Journal Title
XX(X):1–8

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

An Analysis of Flight Delay Propagation through Sequence Mining and Organization Studies

Claudio Teixeira[1]

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 the (combinatorial Sequential Pattern Discovery using Equivalence classes) 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.

Keywords

Sequence Mining, Flight Delay, Air Transportation, Organization Studies

1. Introduction

Flight delays and their propagation cause various inconveniences for airlines, airports and passengers.

According to data from the National Civil Aviation Agency (ANAC), between 2009 and 2015, 22% of domestic flights delay in more than 15 minutes. [Ogasawara et al., 2018].

Delays and interruptions in one airport may cause congestion in the air space and / or impact the operation in other airports, causing delays in the flights of other airlines. [Xu et al., 2005; Pyrgiotis et al., 2013].

The spread of airport delays is an important issue and mainly addressed in Europe and the United States, with application of different methods in order to analyze this phenomenon.

In this context, the Data Mining has an important role in the collection and analysis of large volumes of data with the objective to discover information and knowledge that will support a better decision-making organizations that work on these processes to mitigate this problem.

That said, Data Mining is one of the phases of the knowledge and consists of the application of data analysis and discovery of algorithms with in order to produce a list of patterns (or models) on the data.

Data analysis is performed at the data collection and preprocessing stages while the discovery of the algorithms and generation of the patterns is part of the step of processing the results from mining. [Fayyad et al. 1996].

From the point of view of the data mining phase, we must know that it will not be possible if two prerequisites are not satisfied:

(a) the data needs to know the context in which the data are inserted and how they occur in this context;

(b) the data analyst must perform procedures that make the data set as suitable as possible for the data mining step.

Data Mining becomes very useful when the amount of data available is large and representative - which is why the data collection phase and the sampling task are very important in the process of knowledge discovery.

Considering this scenario, a technique known for extraction knowledge and the mining of Frequent Sequential Patterns or Sequence Mining, used mainly to look for frequent sequential patterns in a large amount of data.

The discovery of new and differentiated sequential patterns a new and interesting knowledge about the data.

Through this work, the finding of frequent sequential patterns is and their respective rules of association between delayed flights and their propagation through the application of the cSPADE algorithm in order to respond to the following:

(i) what rules of delay and propagation of such delays can be found in specific airport?

(ii) what rules of delay and propagation of such delays to discover among the airports?

(iii) what factors are associated with these rules in impact of the operation as a whole? and

(iv) which propositions are suggested to mitigate these problems? Besides this introduction, the work is divided into four more sections.

The section 2 describes the data preprocessing step. Section 3 describes the function of mining of data.

The methodology itself is presented in Section 4 and Finally, the experimental evaluation and the results obtained are analyzed in Section 5, closing this work.

2. Data Preprocessing

The data sample available for analysis may contain a series of inaccuracies and deviations or may be poorly represented.

[1]Federal Center for Technological Education (CEFET-RJ)

Corresponding author:
Claudio Teixeira, CEFET-RJ
Email: claudio_teix@hotmail.com

Examples of common errors include missing values, typos, mixed formats, replicated entries of the same real-world entity, and violations of business rules. [Chu et al., 2016].

These factors negatively influence any type of data analysis, and therefore, pre-processing strategies alleviate such negative effects in addition to the context of data exploration. In the last few years, there has been a different aspects of data cleansing, including new abstractions, interfaces, approaches for scalability and crowdsourcing techniques.

The concepts of descriptive statistics are useful for the planning of strategies preprocessing, since they support the verification of the presence of noise (attributes whose set of values varies above the quartile ends), the need for transformation of values (using the mean and standard deviation) or the utility of the selection of data or attributes (with use of correlation measures), for example.

In context of selection of data or attributes, we can mention the techniques widely used: Information Gain Attribute Ranking, Relief, Principal Component Analysis (PCA), Correlation-based Feature Selection (CFS) and Wrapper Subset Evaluation. [Hall, Holmes, 2003].

Also noteworthy is the utility of the frequency analysis in the preparation of sets of data, as well as, it is interesting to know that measures of central tendency, and others are important for analysis and comparison of results.

One of the main differentiating factors is how to define a data error (or detection of a data error).

The quantitative techniques, widely used for detection of outliers, use statistical methods to identify behaviors and abnormal errors (for example, ”a salary that has three standard deviations of distance of the average salary is an error ”). On the other hand, qualitative techniques use constraints, rules and patterns to detect errors (for example, ”There can not be two functionaries of the same level with different salaries ”.

Once the errors are detected, the can be run using scripts, a large group of experts, or a hybrid of both. [Chu et al., 2016].

The following paragraphs describe cleaning activities and transformation of data: Removal of outliers (subsection 2.1), Replacement of other inconsist (Subsection 2.2), Creation of Variables (Subsection 2.3) and Temporal Aggregation (subsection 2.4), used in this work.

2.1. Removal of outliers

From the point of view of the scenario of commercial flight delays, inconsistent data, incorrect or discrepancies, such as the estimated time of departure or the actual time of departure would be a major problem, since ’and in the difference between these schedules the calculated. Therefore, it is important to treat such data to avoid negative influences on results.

The first step in the data cleansing process is to find discrepant values [Han et al., 2011].

Discrepant values (or outliers) are data that have atypical values in to the rest of the database.

These discrepancies may be for various reasons such as poor modeling of data or human error in the input.

A common way of handling cleaning and removal tasks from the given the tuples that contain discrepancies [Han et al., 2011].

For this work, the outliers are considered to be the negative values of the delays in departures and arrivals, as well as delay values above 240 minutes (or 4 hours) since the number of flights delayed after 4 hours is not relevant and could be removed from the data set without compromising representativeness observations, as shown in **Figure 1**.

2.2. Removal of other inconsistencies

Other inconsistencies that were removed during the experimental evaluation of this work, it was the dates of flight departures that were greater than the dates of arrivals of the same flights and the observations in which there were for the same number flights, departures and arrivals at different airports. To resolve this issue, an increase was made by the number of the flight, followed by an indexation by the number of the flight, by selecting the highest value by means of the max () function for the full hour of the value of the date of departure and / or arrival of that same flight.

2.3. Variable creation

In order to facilitate the process of creating the file with the necessary dataIn the format expected by the cSPADE algorithm and from the dates of the and flight arrivals, both in the date / time format, the variables were created: dates of flight departures and arrivals, in the date format (YYYYMMDD) and the respective flight departures and arrivals in the (HH) format, ignoring the minutes and seconds.

2.4. Temporal aggregation

An aggregation of attributes is a common technique and widely used in data sets that will be mined. It consists of adding attributes by continuity or proximity to zones, ranges or discrete values. An example is a temporal aggregation [TIAO, 1972] which consists of transforming continuous values of timestamps (with hours, minutes and seconds) in discrete hour values. Another example is the binning process, which takes many continuous values and places them in intervals.

In this part of the work, the series of flights becomes a time series hourly observations with flights grouped by the concatenation of the ICAO airport code plus the date of departure or arrival ordered by that grouping and temporal aggregation considering full time. For example, if a flight estimated to arrive at 13:54, it will be concatenated on the 13th. hour, in string format, for the specific date of arrival. The size of each grouping is calculated by the sum of the flights in each grouping.

3. Data Mining Function

Sequential pattern mining discovers frequent sequences and subsequences in an indexed dataset and has been a major focus of data mining research in the last decade. However, it is also a challenging problem, since mining can generate or examine a very large number of combinations

of intermediate subsequences. Literature is abundant and recurrent progress until today. [Aggarwal et al., 2014].

The sequences and subsequences discovered by the algorithms have a frequency not less than a user-specified threshold, called the carrier. The indexed dataset stores events that are time. For this set is the name of transactions. A subsequence, if it occurs frequently, in a set of transactions, is a frequent sequential pattern.

Similar to association rule mining [Agrawal et al., 1993] a sequential pattern mining was initially motivated by the decision support problem in the retail industry and was first addressed by Agrawal and Srikant in [Agrawal et al., 1995].

This problem was defined as follows: Given a set of sequences, where each sequence consists of a list of elements and each element consists of a set of items, and given a user-specified min-limit, the goal of sequential pattern mining is to locate all frequent subsequences, that is, subsequences whose occurrence frequency in the sequence set is not less than min_support. [Agrawal et al., 1995].

Sequential pattern mining algorithms can be categorized into two classes: approaches based on the Apriori algorithm [Agrawal et al., 1995] and pattern growth algorithms [Pei et al., 2001].

For this work, we chose the most recent algorithm in the context of the approach based on the Apriori algorithm, which is the cSPADE algorithm. Until the discovery of the cSPADE algorithm, the existing solutions made repeated scans in the dataset and used complex hash structures to locate the sequences.

CSPADE uses combinatorial properties to decompose the original problem into smaller subproblems, which can be independently solved in main memory using efficient network search techniques, and using simple joining in operations. All sequences are discovered in only three scans in the data set. [Zaki, 2001].

In this work, the data set is in the single format, where all tuples of observations are normalized representing each event at its lowest cardinality, that is, when a set of data is in this format, each observation represents a single item and each item contains a transaction identifier.

To use the cSPADE algorithm, we need a new data set, transformed from the original, where only the columns necessary to create a text file (.txt), also called flat file, will be in a format called basket. When a file is in the basket format each observation represents a transaction where the items are in the same column, that is, for the input of the cSPADE algorithm, there is a list of transactions, where each transaction consists of a sequence id, a event id and a list of items.

The sequence ID identifies the sequence to which the transaction belongs. O event-id can be a timestamp or a temporary ID to that transaction in that sequence.

The item list is a set of items in this transaction.

4. Methodology

The adopted methodology contemplated a bibliographical research for the approach, treatment of the object and its foundation by means of the reading and analysis of books, articles and theses.

A particular and representative case study was also used to substantiate theory and practice in relation to the expected results.

That said, the methodology was based on two main activities: preprocessing and rule generation.

By means of exploratory data analysis and the application of preprocessing functions on the data set obtained by the combination of two databases: Active Regular Flights (VRA) obtained through the export of information contained in the Integrated Civil Aviation Information System (SINTAC) provided by the Agency National Civil Aviation (ANAC) and meteorological conditions collected from the data provided by the company Weather Underground (WU) on its website, a Jupyter Notebook was implemented through the statistical package R.

One can represent the methodology adopted in this work, by means of a pseudo-algorithm that starts with the functionSequences () function that receives a dataset as a parameter.

In this function, a sub-set of data receives the return of the preProcess () function that passes by parameter, the initial data set and returns the delayed flights in 15 minutes or more and all the treatments commented on in the Preprocessing item, including the method of temporal aggregation.

In the end, a set of transactions will be ready by generating a text file (Late flights) that will be loaded through the generateRules () function that will return the rules induced for interpretation and analysis.

The **Figure 2** shows the structure of the pseudo-algorithm.

5. Experimental Evaluation

By analyzing the rules generated, one can try to answer the first questions suggested and to understand a little more about the scenario as a whole.

Initially, frequent sequencing patterns were generated with a minimum support of 50% (ie, the substring occurs in at least 2 input sequences).

In addition, the experiments were subdivided into two parts:

(a) generation of rules for a specific airport; and

(b) generation of rules between airports.

In case (a), the airport of Belm-PA was determined with departures and arrivals in arrears on January 1, 2017.

Considering a support of 50% and confidence of 95%, the following rules were obtained, as shown in **Figure 3**.

(i) The delay of flight 1679 implied the delay of flight 3796;

(ii) The delay of flight 1679, resulted in the joint delay of flights 3233 and 3796; and

(iii) The delay of flight 1679 implied the delay of flight 3233;

Therefore, considering a lift greater than or equal to 1, there is a positive correlation between flights above and therefore the rules indicate a delay spread between those flights.

That is:

The departure of flight 1679 with origin Belm-PA and destination Guarulhos-So Paulo (SP) under the weather conditions of thunderstorm delayed 38 minutes and propagated this delay in the departure of flight 3233 from

Belm-PA and destination Braslia-DF, with a total delay of 123 minutes, as well as the arrival of flight 3796 from Braslia-DF with total delay of 119 minutes. That said, there is also a delay spread between the airports of Belm-PA and Braslia-DF.

For case (b), the delay rules were induced between Guarulhos-So Paulo (SP) and Confins-Belo Horizonte (MG) airports.

And just as in case (a), the departures and arrivals were delayed, and this time, for January 5, 2017.

For the execution of the cSPADE algorithm, in addition to the parameters used in case (a), we also determined the parameters: maxsize = 2 (two items in one element of a sequence) and maxlen = 2 (two elements in a sequence). That said, 952 rules were generated, as shown in **Figure 4**.

For the purpose of analysis, two round-trip flights were selected in which the probability of a delay in this type of route impacts not only this route, but also other routes in which the operation can be compromised between these airports. Analyzing the rules generated for flights 3344 and 3345, it is concluded that flight 3344 departed from Guarulhos-So Paulo (SP) with delay of 21 minutes to Confins-Belo Horizonte (MG) and this delay spread for flights, as shown in **Figure 5**.

We can also visualize this propagation of delays by means of a network analysis, generating a graph, as shown in **Figure 6**.

Applying data indexing techniques combined with rules of association, induced after the discovery of frequent sequential patterns, can be hidden patterns of flight delays and their spread.

Considering the data set of domestic flights and guided by the questions asked regarding the causes, moments, differences and relationships between airports, it is possible to evaluate and quantify attributes that could be related to the delays, showing not only the main standards, but also a subset of occurrences of delay propagations in the network and in some airports.

By means of an exploratory analysis of the data, still in time of preprocessing, it was noticed that the delays and their respective propagations have less incidence in the morning, between 6 am and 12 noon, and which is related to a larger number of people working at that time of day, with greater attention in the operations of the airports, speeding up the execution of the embarkation procedures, landing, dispatch, among other activities, as shown in the histogram shown in **Figure 7**.

Another analysis carried out refers to the days of the week when there is a greater incidence of delays and propagations.

From the graph shown in **Figure 8**, it is concluded that Mondays and Fridays are not the best choices for air transport, since there is a higher incidence of delays due to a larger number of flights days that is explained by the fact that people are transported to work on Mondays and return to their homes on Fridays. By analyzing the meteorological conditions of the delayed flights, a graph was generated with the percentage of delays and or cancellations, per airport, due to adverse climatic conditions, as shown in **Figure 9**. Finally, it is verified where the airports that have more delays or cancellations due to climatic conditions are positioned on the Brazilian map, as presented in **Figure 10**.

Figures and tables

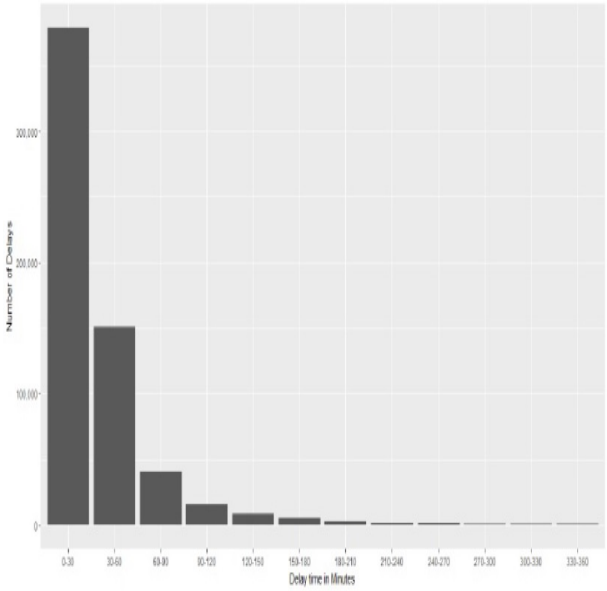

Figure 1. Delayed Flights.

Pseudo-Algorithm: sequenceMining()

```
1: function generateSequences(dataset)
2:   delayedFlights ← preprocess(dataset)
3:   return generateRules(delayedFlights)

1: function preprocess(dataset)
2:   delayedFlights ← removeOutliers(delayedFlights)
3:   return temporalAggregation(delayedFlights)

1: function generateRules(delayedFlights)
2:   rules ← cSPADE(delayedFlights)
3:   return rulesInduction(rules)
```

Figure 2. Pseudo-Algorithm.

```
inspect(sequenceRules)
  lhs         rhs      support confidence lift
1 <{1679}> => <{3796}>       1          1    1
2 <{1679}> => <{3233,        1          1    1
               3796}>
3 <{1679}> => <{3233}>       1          1    1
```

Figure 3. Rules at Belm-PA Airport.

```
inspect(sequenceRules)
     lhs          rhs       support confidence lift
 1 <{1233}> => <{6578}>     0.5           1    2
 2 <{1408}> => <{6578}>     0.5           1    2
 3 <{1677}> => <{6578}>     0.5           1    2
 4 <{2700}> => <{6578}>     0.5           1    2
 5 <{3291}> => <{6578}>     0.5           1    2
 6 <{3299}> => <{6578}>     0.5           1    2
 7 <{3345}> => <{6578}>     0.5           1    2
 8 <{3408}> => <{6578}>     0.5           1    2
 9 <{3804}> => <{6578}>     0.5           1    2
10 <{3895}> => <{6578}>     0.5           1    2
11 <{3896}> => <{6578}>     0.5           1    2
12 <{5019}> => <{6578}>     0.5           1    2
13 <{5187}> => <{6578}>     0.5           1    2
14 <{6305}> => <{6578}>     0.5           1    2
15 <{6312}> => <{6578}>     0.5           1    2
16 <{6577}> => <{6578}>     0.5           1    2
17 <{3895,
    6305}> => <{6578}>      0.5           1    2
18 <{5019,
    5187}> => <{6578}>      0.5           1    2
19 <{3299,
    3896}> => <{6578}>      0.5           1    2
20 <{3408,
    1677}> => <{6578}>      0.5           1    2
21 <{3345,
    1233}> => <{6578}>      0.5           1    2
22 <{6577}> => <{4946,
    6578}>                  0.5           1    2

945 <{4946,
     6578}> => <{1095}>     0.5           1    2
946 <{3516,
     6315}> => <{1095}>     0.5           1    2
947 <{3895,
     6305}> => <{1095}>     0.5           1    2
948 <{5019,
     5187}> => <{1095}>     0.5           1    2
949 <{3562,
     4488}> => <{1095}>     0.5           1    2
950 <{3299,
     3896}> => <{1095}>     0.5           1    2
951 <{3408,
     1677}> => <{1095}>     0.5           1    2
952 <{3345,
     1233}> => <{1095}>     0.5           1    2
```

Figure 4. Rules between Airports.

Departures

#flightNumber	#from	#destiny
4120	Confins-MG(SBCF)	Campinas-Vira Copos-SP(SBKP)
3378	Confins-MG(SBCF)	Recife-PE(SBRF)
3326	Guarulhos-SP(SBGR)	Confins-MG(SBCF)
2185	Confins-MG(SBCF)	Galeão-RJ(SBGL)
1913	Confins-MG(SBCF)	Galeão-RJ(SBGL)
1703	Confins-MG(SBCF)	Brasília-DF(SBBR)

Arrivals

#flightNumber	#from	#destiny
1833	Salvador-BA(SBSV)	Confins-MG(SBCF)
1700	Brasília-DF(SBBR)	Confins-MG(SBCF)

Figure 5. Delay Propagation.

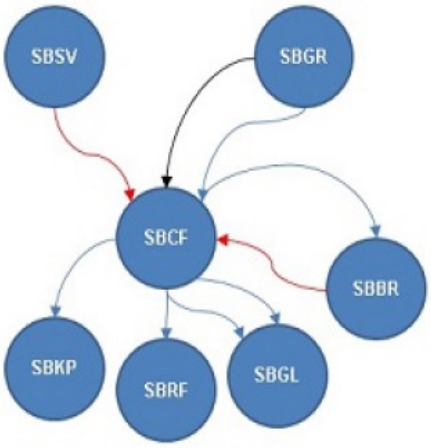

Figure 6. Delay Propagation Graph.

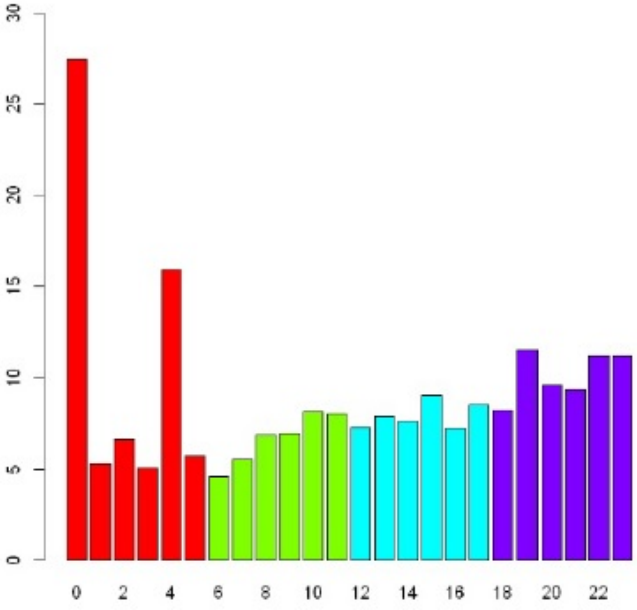

Figure 7. Delays by Period of the day.

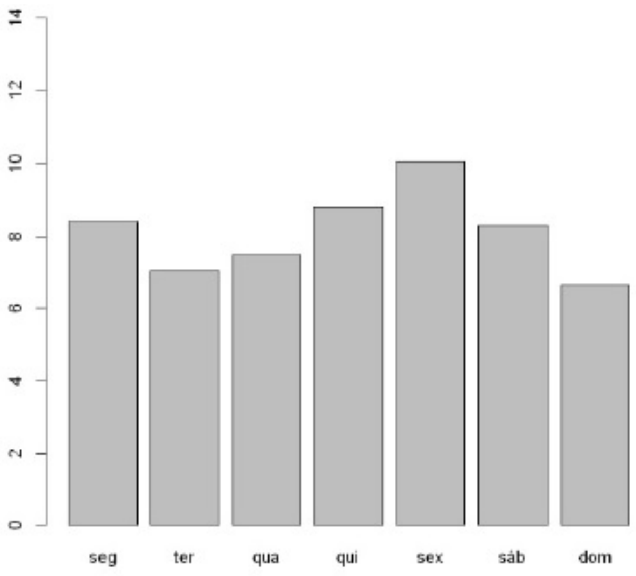

Figure 8. Delays by Days of Week.

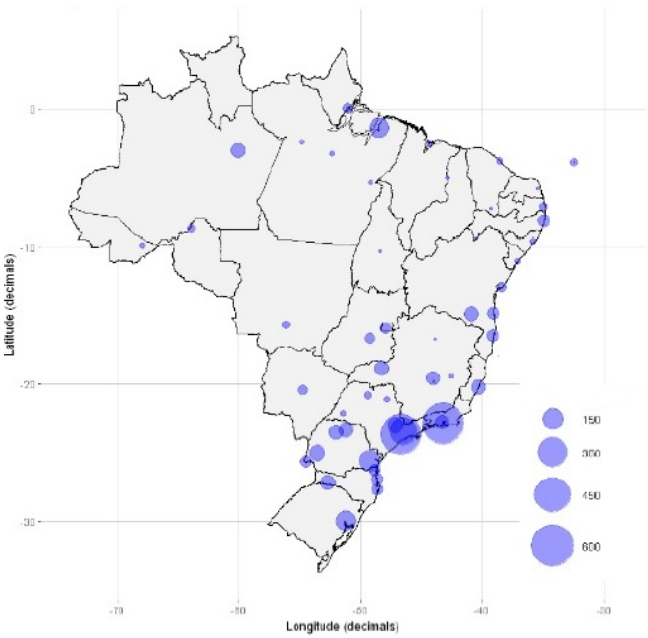

Figure 10. Delays by Region for climatic reasons.

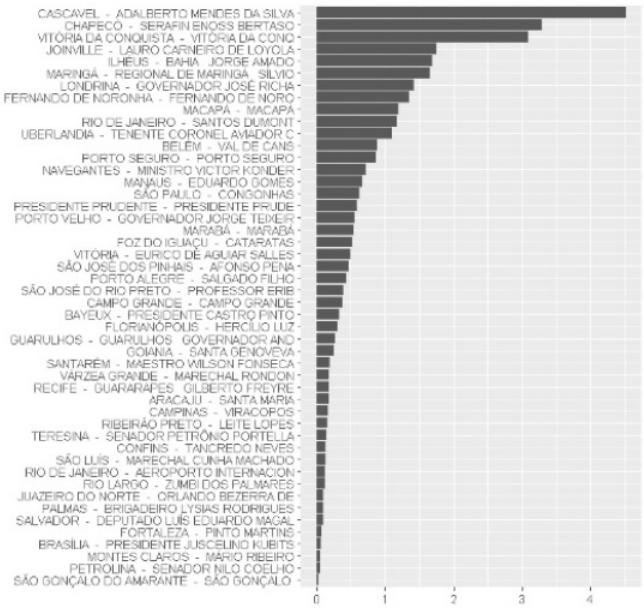

Figure 9. Airport delays due to weather conditions.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

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. 207216.

Agrawal, R. and Srikant, R., (1995), Mining sequential patterns, in ICDE Conference, pp. 314.

Chu, X., Ilyas, I., Krishnan, S., Wang. J., (2016), Data Cleaning: Overview and Emerging Challenges, SIGMOD16, 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. 215224.

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):525531.

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.

Author biography

Claudio Teixeira holds a Master's Degree in Production Engineering, in the research line of Optimization Methods and Network Problems, and currently works on a research project on Data Mining Applied to Products and Processes.

In addition, he has a post-graduate degree in Business Management (Executive MBA) since 2017 by UNESA - Estcio de S University, Bachelor in Informatics and Information Technology (2001) from UERJ - Rio de Janeiro State University, Project Management Professional) by the PMI - Project Management Institute (2008), Information Technology Infrastructure Library (ITIL) by the EXIN - Examinations Institute (2012), PMI Member 1210403 with 20+ years of professional experience in Information Technology with extensive experience in operations and commercial management . Being 10 years in Large Accounts / Projects Management in the Oil and Gas, Logistics and Manufacturing sectors and 8 years in Project Management / Leadership, Change Management, Business Cases Development, Requirements Survey, Modeling, Analysis and Systems Development for large companies in the Manufacturing, Insurance, Telecom, Energy, Government and Food sectors, highlighting major projects of Business Intelligence. People management, experience with more than 150 employees Financial management of contracts (revenue, revenue, operating cost, operating margin, EBITDA, current result and projection) Extensive experience in activities involving the implementation of IT solutions, outsourcing, supplier management and software factory (contract measurements, issuance and control of invoice payments, analysis and negotiation of contract terms and values) Program, Portfolio and Project Management Relationship with internal and external clients Elaboration of commercial strategies in accounts and proposals Identification, qualification and development of opportunities and commercial activity in bidding processes.

Copyright statement

Please be aware that the use of this LaTeX 2$_\varepsilon$ class file is governed by the following conditions.

Copyright

Rules of use

This class file is made available for use by authors who wish to prepare an article for publication in a *SAGE Publications* journal. The user may not exploit any part of the class file commercially.

This class file is provided on an *as is* basis, without warranties of any kind, either express or implied, including but not limited to warranties of title, or implied warranties of merchantablility or fitness for a particular purpose. There will be no duty on the author[s] of the software or SAGE Publications Ltd to correct any errors or defects in the software. Any statutory rights you may have remain unaffected by your acceptance of these rules of use.

Acknowledgements

This class file was developed by Sunrise Setting Ltd, Brixham, Devon, UK.

Website: http://www.sunrise-setting.co.uk

References

Kopka H and Daly PW (2003) *A Guide to LaTeX*, 4th edn. Addison-Wesley.

Lamport L (1994) *LaTeX: a Document Preparation System*, 2nd edn. Addison-Wesley.

Mittelbach F and Goossens M (2004) *The LaTeX Companion*, 2nd edn. Addison-Wesley.