Uma Análise da Propagação dos Atrasos de Voos por meio da Mineração de Sequência

Claudio Teixeira[1], **Gustavo Epifânio**[1], **Jefferson Colares**[1], **Raphael Fialho**[1]

[1]Centro Federal de Educacação Tecnológica Celso Suckow da Fonseca - CEFET/RJ

claudio_teix@hotmail.com

gustavo.p.epi@gmail.com

jcolares@gmail.com

raphael.fialho@eic.cefet-rj.br

Abstract. *Delays and cancellations are two of the top concerns in commercial aviation. Its impacts and high costs are felt by airports, airlines and especially by passengers. This work aims to identify frequent sequential patters with the application and comparison of the results obtained through cSPADE algorithm on a set of flight data and meteorological conditions of the airports, from the National Civil Aviation Agency (ANAC). In this way, it is sought to mitigate such occurrences and to improve the process of normalizing flights to an air transport system.*

Resumo. *Atrasos e cancelamentos são duas das principais preocupações na aviação comercial. Seus impactos e custos elevados são sentidos por aeroportos, companhias aéreas e principalmente, pelos passageiros. Este trabalho tem por objetivo, identificar padrões sequenciais frequentes com a aplicação e comparação dos resultados obtidos por meio do algoritmo cSPADE sobre um conjunto de dados de voos e condições meteorológicas dos aeroportos, provenientes da Agência Nacional de Aviação Civil (ANAC). Deste modo busca-se mitigar tais ocorrências e aperfeiçoar o processo de normalização dos voos para um sistema de transporte aéreo.*

1. Introdução

Atrasos de voos e suas propagações causam vários inconvenientes para companhias aéreas, aeroportos e passageiros. De acordo com os dados provenientes da Agência Nacional de Aviação Civil (ANAC), entre 2009 e 2015, 22% dos voos domésticos sofreu atraso em mais de 15 minutos. [Ogasawara et al., 2018]. Atrasos e interrupções em um aeroporto podem gerar congestionamento no espaço aéreo e ou impactar a operação em outros aeroportos, propagando atrasos nos voos de outras companhias aéreas. [Xu et al., 2005; Pyrgiotis et al., 2013]. A propagação de atrasos entre aeroportos é um tema importante e abordado principalmente na Europa e nos Estados Unidos, com aplicação de diversos métodos a fim de analisar esse fenômeno. Nesse contexto, a Mineração de Dados tem um importante papel na coleta e análise de grande volume de dados com o objetivo de descobrir informações e conhecimento que subsidiarão uma melhor tomada de decisão pelas organizações que atuam nesses processos para mitigar esse problema.

Isso posto, a Mineração de Dados é uma das fases do processo de descoberta de conhecimento e consiste na aplicação de análise de dados e descoberta de algoritmos com o intuito de produzir uma lista de padrões (ou modelos) sobre os dados.

A análise de dados é realizada nas etapas de coleta de dados e pré-processamento de dados, enquanto que a descoberta dos algoritmos e geração dos padrões faz parte da etapa de pós-processamento dos resultados provenientes da mineração. [Fayyad et al., 1996].

Sob o ponto de vista da fase de mineração de dados, é preciso saber que não será possível obter bons resultados se dois pré-requisitos não forem atendidos: (a) o analista de dados precisa conhecer o contexto em que os dados estão inseridos e como eles ocorrem nesse contexto; (b) o analista de dados precisa executar procedimentos que tornem o conjunto de dados o mais adequado possível para a etapa de mineração de dados.

A Mineração de Dados se faz muito útil quando a quantidade de dados disponível é grande e representativa – motivo pelo qual a fase de coleta de dados e a tarefa de amostragem são muito importantes no processo de descoberta de conhecimento.

Considerando-se este cenário, uma técnica bastante conhecida para extração de conhecimento é a mineração de Padrões Sequenciais Frequentes ou Mineração de Sequência, utilizada principalmente para procurar padrões sequenciais frequentes em uma grande quantidade de dados. A descoberta de novos e diferenciados padrões sequenciais frequentes, possibilita um novo e interessante conhecimento sobre os dados.

Vislumbra-se por meio deste trabalho, a descoberta de padrões sequenciais frequentes e suas respectivas regras de associação entre os voos atrasados e suas propagações por meio da aplicação do algoritmo cSPADE com o objetivo de responder as seguintes questões: (i) quais regras de atrasos e propagações desses atrasos pode-se descobrir em um aeroporto específico?, (ii) quais regras de atrasos e propagações desses atrasos pode-se descobrir entre os aeroportos? (iii) quais fatores estão associados a estas regras em geral e que impactam a operação como um todo? e (iv) quais proposições são sugeridas para mitigar-se esses problemas?

Além dessa introdução, o trabalho divide-se em mais quatro seções. A Seção 2 descreve a etapa de preprocessamento dos dados. A Seção 3 descreve a função de mineração de dados. A metodologia propriamente dita é apresentada na Seção 4 e por fim, a avaliação experimental e os resultados encontrados são analisados na Seção 5, encerrando este trabalho.

2. Preprocessamento dos Dados

A amostra de dados disponível para análise pode conter uma série de imprecisões e desvios ou pode estar representada de maneira inadequada. Exemplos de erros comuns incluem valores ausentes, erros de digitação, formatos mistos, entradas replicadas da mesma entidade do mundo real, e violações de regras de negócios. [Chu et al., 2016]. Esses fatores influenciam negativamente qualquer tipo de análise de dados, e, portanto, estratégias de pré-processamento de dados amenizam tais efeitos negativos somados ao contexto da exploração de dados. Nos últimos anos, tem havido um enorme interesse em diferentes aspectos da limpeza de dados, incluindo novas abstrações, interfaces, abordagens para técnicas de escalabilidade e crowdsourcing.

Os conceitos de estatística descritiva são úteis para o planejamento de estratégias de preprocessamento, pois suportam a verificação da presença de ruídos (atributos cujo conjunto de valores varia acima das extremidades dos quartis), a necessidade de transformação de valores (com uso da média e desvio-padrão) ou a utilidade da seleção de dados ou atributos (com uso de medidas de correlação), por exemplo. No contexto de seleção de dados ou atributos, podem-se mencionar as técnicas amplamente utilizadas: Information Gain Attribute Ranking, Relief, Principal Component Analysis (PCA), Correlation-based Feature Selection (CFS) e Wrapper Subset Evaluation. [Hall, Holmes, 2003]. Destaca-se também a utilidade da análise de frequência na preparação de conjuntos de dados, assim como, faz-se interessante saber que medidas de tendência central, dispersão e outras são importantes para análises e comparação de resultados.

Um dos principais fatores diferenciadores é como definir um erro de dado (ou seja, detecção de um erro de dado). As técnicas quantitativas, amplamente utilizadas para detecção de outliers, empregam métodos estatísticos para identificar comportamentos e erros anormais (por exemplo, "um salário que tem três desvios-padrão de distância do salário médio é um erro"). Por outro lado, técnicas qualitativas usam restrições, regras e padrões para detectar erros (por exemplo, "Não podem existir dois funcionários do mesmo nível com salários distintos". Uma vez que os erros são detectados, o reparo pode ser executado usando scripts, um grande grupo de especialistas, ou um híbrido de ambos. [Chu et al., 2016]. Os próximos parágrafos descrevem as atividades de limpeza e transformação de dados: Remoção de outliers (subseção 2.1), Remoção de outras inconsistências (subseção 2.2), Criação de variáveis (subseção 2.3) e Agregação temporal (subseção 2.4), utilizados neste trabalho.

2.1. Remoção de outliers

Sob o ponto de vista do cenário dos atrasos dos voos comerciais, dados inconsistentes, incorretos ou discrepantes, como o horário previsto de partida ou o horário real de partida seria um grande problema, pois é na diferença entre esses horários que o atraso é calculado.

Logo, é importante tratar esses dados para evitar influências negativas nos resultados. O primeiro passo do processo de limpeza de dados é encontrar valores discrepantes [Han et al., 2011].

Valores discrepatantes (ou outliers) são dados que possuem valores atípicos em relação às demais observações da base de dados. Essas discrepâncias podem ser ocasionadas por diversos motivos como uma modelagem ruim dos dados ou erro humano na entrada do dado.

Uma maneira comum de lidar com as tarefas de limpeza é remover da base de dados as tuplas que contém discrepâncias [Han et al., 2011].

Para este trabalho, considera-se os outliers como sendo os valores negativos dos atrasos dos voos nas partidas e nas chegadas, assim como os valores de atraso acima de 240 minutos (ou 4 horas) visto que o número de voos atrasados após 4 horas não é relevante e poderia ser removido do conjunto de dados sem comprometer a representatividade das observações, conforme mostra a Figura 1 abaixo.

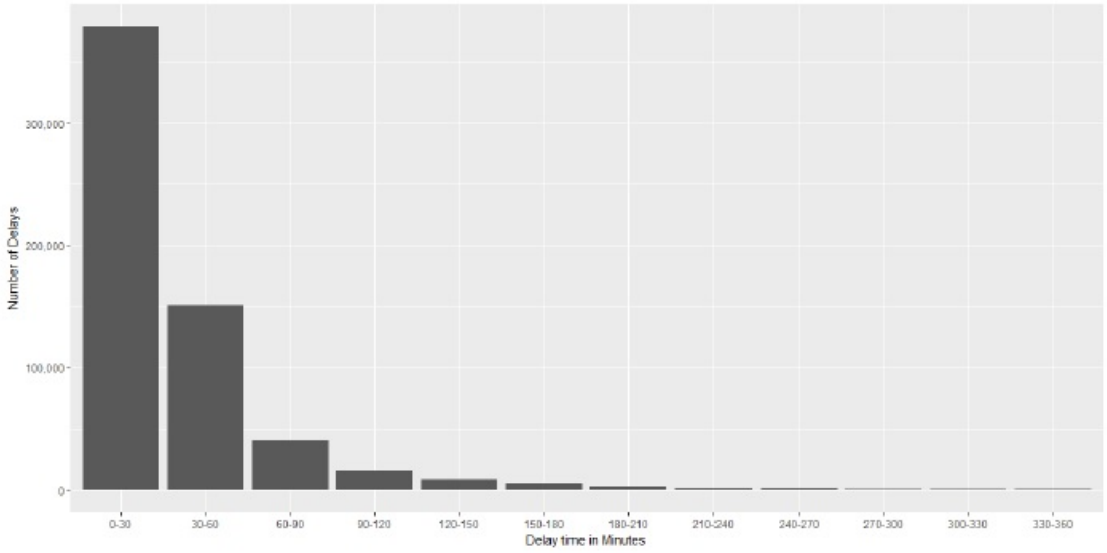

Figura 1. Number of Delayed Flights

2.2. Remoção de outras inconsistências

Outras inconsistências que foram removidas durante a avaliação experimental deste trabalho, foram as datas das partidas dos voos que eram maiores do que as datas das chegadas desses mesmos voos e as observações em que havia para um mesmo número de voo, partidas e ou chegadas em aeroportos distintos. Para solucionar este problema, realizou-se uma agregação pelo número do voo, seguida de uma indexação também pelo número do voo, selecionando-se o maior valor por meio da função max() para a hora cheia do valor da data de partida e ou chegada desse mesmo voo.

2.3. Criação de variáveis

Com o objetivo de facilitar o processo de criação do arquivo com os dados necessários no formato esperado pelo algoritmo cSPADE e a partir das datas das partidas e chegadas dos voos, ambas no formato data/hora, foram criadas as variáveis: datas de partidas e chegadas dos voos, no formato data (AAAAMMDD) e as respectivas horas de partidas e chegadas dos voos, no formato (HH) ignorando-se os minutos e os segundos.

2.4. Agregação temporal

Uma agregação de atributos é uma técnica comum e muito usada em conjuntos de dados que serão minerados. Consiste em agregar atributos contínuos por similaridade ou proximidade em zonas, intervalos ou valores discretos. Um exemplo é uma agregação temporal [TIAO, 1972] que consiste na transformação de valores contínuos de timestamps (com horas, minutos e segundos) em valores discretos de hora. Outro exemplo é o processo de binning, que leva muitos valores contínuos e coloca-os em intervalos.

Nesta parte do trabalho, transforma-se a série de voos em uma série temporal com observações hora a hora com os voos agrupados pela concatenação do código ICAO do aeroporto mais a data de realização da partida ou da chegada, ordenados por esse agrupamento e pela agregação temporal considerando-se a hora cheia. Por exemplo, se um voo estimado para chegar às 13:54, será concatenado na 13a. hora, em formato string,

para a data específica de chegada. Calcula-se o tamanho de cada agrupamento por meio do somatório dos voos em cada agrupamento.

3. Função de Mineração de Dados

A mineração de padrões sequenciais descobre sequências e subsequências frequentes em um conjunto de dados indexado e tem sido um dos principais focos de pesquisa em mineração de dados na última década. No entanto, também é um problema desafiador, pois a mineração pode gerar ou examinar um elevadíssimo número de combinações de subsequências intermediárias. A literatura é abundante e progressos recorrentes acontecem até hoje. [Aggarwal et al., 2014]

As sequências e subsequencias descobertas pelos algoritmos possuem uma frequência não inferior a um limite especificado pelo usuário, chamado de suporte. O conjunto de dados indexado, armazena eventos que estão ou não relacionados com o tempo. Para este conjunto dá-se o nome de transações. Uma subsequência, se ocorrer com frequência, em um conjunto de transações, é um padrão sequencial frequente.

Semelhante à mineração de regras de associação [Agrawal et al., 1993] a mineração de padrões sequenciais foi inicialmente motivada pelo problema de suporte à decisão na indústria do varejo e foi abordada pela primeira vez por Agrawal e Srikant em [Agrawal et al., 1995]. Esse problema foi definido da seguinte forma: Dado um conjunto de sequências, onde cada seqüência consiste em uma lista de elementos e cada elemento consiste em um conjunto de itens, e dado um limite min_support especificado pelo usuário, o objetivo da mineração de padrões sequenciais é localizar todas as subsequências frequentes, ou seja, as subsequências cuja freqüência de ocorrência no conjunto de seqüências não é menor que min_support. [Agrawal et al., 1995].

Os algoritmos de mineração de padrões sequenciais podem ser categorizados em duas classes: abordagens baseadas no algoritmo Apriori [Agrawal et al., 1995] e algoritmos de crescimento de padrões [Pei et al., 2001].

Para este trabalho, escolheu-se o algoritmo mais recente no contexto da abordagem baseada no algoritmo Apriori que é o algoritmo cSPADE. Até a descoberta do algoritmo cSPADE (combinatorial Sequential Pattern Discovery using Equivalence classes), as soluções existentes faziam varreduras repetidas no conjunto de dados e usavam estruturas de hash complexas para localização das sequências.

O cSPADE utiliza propriedades combinatórias para decompor o problema original em subproblemas menores, que pode ser independentemente resolvido na memória principal usando técnicas eficientes de pesquisa de rede, e usando junção simples nas operações. Todas as sequências são descobertas em apenas três varreduras no conjunto de dados. [Zaki, 2001].

Neste trabalho, o conjunto de dados encontra-se no formato single, onde todas as tuplas de observações estão normalizadas representando cada evento na sua menor cardinalidade, ou seja, quando um conjunto de dados está neste formato, cada observação representa um único item e cada item contém um identificador de transação.

Para utilizarmos o algoritmo cSPADE, necessita-se de um novo conjunto de dados, transformado do original, onde apenas as colunas necessárias para a criação de um arquivo TXT (flat file) estarão em um formato chamado basket. Quando um arquivo

está no formato basket cada observação representa uma transação onde os itens estão em uma mesma coluna, ou seja, para a entrada do algoritmo cSPADE, tem-se uma lista de transações, onde cada transação consiste em um id de sequência, um id de evento e uma lista de itens. O ID da sequência identifica a seqüência à qual a transação pertence. O event-id pode ser um timestamp ou uma identificação temporal a essa transação nessa sequência. A lista de itens é um conjunto dos itens dessa transação.

4. Metodologia

A metodologia adotada contemplou uma pesquisa bibliográfica para a abordagem, tratamento do objeto e sua fundamentação por meio da leitura e análise de livros, artigos e teses. Utilizou-se também de um estudo de caso particular e representativo para fundamentar a teoria e a prática em relação aos resultados esperados. Isso posto, contemplou-se a metodologia baseada em duas atividades principais: pré-processamento e geração de regras. Por meio de análise exploratória dos dados e a aplicação de funções de pré-processamento sobre o conjunto de dados obtido pela junção de duas bases de dados: Voos Regulares Ativos (VRA) obtido por meio de exportação das informações contidas no Sistema Integrado de Informações da Aviação Civil (SINTAC) disponibilizado pela Agência Nacional de Aviação Civil (ANAC) e condições meteorológicas (WU) coletados a partir dos dados disponibilizados pela empresa Weather Underground em seu website, criou-se um Jupyter Notebook implementado por meio do pacote estatístico R. Pode-se representar a metodologia adotada neste trabalho, por meio de um pseudo-algoritmo que inicia-se com a função gerarSequencias() que recebe um conjunto de dados como parâmetro. Nesta função, um sub-conjunto de dados recebe o retorno da função preProcessar() que passa por parâmetro, o conjunto de dados inicial e retorna os voos atrasados em 15 minutos ou mais e todos os tratamentos comentados no item de Preprocessamento, inclusive o método de agregação temporal. Ao final, um conjunto de transações estarão prontas por meio da geração de um arquivo TXT (voosAtrasados) que será carregado na função gerarRegras() que retornará as regras induzidas para interpretação e análise. A figura 2 abaixo, mostra a estrutura do pseudo-algoritmo.

Pseudo-Algoritmo: MINERACAOSEQUENCIAL():

```
1: function GERARSEQUENCIAS(dataSet)
2:     voosAtrasados <- PREPROCESSAR(dataSet)
3:     return GERARREGRAS(voosAtrasados)

1: function PREPROCESSAR(dataSet)
2:     voosAtrasados <- REMOVEROUTLIERS(voosAtrasados)
3:     return AGREGACAOTEMPORAL(voosAtrasados)

1: function GERARREGRAS(voosAtrasados)
2:     regras <- cSPADE(voosAtrasados)
3:     return INDUZIRREGRAS(regras)
```

Figura 2. Pseudo-Algoritmo

5. Avaliação Experimental

Analisando-se as regras geradas, pode-se tentar responder as primeiras questões sugeridas e entender um pouco mais sobre o cenário como um todo.

Inicialmente, os padrões sequenciais frequentes foram gerados com um suporte mínimo de 50% (ou seja, a sub-sequência ocorre em no mínimo 2 sequências de entrada).

Além disso, os experimentos subdividiu-se em duas partes:

(a) geração de regras para um aeroporto específico e; (b) geração de regras entre aeroportos.

Para o caso (a), determinou-se o aeroporto de Belém-PA com as partidas e chegadas em atraso no dia primeiro de janeiro de 2017.

Considerando-se um suporte de 50% e confiança de 95%, obteve-se as seguintes regras:

(i) O atraso do voo 1679 implicou no atraso dos voo 3796; (ii) O atraso do voo 1679, implicou no atraso conjunto dos voos 3233 e 3796; e (iii) O atraso do voo 1679 implicou no atraso do voo 3233;

Logo, considerando-se um lift maior ou igual 1 tem-se uma correlação positiva entre os voos acima e por conseguinte, as regras indicam uma propagação de atraso entre esses voos.

Ou seja:

A partida do voo 1679 com origem Belém-PA e destino Guarulhos-São Paulo(SP) sob as condições climatológicas de trovoada atrasou 38 minutos e propagou esse atraso na partida do voo 3233 com origem Belém-PA e destino Brasília-DF com atraso total de 123 minutos, assim como, na chegada do voo 3796 vindo de Brasília-DF com atraso total de 119 minutos. Isso posto, tem-se também uma propagação de atraso entre os aeroportos de Belém-PA e Brasília-DF.

Para o caso (b), induziu-se a geração das regras de atraso entre os aeroportos de Guarulhos-São Paulo(SP) e Confins-Belo Horizonte(MG).

E assim como no caso (a), contemplou-se as partidas e chegadas em atraso e, desta vez, para o dia 05 de janeiro de 2017.

Para a execução do algoritmo cSPADE, além dos parâmetros utilizados no caso (a), determinou-se também os parâmetros: maxsize = 2 (dois itens em um elemento de uma sequência) e maxlen = 2 (dois elementos de uma sequência). Isso posto, 952 regras foram geradas.

Para fins de análise, selecionou-se dois voos do tipo ”ida e volta”em que a probabilidade de um atraso neste tipo de rota, impacta não somente esta rota, mas também outras rotas em que a operação possa ser comprometida entre esses aeroportos.

Analisando-se as regras geradas para os voos 3344 e 3345, conclui-se que o voo 3344 partiu de Guarulhos-São Paulo(SP) com atraso de 21 minutos com destino a Confins-Belo Horizonte(MG) e esse atraso propagou-se para os voos, conforme mostrado na Figura 3.

Partidas

#Número do Voo	#Origem	#Destino
4120	Belo Horizonte-Confins-MG(SBCF)	Campinas-Vira Copos-SP(SBKP)
3378	Belo Horizonte-Confins-MG(SBCF)	Recife-PE(SBRF)
3326	Guarulhos-SP(SBGR)	Belo Horizonte-Confins-MG(SBCF)
2185	Belo Horizonte-Confins-MG(SBCF)	Galeão-RJ(SBGL)
1913	Belo Horizonte-Confins-MG(SBCF)	Galeão-RJ(SBGL)
1703	Belo Horizonte-Confins-MG(SBCF)	Brasília-DF(SBBR)

Chegadas

#Número do Voo	#Origem	#Destino
1833	Salvador-BA(SBSV)	Belo Horizonte-Confins-MG(SBCF)
1700	Brasília-DF(SBBR)	Belo Horizonte-Confins-MG(SBCF)

Figura 3. Propagação dos Atrasos

Pode-se também visualizar esta propagação de atrasos, por meio de uma análise de rede, gerando-se um grafo, conforme mostra a Figura 4.

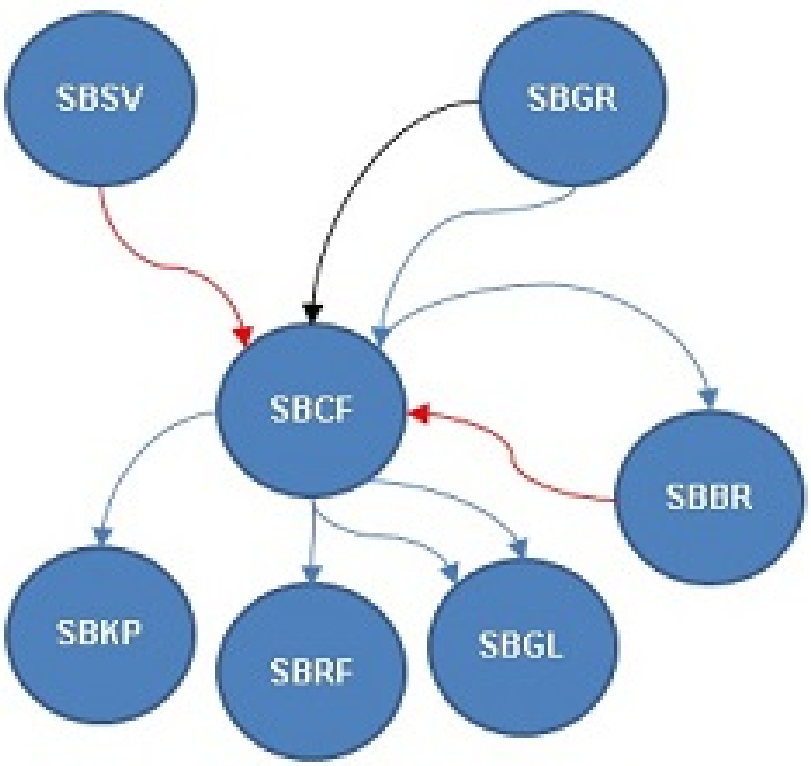

Figura 4. Grafo de Propagação dos Atrasos

Aplicando-se técnicas de indexação de dados combinadas com regras de associação, induzidas após a descoberta de padrões sequenciais frequentes, pode-se revelar padrões escondidos de atrasos de voos e suas propagações.

Considerando-se o conjunto de dados dos voos domésticos e guiando-se pelas perguntas formuladas relacionadas às causas, momentos, diferenças e relacionamentos entre aeroportos, pode-se avaliar e quantificar atributos que poderiam estar relacionados aos atrasos, mostrando-se não somente os principais padrões, mas também um subconjunto de ocorrências de propagações de atrasos na rede e em alguns aeroportos.

Referências

Aggarwal, C. and Han, J., (2014), "Frequent Pattern Mining.", Springer.

Agrawal, R., Imielinski, T. and Swami, A., (1993), "Mining association rules between sets of items in large databases," in ACM SIGMOD conference, pp. 207–216.

Agrawal, R. and Srikant, R., (1995), "Mining sequential patterns," in ICDE Conference, pp. 3–14.

Chu, X., Ilyas, I., Krishnan, S., Wang. J., (2016), "Data Cleaning: Overview and Emerging Challenges", SIGMOD'16, June 26-July 01, 2016, San Francisco, CA, USA.

Fayyad et al., (1996), "Advances in Knowledge Discovery and Data Mining.". AAAI Press.

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

Ogasawara, E., Soares, J., Moreira, L., Oliveira, L. e Dantas, C., (2018), "On Evaluating Data Preprocessing Methods for Machine Learning Models for Flight Delays.", International Joint Conference on Neural Networks (IJCNN) 2018.

Pei, J., Han, J., Mortazavi-asl, B., Pinto, H., Chen, Q., Dayal, U. and Chun Hsu, M., (2001), "Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth," in ICDE Conference, pp. 215–224.

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

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

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

Zaki, M., (2001), "SPADE: An Efficient Algorithm for Mining Frequent Sequences".