Integrated Dataset of Brazilian Flights*

Claudio Teixeira[1], Lucas Giusti[1], Jorge Soares[1], Joel dos Santos[1], Glauco Amorim[1], Eduardo Ogasawara[1]

[1]CEFET/RJ - Federal Center for Technological Education of Rio de Janeiro

{claudio.teixeira,lucas.giusti}@eic.cefet-rj.br

{jorge.soares,joel.santos,glauco.amorim}@cefet-rj.br, eogasawara@ieee.org

Abstract. *The Brazilian commercial aviation system achieved the first position among Latin American countries and the fifteenth place worldwide on the Revenue Passenger-Kilometer ranking. The availability of flight information, including meteorological conditions, enables studies about the Brazilian flight system, such as flight delays and timetabling. Therefore, this paper contributes to such studies by offering an integrated dataset containing data on departure and arrival for flights departing and arriving at Brazilian airports comprising the period from* 2000 *to* 2019*. This paper presents a dataset composed of* $15,505,922$ *records of flight data, each containing* 45 *attributes. The attributes include data regarding the airline, flight, airports, meteorological conditions, scheduled and elapsed times for departure and arrival.*

1. Introduction

The Brazilian commercial aviation system contains more than one hundred airports. It transported 95.9 million revenue passengers during 2014. It achieved the first position among Latin American countries and the fifteenth place worldwide on the Revenue Passenger-Kilometer (RPK) ranking [ICAO, 2015]. The commercial aviation network in Brazil is organized towards regional hubs in contrast to airline hubs. The main reason is the Brazilian territorial extension and that few Brazilian states have more than one major airport. One exception to this rule is Campinas (in the state of São Paulo), where airline company *Azul* holds 77% of its commercial flights. Besides, due to market deregulation instituted in 2005, the Brazilian commercial aviation system experienced significant changes in its players, leading to market share changes and flight availability.

The National Civil Aviation Agency (ANAC) regulates and supervises Brazilian civil aviation activities. Since 2000, ANAC keeps track of departure and arrival data for Brazilian flights in its Active Regular Flight (VRA) dataset [ANAC, 2015]. The data available in VRA are registered by the airlines and consolidated by ANAC. It contains data about each flight stage, *i.e.*, the aircraft's necessary steps from its takeoff to the next landing. These steps are established regardless of where the object of transport has been loaded or unloaded. For each flight step, VRA provides data such as airline, flight number, type (such as international, domestic, and cargo), class (such as regular, extra, charter, and instruction), airports, and scheduled and elapsed times for departure and arrival. ANAC monthly provides VRA data on its webpage.

*The authors would like to thank CNPq, CAPES (finance code 001), and FAPERJ for the partial funding of the research.

VRA enables studying the Brazilian commercial aviation system. Examples of studies are flight delay patterns [Sternberg et al., 2016b] and their prediction [Moreira et al., 2018; Scarpel and Pelicioni, 2018]. Although meteorological conditions play an essential role in analyzing flight information, such data is not present in VRA. Thus, this paper presents a dataset that integrates Brazilian flight data. It fuses all monthly data available in VRA. It enriches it with meteorological data from the ASOS (Automated Surface Observing Systems) dataset [ASOS, 2000] provided by the IOWA University in the USA. ASOS contains weather sensor data from airports around the world. Data cleaning and data preprocessing techniques were also applied to improve data quality during the entire data integration.

2. Data Acquisition

According to the flight regulation of ANAC, commercial airline companies must register flight metadata indicating changes in flight time, either delay, anticipation, or canceling. They have to log the time a flight happened and a justification for the alteration. Table 1 indicates the flight metadata together with their semantics.

Table 1. Flight metadata registered by airline companies available in VRA

Attribute	Description
Airline	ICAO code representing the airline company
Flight	Flight number
Authorization code	Identifies the authorization type for each flight step
Flight type	Identifies the type of operation performed
Origin	ICAO code of origin airport
Destination	ICAO code of destination airport
Expected Departure	Date and time of scheduled departure
Real departure	Date and time of departure performed informed by the airline
Estimated Arrival	Date and time of estimated arrival
Real Arrival	Date and time of arrival, informed by the airline
Flight status	Informs if the flight was performed or canceled
Justification Code	Register delay, cancellation, or changes regarding planning

According to the regulation of ANAC, the metadata indicated in Table 2 must be registered in paper form, either typed or handwritten. ANAC then consolidate the data sent by the airline companies into the VRA dataset. VRA is published monthly, comprising all flight steps expected to depart in a given month.

The primary goal of ANAC is to use the recorded metadata to compute the punctuality rate of airlines. Thus, sector regulation obliges airline companies to provide the data presented in Table 1. Therefore, it comprises all flight steps that took place in a given period. However, around 20% of the records may be considered inconsistent due to errors while filling the report form. As will be presented in Section 3.1, the causes of errors include arrival time before departure or flight duration inconsistent with the regulation of ANAC.

Meteorological conditions play an important role in aviation operations. The Automated Surface Observing Systems (ASOS) is a program that involves several American government agencies. It was created to become an official network of meteorological information to support primarily aviation entities. It includes meteorological, climatolog-

ical, and hydrological components. ASOS data come from weather sensors in locations all over the planet. In Brazil, ASOS covers all 154 airports available in VRA.

The Department of Agronomy at Iowa State University, in the United States, compiles daily information from the US ASOS system. It creates an hourly report of meteorological observations in all of its sites. Table 2 indicates the meteorological data together with their semantics.

Table 2. ASOS meteorological data

Attribute	Description
Sky condition	Cloud height and amount (clear, scattered, broken, overcast)
Visibility	To at least ten statute miles
Weather	Type and intensity for rain, snow, and freezing rain.
Obstructions to vision	Fog, haze
Pressure	Sea-level pressure, altimeter setting
Temperature	Ambient and dew point temperature
Wind	Direction, speed, and character (gusts, squalls)

3. Integrated Dataset

The integrated *Brazilian Flight Dataset* (BFD) presented in this paper includes both the flight data present in VRA and meteorological information present in ASOS. It is intended to enable studies regarding the Brazilian commercial aviation system. BFD is composed of $15,505,922$ records of flight data, each containing 45 attributes. The dataset, together with its integration process description and R scripts, is available on IEEE DataPort[1].

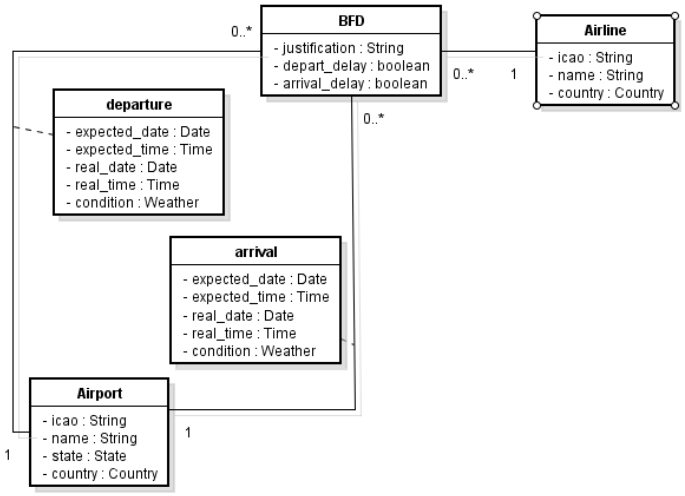

Figure 1. The data model for the BFD

Figure 1 presents the data model of BFD. It is detailed in the following sections. As can be seen, BFD aggregates data from VRA and ASOS for flight information and meteorological information, respectively. It also includes data currently unavailable in VRA, such as describing the justification codes of ANAC, airline and airport names, and ISO codes for country names.

[1]Dataset is available at http://dx.doi.org/10.21227/k10b-qn21.

BFD focus on flight data regarding flights that departed or arrived in Brazil. When both origin and destination airports are located in Brazil, those flights are considered domestic flights. Conversely, when either the origin or the destination airport is located outside of Brazil, it is considered international. The data integration process for creating BFD was organized into three main activities: (i) data preprocessing, (ii) data enrichment, and (iii) data fusion. Those activities resemble the traditional Extraction, Transformation, and Load (ETL) process [Vassiliadis, 2009].

3.1. Data Preprocessing

The preprocessing stage was performed in three parts. First, VRA attribute names were translated from Brazilian Portuguese to English. It was unnecessary to translate the acronyms used in each variable since they were already following the International Civil Aviation Organization (ICAO) standards. It was necessary to convert temperature and dew point data to the International System of Units regarding the ASOS data. Data from ASOS was filtered to consider the 154 airports available in VRA.

The second part consisted of data cleaning for both VRA and ASOS datasets. Given that flight information is usually recorded by hand, VRA data was cleaned to remove inconsistent data. During cleaning, records with missing variables were removed. Also, records with departure time (either elapsed or expected) greater or equal to arrival time were removed. They corresponded to approximately 0.02% of the records. Approximately 3.77% of VRA records were removed for being out of BFD scope, *i.e.*, with origin and destination out of Brazil. Finally, the regulation of ANAC prohibits delays higher than 24 hours. Thus, during cleaning records with departure or arrival delays exceeding this norm were removed. The complete data cleaning removed 21.07% of VRA records.

The third part of the preprocessing stage consisted of removing outliers. For each pair of airports $\langle o, d \rangle$ in VRA, it was considered both the expected and elapsed duration of a flight from origin o and destination d. Flights whose duration (either elapsed or expected) were not in the interval $[Q_1 - 3 \cdot IQR, Q_3 + 3 \cdot IQR]$ were considered outliers. These values were used because we follow a more conservative approach [2]. They corresponded to 2.76% of VRA records. The preprocessing step resulted in $15,505,922$ flight records from VRA to be used in the fusion stage.

3.2. Data Enrichment

After preprocessing, the dataset is enriched as follows. The dataset schema is changed by separating departure and arrival data attributes (see Table 1 into hour and date attributes. Besides, it included attributes related to flight duration, departure, and arrival delays.

Additionally, two discrete attributes were included for the time of the day for departures and arrivals. It divides the time attribute into seven ranges, as presented in Table 3.

Two discrete attributes are included in ASOS while enriching the dataset. The use the wind velocity in knots to include the wind intensity using a Beaufort Scale. The second uses the wind direction in degrees to include the wind direction using Wind Rose

[2]https://www.sciencedirect.com/topics/mathematics/extreme-outlier

Table 3. Time attribute discretization

Period	Start Time	End Time
Night	23:00	04:00
Early Morning	05:00	08:00
Mid Morning	09:00	10:00
Late Morning	11:00	12:00
Afternoon	13:00	16:00
Early Evening	17:00	19:00
Late Evening	20:00	22:00

with 16 cardinal directions (derived from N, E, S, W)[3].

Both discretized data (window intensity and wind direction) are related to takeoff and landing safety. Takeoff and landing are critical phases in a flight. According to ANAC, the ideal wind for landing and takeoff is always the headwind, never the tailwind. The headwind is the wind contrary to the aircraft. When the wind changes direction, air traffic control agencies usually change the takeoff and landing operations, using another headland. Up to five knots, the wind would be considered harmless for takeoff and landing operations and does not require a headland change. Finally, there are the gusts of wind. When the average wind speed is exceeded by ten or more knots for at least 20 seconds, it is considered a gust of wind, and procedures must be adopted to ensure takeoffs and landings.

3.3. Data Fusion

Data fusion was applied over VRA data from 2000 to 2019, except for June, July 2014, and March 2018, when ANAC did not collect the data. It is worth mentioning that ASOS provides hourly meteorological data.

During the fusion process for the meteorological and flight data, it was necessary to group all flight data in a given hour. The grouping was performed for each elapsed departure and arrival of the flight to determine its meteorological information. The process needed to select the flights in a discrete way, hour by hour, because the weather conditions in the ASOS system are collected hourly on each station (airport).

Furthermore, the fusion stage resolved airport and airline names from VRA data. It also included an ISO code for country names whenever the flight departs or arrives at a non-Brazilian airport. Finally, the justification codes for flight delay were also expanded to their descriptions.

4. Dataset Usage

BFD allows for studies regarding the Brazilian commercial aviation system. This section presents previous and ongoing work conducted on top of BFD together with an exploratory analysis of BFD data. To show the importance of using the database, we conduct an exploratory analysis and mention studies that used the data in their research.

As discussed before, the Brazilian flight system is oriented towards regional hubs instead of company hubs. Figure 2 presents the number of flights per airport, considering

[3]Wind Rose Data - US Department of Agriculture - Natural Resources Conservation Service (NRCS) available at https://www.wcc.nrcs.usda.gov/climate/windrose.html

just the 25 biggest airports on flights. It also divides flights into domestic (D), international (I), and cargo (C) flights.

As can be seen in Figure 2, in the top five busiest airports, the first two are in São Paulo (SBSP and SBGR), the third in Brasília (SBBR), and the last two in Rio de Janeiro (SBGL and SBRJ). Rio and São Paulo are the two higher Gross Domestic Products (GDPs) in Brazil. They are two major gateways for flights coming and exiting Brazil. Approximately one-third of Guarulhos Airport (SBGR) and Galeão Airport (SBGL) flights are international flights.

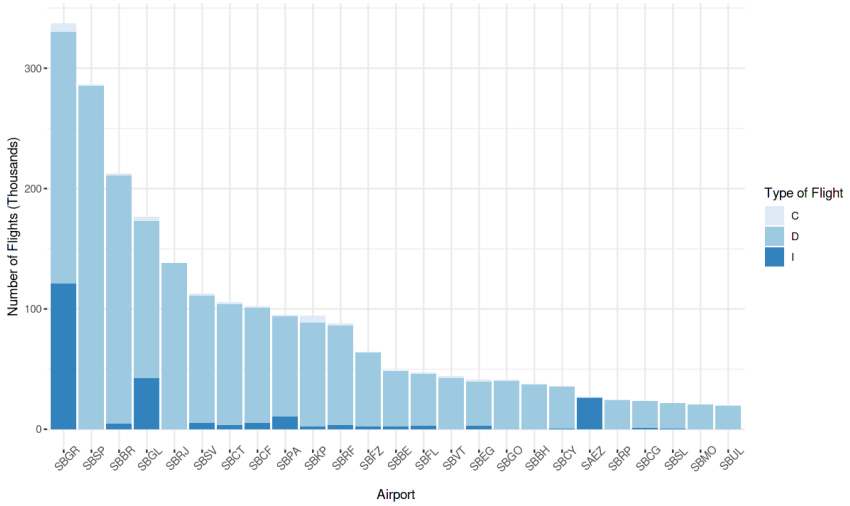

Figure 2. Number of flights per airport, for the top-25 most active airports

Brasilia is the capital of the country and is located in the middle of Brazil. It acts as a hub for flights from and to cities in the north and northeast regions. It can be seen, however, that it has few international flights.

Brazil and Argentina have solid touristic relations. Thus we can see the Buenos Aires international airport (SAEZ) in the top-25 busiest airports. Since BFD has only flights from and to Brazil, SAEZ has only international and cargo flights.

Figure 3 presents the distribution of flights according to the period of the day. As shown, most of the flight departures (Figure 3.a) occur in the afternoon and early evening. Most arrivals (Figure 3.b) occur in the afternoon and early morning. During the mid and late morning, the number of flights decreases significantly for both departure and arrival.

According to ANAC regulation, a flight is considered to be delayed when its departure or arrival time surpasses, respectively, the expected departure or arrival by more than 30 minutes. Figure 4 presents the punctuality rate considering all the Brazilian flight systems per year, from 2000 to 2019. It is possible to observe that the Brazilian flight crises that occurred in 2007 interfered with both punctuality rates and mean delay [Times, 2007].

Figure 5 plots the monthly behavior of the Brazilian systems. Historically, months of school break (December, January, and July) have the lowest punctuality rates and the highest mean delay. August is the month with the highest level of punctuality and lowest mean delay.

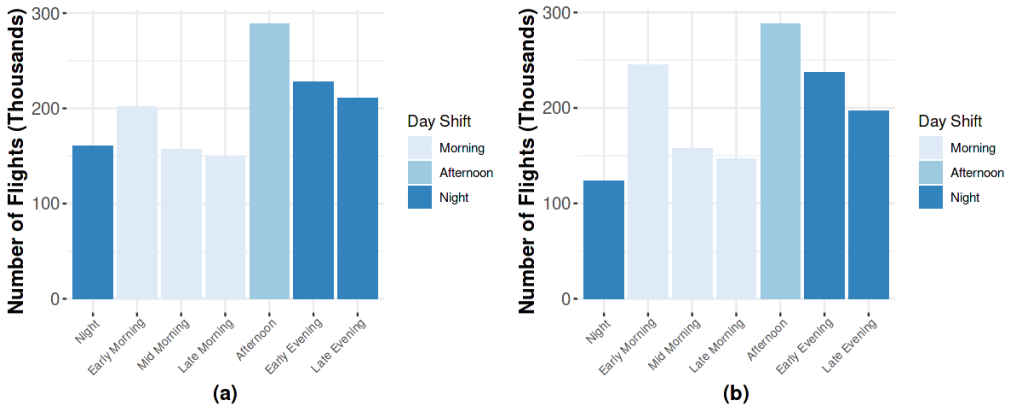

Figure 3. Number of flights per period of the day: (a) departure; (b) arrival

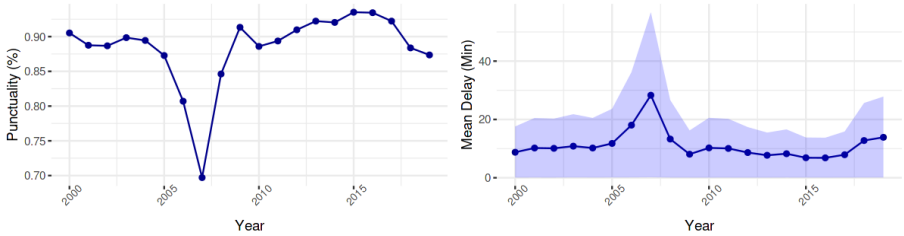

Figure 4. Punctuality rate and mean delay per year. The charts present the mean delay together with its confidence interval of 95%

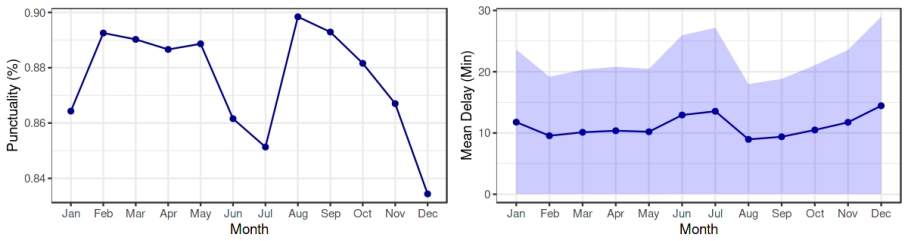

Figure 5. Punctuality rate and mean delay per month of the year

Given the various inconveniences for airlines, airports, and passengers caused by flight delays, it is fundamental to mitigate their occurrence and optimize an air transport system's decision-making process. Airlines, airports, and users may be more interested in when delays are likely to occur than the accurate prediction of the absence of delays. In that context, Moreira et al. [2018] use BFD to analyze Flight delays in the period between 2009 and 2015. The authors present a classification model capable of predicting delays, getting about 60% of hits.

Flight delays fall into two main categories: root delay and delay propagation. Root delays are related to events that are intrinsic to a particular flight. In delay propagation, it is presumed that a delay has already occurred at some point in the network, *i.e.*, new delays occur due to previous delays. The understanding of delay propagation patterns among airports is essential for decision-making processes.

That study may devise patterns in flight delays and the way the system recovers

from them. Focusing on unveiling those patterns, Sternberg et al. [2016a] apply data indexing techniques combined with BFD data association rules. The authors observed that the Brazilian flight system has difficulties recovering from previous delays when operating under adverse meteorological conditions, when delays may increase up to 216%.

5. Conclusion

This work aimed to create a reliable and enriched database on national and international flights that arrived and departed from Brazilian airports. With the data offered by this database, it is possible to carry out several studies to aid the decision-making process. For example, it is possible to answer the following questions: (i) "Which airport suffers the most delays?"; (ii) "What month of the year is an airport most likely to be delayed?"; or (iii) "What part of the day is a particular airport most likely to experience a delay in departure?" The answers to these questions can help companies and governments review their protocols and optimize their services.

As a data limitation, we can mention the manual filling in delays, flight cancellations, and schedule updates. This situation contributes to the enormous value of inaccurate information that has been removed from the dataset.

As future work, we intend to update this dataset yearly, conducting the entire data integration.

References

ANAC. Agência Nacional de Aviação Civil. Technical report, https://www.gov.br/anac/pt-br, 2015.

ASOS. Automated Surface Observing System. Technical report, https://mesonet.agron.iastate.edu/ASOS/, 2000.

ICAO. Annual Report of the Council 2014. Technical report, http://www.icao.int/annual-report-2014/Pages/default.aspx, 2015.

L. Moreira, C. Dantas, L. Oliveira, J. Soares, and E. Ogasawara. On Evaluating Data Preprocessing Methods for Machine Learning Models for Flight Delays. In *Proceedings of the International Joint Conference on Neural Networks*, volume 2018-July, 2018.

R. Arnaldo Scarpel and L.C. Pelicioni. A data analytics approach for anticipating congested days at the São Paulo International Airport. *Journal of Air Transport Management*, 72:1–10, 2018.

A. Sternberg, D. Carvalho, L. Murta, J. Soares, and E. Ogasawara. An analysis of Brazilian flight delays based on frequent patterns. *Transportation Research Part E: Logistics and Transportation Review*, 95:282–298, 2016a.

Alice Sternberg, Diego Carvalho, Leonardo Murta, Jorge Soares, and Eduardo Ogasawara. Experimental Evaluation. Technical report, https://eic.cefet-rj.br/~dal/an-analysis-of-brazilian-flight-delays-based-on-frequent-patterns/, 2016b.

New York Times. Brazil Demands Solution to Aviation Crisis. Technical report, https://www.nytimes.com/2007/07/19/world/americas/19brazil.html, 2007.

P. Vassiliadis. A survey of extract-transform-load technology. *International Journal of Data Warehousing and Mining*, 5(3):1–27, 2009.