights Delay Based on Machine Learning. In: *Knowledge Acquisition and Modeling, 2008. KAM '08. International Symposium on*. [S.l.: s.n.], 2008. p. 589–592.

LV, X.; WANG, H. Flight Delay Alarming Analysis for an Airport Based on Markov. In: *Education Technology and Computer Science, 2009. ETCS '09. First International Workshop on*. [S.l.: s.n.], 2009. v. 1, p. 685–688.

MATSUDAIRA, K. The science of managing data science. *Communications of the ACM*, v. 58, n. 6, p. 44–47, maio 2015.

MITRA, S.; PAL, S. K.; MITRA, P. Data mining in soft computing framework: a survey. *IEEE transactions on neural networks*, v. 13, n. 1, p. 3–14, 2002.

MORRISON, S. A.; WINSTON, C. The effect of {FAA} expenditures on air travel delays. *Journal of Urban Economics*, v. 63, n. 2, p. 669 – 678, 2008.

MUELLER, E. R.; CHATTERJI, G. B. Analysis of aircraft arrival and departure delay characteristics. In: *AIAA aircraft technology, integration and operations (ATIO) conference*. [S.l.: s.n.], 2002.

NOAA. *National Oceanic and Atmospheric Administration*. [S.l.], 2015. Disponível em: <http://www.noaa.gov/>. Acesso em: 31 mar. 2015.

PAI, V. On the factors that affect airline flight frequency and aircraft size. *Journal of Air Transport Management*, v. 16, n. 4, p. 169 – 177, 2010.

PATHOMSIRI, S.; HAGHANI, A.; DRESNER, M.; WINDLE, R. J. Impact of undesirable outputs on the productivity of {US} airports. *Transportation Research Part E: Logistics and Transportation Review*, v. 44, n. 2, p. 235 – 259, 2008.

PEJOVIC, T.; NOLAND, R. B.; WILLIAMS, V.; TOUMI, R. A tentative analysis of the impacts of an airport closure. *Journal of Air Transport Management*, v. 15, n. 5, p. 241 – 248, 2009.

PostgreSQL GLOBAL DEVELOPMENT GROUP. *PostgreSQL*. [S.l.], 2015. Disponível em: <http://www.postgresql.org/>. Acesso em: 20 aug. 2015.

PYRGIOTIS, N.; MALONE, K. M.; ODONI, A. Modelling delay propagation within an airport network. *Transportation Research Part C: Emerging Technologies*, v. 27, n. 0, p. 60 – 75, 2013.

R CORE TEAM. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2014. Disponível em: <http://www.R-project.org/>. Acesso em: 25 aug. 2014.

REBOLLO, J. J.; BALAKRISHNAN, H. Characterization and prediction of air traffic delays. *Transportation Research Part C: Emerging Technologies*, v. 44, n. 0, p. 231 – 241, 2014.

REYNOLDS-FEIGHAN, A. J.; BUTTON, K. J. An assessment of the capacity and congestion levels at European airports. *Journal of Air Transport Management*, v. 5, n. 3, p. 113 – 134, 1999.

RYERSON, M. S.; HANSEN, M.; BONN, J. Time to burn: Flight delay, terminal efficiency, and fuel consumption in the National Airspace System. *Transportation Research Part A: Policy and Practice*, v. 69, n. 0, p. 286 – 298, 2014.

SCHAEFER, L.; MILLNER, D. Flight delay propagation analysis with the Detailed Policy Assessment Tool. In: *Systems, Man, and Cybernetics, 2001 IEEE International Conference on*. [S.l.: s.n.], 2001. v. 2, p. 1299–1303 vol.2.

SILBERSCHATZ, A.; KORTH, H.; SUDARSHAN, S. *Database System Concepts*. 6 edition. ed. New York: McGraw-Hill Science/Engineering/Math, 2010.

SIMIĆ, T. K.; BABIĆ, O. Airport traffic complexity and environment efficiency metrics for evaluation of ATM measures. *Journal of Air Transport Management*, v. 42, n. 0, p. 260 – 271, 2015.

SOOMER, M. J.; FRANX, G. J. Scheduling aircraft landings using airlines' preferences. *European Journal of Operational Research*, v. 190, n. 1, p. 277 – 291, 2008.

STERNBERG, A.; CARVALHO, D.; MURTA, L.; SOARES, J.; OGASAWARA, E. *Arules Package for R : Association Rules to Understanding Flight Delays*. [S.l.], 2016. Disponível em: <https://sourceforge.net/p/gpca/wiki/FlightDelayAssociationRules/>. Acesso em: 21 mar. 2016.

STORM PREDICTION CENTER. *Beaufort Wind Scale*. [S.l.], 2015. Disponível em: <http://www-.spc.noaa.gov/faq/tornado/beaufort.html>. Acesso em: 05 set. 2015.

THE UNITED STATES DEPARTMENT OF TRANSPORTATION. *Aviation Weather Formats: METAR/TAF*. [S.l.], 1999. Disponível em: <http://www.uscg.mil/auxiliary/missions/auxair-/metar_taf.pdf>. Acesso em: 05 set. 2015.

THE UNITED STATES DEPARTMENT OF TRANSPORTATION. *Air Travel Consumer Report 2014*. [S.l.], 2015. Disponível em: <http://www.dot.gov/airconsumer/air-travel-consumer-reports>. Acesso em: 02 fev. 2016.

TIAO, G. C. Asymptotic behaviour of temporal aggregates of time series. *Biometrika*, v. 59, n. 3, p. 525–531, dez. 1972.

TU, Y.; BALL, M. O.; JANK, W. S. Estimating flight departure delay distributions—a statistical approach with long-term trend and short-term pattern. *Journal of the American Statistical Association*, v. 103, n. 481, p. 112–125, 2008.

TWC. *The Weather Company*. [S.l.], 2015. Disponível em: <http://www.theweathercompany-.com/>. Acesso em: 31 mar. 2015.

VLACHOS, I.; LIN, Z. Drivers of airline loyalty: Evidence from the business travelers in China. *Transportation Research Part E: Logistics and Transportation Review*, v. 71, n. 0, p. 1 – 17, 2014.

WANG, P.; SCHAEFER, L.; WOJCIK, L. Flight connections and their impacts on delay propagation. In: *Digital Avionics Systems Conference, 2003. DASC '03. The 22nd.* [S.l.: s.n.], 2003. v. 1, p. 5.B.4–5.1–9 vol.1.

WIELAND, F. Limits to growth: results from the detailed policy assessment tool [air traffic congestion]. In: *Digital Avionics Systems Conference, 1997. 16th DASC., AIAA/IEEE.* [S.l.: s.n.], 1997. v. 2, p. 9.2–1–9.2–8 vol.2.

WONG, J.-T.; TSAI, S.-C. A survival model for flight delay propagation. *Journal of Air Transport Management*, v. 23, n. 0, p. 5 – 11, 2012.

WU, C.-L. Inherent delays and operational reliability of airline schedules. *Journal of Air Transport Management*, v. 11, n. 4, p. 273 – 282, 2005.

XIONG, J.; HANSEN, M. Modelling airline flight cancellation decisions. *Transportation Research Part E: Logistics and Transportation Review*, v. 56, n. 0, p. 64 – 80, 2013.

XU, N.; DONOHUE, G.; LASKEY, K. B.; CHEN, C.-H. Estimation of delay propagation in the national aviation system using Bayesian networks. In: *6th USA/Europe Air Traffic Management Research and Development Seminar.* [S.l.]: Citeseer, 2005.

ZAKI, M. J.; PARTHASARATHY, S.; OGIHARA, M.; LI, W.; others. New Algorithms for Fast Discovery of Association Rules. In: *KDD.* [S.l.: s.n.], 1997. v. 97, p. 283–286.

ZOU, B.; HANSEN, M. Flight delays, capacity investment and social welfare under air transport supply-demand equilibrium. *Transportation Research Part A: Policy and Practice*, v. 46, n. 6, p. 965 – 980, 2012.

ZOU, B.; HANSEN, M. Flight delay impact on airfare and flight frequency: A comprehensive assessment. *Transportation Research Part E: Logistics and Transportation Review*, v. 69, n. 0, p. 54 – 74, 2014.