Development of a data processing system from OpenSky Network database for Lisbon region

Miguel Barrameda Leite Salazar[a]

[a] *ISEC Lisboa, Lisbon, Portugal*

Abstract: Obtaining data stored from the OpenSky Network originated by ADS-B exchanged messages allows analysis and studies to understand the behaviour of air traffic flow. This project – destined to the ISEC Lisboa and the aviation community – aims to create a volume-oriented tool for air traffic in Lisbon Airport (LPPT) focused on Air Traffic Management optimization. Through a dashboard environment, it will be able to treat and present flight information such as arrivals and departure volumes, time-focused and geography-focused analysis and a trajectory visualization, for example. From the database, using the R language and R Studio via libraries such as "openskyr" and "rshiny", this project will establish a different approach for using the information available on databases such as OpenSky Network, using Big Data concepts, such as Data Engineering, Data Science and Data Visualization

Keywords: OpenSky Network, Air Traffic Management, R language, openskyr, Big Data.

1. Introduction

The increase on air traffic flow throughout the world brings with it the need to develop technologies that can optimize air operations and Air Traffic Control (ATC) procedures (Schäfer et al. 2014). Air Traffic Surveillance even though is not recent technology, it is constantly modernizing itself to respond to the greater traffic demand, and such process was proven successful due to increase on precision, volume and speed of messages exchanges and transmissions (Seymour 2018), which translates on extremely reliable and precise operations. This paper relies on 3 main pillars: ADS-B, which is considered a NextGen Surveillance System created to substitute RADAR technology (Strohmeier et al. 2014); Open Sky Network, a website capable of "translating" ADS-B messages into live traffic data and also storage this data into accessible Data Sets (Open Sky Network 2022); and the R language, which is the conductive thread of this work, since is able to gather the Data from OpenSky Network and model it into visual-friendly analysis regarding air traffic behaviours, patterns and history (Ayala et al. 2021).

Firstly, analysing Aircraft Surveillance systems throughout modern aviation is important to understand how ADS-B was conceived. The main system used in aviation nowadays is the RADAR (Radio Detection and Ranging), which is generally composed of a Primary Surveillance Radar (PSR) and a Secondary Surveillance Radar (ICAO Asia 2007; Abdulaziz et al. 2019). The PSR consist in the transmission of high-frequency signals that, when encounter an aircraft, determine its position in relation to the station and its heading (Wesson, Humphreys, and Evans 2014; Gilbert 1973). It based off the Doppler Effect, that is, the wave is transmitted and the going "back and forth" is what allows it to determine position and direction. Therefore, it is considered an independent system and, despite providing little information and not guaranteeing such precision, it is still used to provide trustworthy readings (Miranda 2022a; Strohmeier et al. 2014). The second one, SSR, has a different operation, while the ground station sends "questions" to the aircraft within its reach, on a 1030 MHz frequency, and waits for the answer, on the 1090 MHz bandwidth (question-answer system), to acquire flight information such as aircraft identification, position, altitude. Therefore, it depends on an onboard device– the Transponder – to understand the question asked by the station and respond with the desired information (Miranda 2022b; Rodrigues 2010). Figure 1 presents both methods visually.

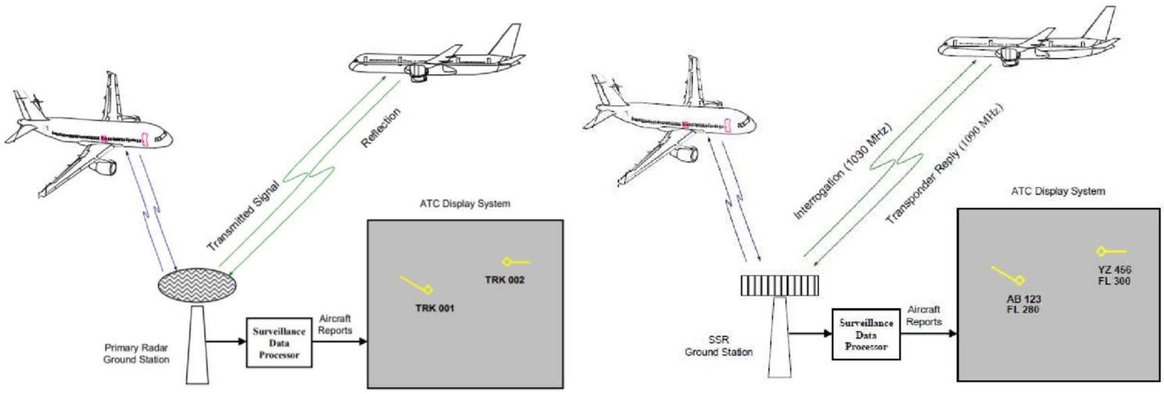


Fig 1. Principle of PSR (left) and SSR (right) (ICAO Asia 2007).

Considering the increased flow of air traffic and inherent limitations of the RADAR system, it became overloaded and subject to problems, being unable to meet the high demand. Therefore, the emergence of the Automatic Dependent Surveillance-Broadcast (ADS-B) technology intended to meet the needs of airspace (Abdulaziz et al. 2019). ADS-B is a system powered by GNSS data (GPS, Galileo, GLONASS, for example) from the aircraft to determine its position and generate a message (Baek and Bang 2012). These messages are constantly transmitted to stations on the ground (ADS-B OUT) and to other aircraft (ADS-B IN) (Rodrigues 2010). Therefore, its purpose is to determine the position of the aircraft and, automatically and repetitively, transmit the information together with callsign, altitude, heading and identification of the aircraft, as shown in Figure 2. One of the great advantages of the ADS-B system is that it is compatible with the Mode S system, while the transmission of information also takes place in the 1090 MHz frequency, greatly facilitating the application of this new method (Rodrigues 2010; Schäfer et al. 2014).

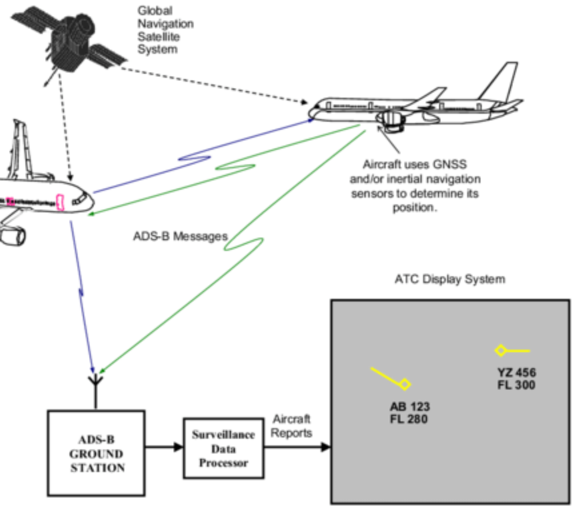


Fig 2. Principle of ADS-B (ICAO Asia 2007).

The ADS-B mechanism, however, has several issues related to Security, given that it was designed with greater focus on Safety (Seymour 2018). The system has a number of vulnerabilities, as the transmitted messages do not have any encryption or digital signature and, therefore, new messages can be inserted, or these can be modified and even deleted (Strohmeier et al. 2014; Seymour 2018). Initiatives for establishing layers of security for ADS-B often encounter several regulatory and technical issues. In addition, any modification in ADS-B will force a decrease messages transmissions capabilities and will always put ADS-B under “surveillance” and approval of RADAR, even though ADS-B is the system that was supposed to replace the conventional PSR and SSR (Wesson, Humphreys, and Evans 2014). Despite having security issues for full applicability in aeronautical operations, the messages exchanged from the ADS-B system are stored in large databases and, considering the nature of the messages and the frequency of transmissions, possess large analytical value. For this same reason, different organizations and projects use this large database to develop studies and tools to the aeronautical community, both in the organization and availability and in the presentation of this data. This is the case of the OpenSky Network, a virtual community that identifies and positions on the map aircrafts that transmit from the ADS-B system (as well as Mode S, TCAS and FLARM) live, and provides specific information for each aircraft, such as for example the origin, destination, equipment, transmission source, altitude (Open Sky Network 2022). In addition to providing the live map, as shown in Figure 3, OpenSky Network creates a database that stores the most varied data about flights performed: estimated and actual take-off time, expected and actual landing time, delays, flight path, altitude and speed profile along the route, airports, traffic volume.

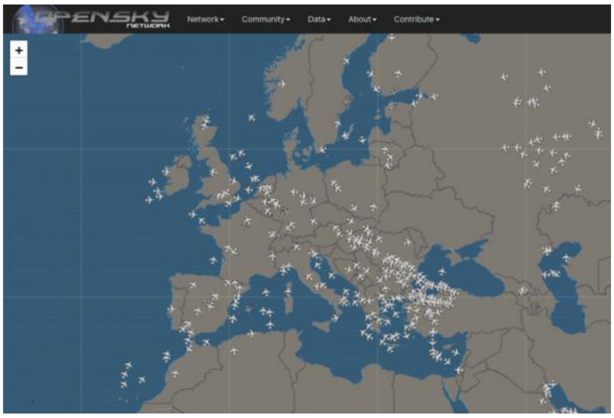


Fig 3. OpenSky Network main window (Open Sky Network 2022).

The objective of this work is to use the data collected by the Open Sky Network, applying the concepts and processes of Big Data (Data Engineering, Data Science and Data Visualization), culminating in the creation of a tool capable of analysing and presenting the data obtained through this network. Thus, such a system will serve the aeronautical community as it can transform large amounts of data into inferences and valuable information. These can be used, at the user's discretion, in the search for Air Traffic Management optimization, since the greater the dynamism of analysis capacity, greater the capacity of a given airspace to hold traffic. Fulfilling this objective will not only affect ISEC Lisboa, but also the aeronautical community regarding the ability to understand behaviours, trends, histories and to carry out analyses based on large-scale data and high informative and analytical value.

2. Case Study

The second chapter aims to develop everything that has already been discussed, studied and presented within the academic universe on the scope of the main themes of this work, namely the ADS-B system, Open Sky Network and the R language, R Studio and R Shiny. For this, it resorts to different articles, books and other academic sources, approaching such studies regarding its premises, hypotheses, and results. In this way, it is possible to better understand not only the relevance of this theme, but also the importance of carrying out this work.

Firstly, it is extremely important to refer to the process of implementation and consolidation of ADS-B as a Next Generation system for tracking and managing air traffic. As already discussed in this document, ADS-B operates via automatic and constant transmission of messages containing flight information between aircraft in the same airspace. This method significantly reduces the intervention of the Air Traffic Control (ATC), in addition to reducing fuel consumption and costs, especially those associated with the installation and maintenance of the structures that make up the RADAR system.

According to the Thesis “ADS-B - Automatic Dependent Surveillance Broadcast: estudo de impacto em Portugal”, the implementation of ADS-B in Europe was carried out after a series of measures for regulating and unifying European airspace as a single one. The launch of Single European Sky in 2004 marks the beginning of the foundations for establishing ADS-B. Together with the Single European Sky, the Single European Sky ATM Research (SESAR) was created, which is the technological division of the Single Sky; and also the Co-operative ATS through Surveillance & Communication Applications Deployed in ECAC (CASCADE), a division of SESAR focused on analysis, security and sustainability in the approval processes of ADS-B as an aerial surveillance tool (Rodrigues 2010). Currently, ADS-B has wide coverage throughout Europe and therefore, in addition to being extremely automated and cheap, it is also capable of generating large amounts of data and information for analysis, especially in the academic field given the ease of access to these (Rodrigues 2010; Schultz, Rosenow, and Olive 2022), even if it has limitations in terms of Security and Safety for it to be fully integrated into the NextGen ATC system (Strohmeier et al. 2014).

The article “Surveillance Radar System Limitations and the Advent of the Automatic Dependent Surveillance Broadcast system for Aircraft Monitoring” describes not only the system that preceded ADS-B, but also addresses the advantages, disadvantages and challenges associated with its implementation. According to Abdulaziz et al., 2019, the advantages of ADS-B are the ability to provide signal coverage in places that RADAR was not able to reach, namely in transoceanic navigations; increased efficiency and dynamism in operations; increase in Safety; improves visibility, since the transmission signal does not deteriorate with distance (either the aircraft is in range, or it is not); less impact on the environment, due to the reduction in physical facilities. Therefore, the system allows ATC to reduce the separation between landings and take-offs without compromising Safety, thus ensuring fewer occurrences of waiting to perform approaches, resulting in the reduction of associated costs (Abdulaziz et al. 2019). On the other hand, the study considers also the disadvantages associated with the ADS-B system: dependence on transponders; possible loss of GPS position due to unavailable satellites; issues related to stability, integrity and credibility related to the content of the ADS-B system (Abdulaziz et al. 2019; Seymour 2018).

Considering the context that culminated in the creation of ADS-B and its subsequent implementation as an air surveillance system, in addition to its characteristics and barriers, the document "Realities and Challenges of NextGen Air Traffic Management: The Case of ADS-B" studies the quality and reliability of messages transmitted by ADS-B in the light of Security. That is, whether the methods used by this system are capable of protecting the operation from illicit acts (Seymour 2018). In order to accurately assess the ADS-B communication channel and its messages, 53,626,642 communication records were used (Strohmeier et al. 2014). It was found that there is a great loss of messages, so in an attempt to determine possible weaknesses of the ADS-B system, these losses were associated to the distance from the receiver, time of day and number of aircraft within the same range. Figure 4 represents the studies for each parameter: regarding distance, there is considerable loss up to 50 km and also after 280 km (a). With regard to the time of day, losses are lower during the morning and evening, that is, at off-peak times, when there are fewer aircraft operating and transmitting within the same airspace (b). Finally, regarding the number of aircraft transmitting and receiving ADS-B in the same space, losses grow as traffic increases (c).

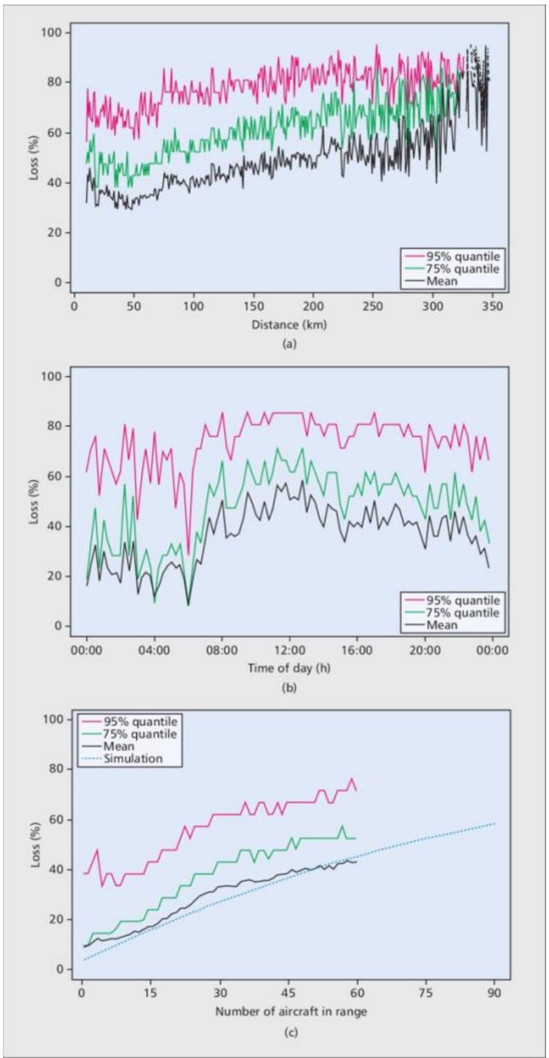


Fig 4. ADS-B message lost (%) according to each metric (Strohmeier et al. 2014).

Considering issues related to the ability of the ADS-B system to protect itself from illicit interventions and cyberattacks, the academic sphere and ATC institutions have focused on the weaknesses of the ADS-B to create mitigation measures in favour of Security (Seymour 2018). The document "Experimental Analysis of Attacks on Next Generation Air Traffic Communication" focuses on attacks to the ADS-B system, as it analyses, simulates and measures the impacts that each type of security breach can bring. Initially, these attacks are divided into passive and active. The former is characterized and originated by the ease in obtaining the data transmitted by ADS-B, which can later be used for a planned and targeted attack based on the information carried by the data, which brings up discussions related to privacy and, consequently, espionage (Schäfer, Lenders, and Martinovic 2013).

On the other hand, active type attacks have more severity as they can directly affect air traffic safety, namely in the visualization of traffic by ATC as well as systems such as TCAS. These attacks are always within the scope of injecting, deleting and modifying messages (Schäfer, Lenders, and Martinovic 2013). Ghost Aircraft Injection is the attack characterized by presenting any ADS-B receiver with an aircraft that does not exist but which, under certain conditions such as low visibility, can create constraints related to landings, take-offs, separations and even in ground Traffic (Schäfer, Lenders, and Martinovic 2013); Ghost Aircraft Flooding consists of the same principle as the previous one, but multiple aircraft simultaneously, which causes not only visual pollution for the ATC but also the inability of the airspace to receive more traffic even if, in the reality, there is no such quantity (Schäfer, Lenders, and Martinovic 2013; Strohmeier et al. 2014); Virtual Trajectory Modification are attacks that either modify the position transmitted by an aircraft, or delete the message containing the correct information and insert a new message with the false information. It takes into account the inaccuracy of a system, for example ADS-B+PSR, and even if it is a small change, it can alter the ATC instructions for the aircraft (Schäfer, Lenders, and Martinovic 2013; Miranda 2022a); False Alarm Attacks consist on modifying or deleting and inserting a transmission in order to fake a dangerous situation in a certain aircraft, for example a hijacking, in order to divert the attention of Air Traffic Control Operators (ATCOs) and tamper with the normal workflow (Schäfer, Lenders, and Martinovic 2013); Ground Station Flooding is described by ATCOs as devastating since, in heavily trafficked areas, these are able to prevent collision warning systems from acting (Schäfer, Lenders, and Martinovic 2013); Aircraft Disappearance consists of deleting all messages transmitted by a given aircraft, which prevents ATC, other aircraft or TCAS from taking that aircraft into account (Schäfer, Lenders, and Martinovic 2013); Aircraft Spoofing is about spoofing the 24bit ICAO address in order to change the data of an aircraft (Schäfer, Lenders, and Martinovic 2013).

Taking in consideration the existence of several methods to attack the vulnerabilities of the ADS-B system, the report "Can Cryptography Secure Next Generation Air Traffic Surveillance?" seeks to present ways to arm ADS-B with defences capable of resisting cyberattacks. It is necessary to take into account that any change or renewal of encryption protocols to protect ADS-B can run into a series of barriers, such as: ADS-B being an international protocol, and therefore it must be compatible with different policies and technological availability ; ADS-B operates under a limited frequency spectrum; ADS-B is limited and subject to interference by Mode S transmissions, as they propagate on similar frequencies, and any modulation of the ADS-B message can result in increased interference and reduced operational capacity (Schäfer, Lenders, and Martinovic 2013; Wesson, Humphreys, and Evans 2014); ADS-B operates in an environment that, from a cryptographic point of view, is unreliable and, therefore, any security software will be accessible by anyone (Wesson, Humphreys, and Evans 2014). Even so, from need to insert cryptography for ADS-B security point of view, the most viable would be Asymmetric-Key Cryptography, even if it reduces the capacity of the 1090 MHz frequency and creates the need for key management, which that would hardly be accepted by regulatory agencies, commercial companies and general aviation enthusiasts. In the end, ADS-B will continue to rely on information coming from the RADAR system for authentication (Abdulaziz et al., 2019; Strohmeier et al., 2014; Wesson et al., 2014).

The advent of ADS-B facilitated data gathering for the creation of tools capable of presenting information in the most varied formats. The Open Sky Network, which is the second foundation for carrying out this work, is a network capable of capturing ADS-B messages and displaying aircraft live, in addition to creating a database containing information such as the flight identifier, aircraft, origin, destination, altitude and speed profile, trajectory (Schäfer et al. 2014; Open Sky Network 2022). Eventually, the amount of data collected by OpenSky Network started to be used as basis for different studies related to, for example, operational efficiency, ecological footprint and emissions, study of trajectories and even the impact of the COVID-19 pandemic on air transport(Sun et al. 2022; Perrichon, Gendre, and Klein 2022; Koelle and Barbosa 2021; Polishchuk, Lemetti, and Sáez 2019).

The first case study to be addressed was presented in the article "Evaluation of Flight Efficiency for Stockholm Arlanda Airport using Open Sky Network Data" and deals with the evaluation of the efficiency of the airspace at Stockholm Airport in the light of fuel consumption and vertical profile from descent to approach. Furthermore, the study uses data from OpenSky Network as well as data from the EUROCONTROL Demand Data Repository (DDR2), to also compare the accuracy and quality of data stored for the two different sources. Carrying out this analysis presupposes the determination and use of Key Performance Indicators (KPIs) to assist in choosing the metrics to be used. In this way, the average delay, percentage of delayed flights, vertical flight efficiency (VFE) in descent (Figure 5), additional time in Terminal Manoeuvring Area (TMA) and fuel-based Performance Indicator were used to determine additional fuel consumption, as shown in Figure 7 (Polishchuk, Lemetti, and Sáez 2019). Both OpenSky Network data and DDR2 data provide lateral and vertical profiles, which allow you to graphically visualize the VFE. The analysis of additional fuel consumption is performed from the difference between the flight path performed and the ideal flight path to reduce consumption, based on the lateral profile of the approach at the TMA (Figure 5), and reflects an inefficient VFE (Polishchuk, Lemetti, and Sáez 2019; Sun et al. 2022).

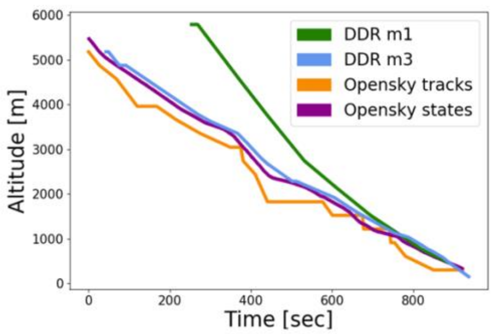


Fig 5. VFE on the TMA for SAS410 flight (Polishchuk, Lemetti, and Sáez 2019).

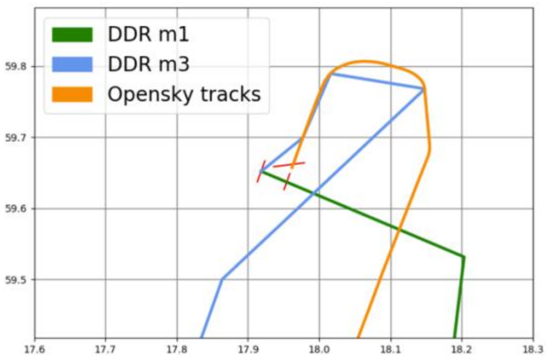


Fig 6. Lateral Profile on TMA for SAS964 flight (Polishchuk, Lemetti, and Sáez 2019).

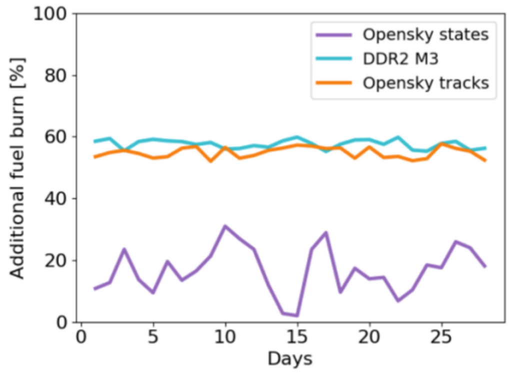


Fig 7. Additional fuel consumption per day (Polishchuk, Lemetti, and Sáez 2019).

From the of flight paths point of view and considering OpenSky Network data as a source of study, “A Geometric Approach to Study Aircraft Trajectories: The Benefits of OpenSky Network ADS-B Data” compares data from EUROCONTROL and OpenSky Network to compare data and perform analyses. The geometric framework of this work aims, primarily, to compare the difference in the shape of the curves and trajectories between the two data sources, as seen in Figure 8 and 9. For this, it uses Paris-Orly Airport (LFPO) and Toulouse-Blagnac Airport (LFBO) as a reference (Polishchuk, Lemetti, and Sáez 2019).

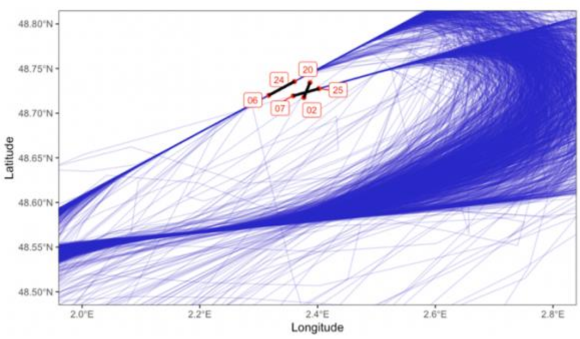


Fig 8. Lateral Profile on LFPO TMA from ADS-B data source (Polishchuk, Lemetti, and Sáez 2019).

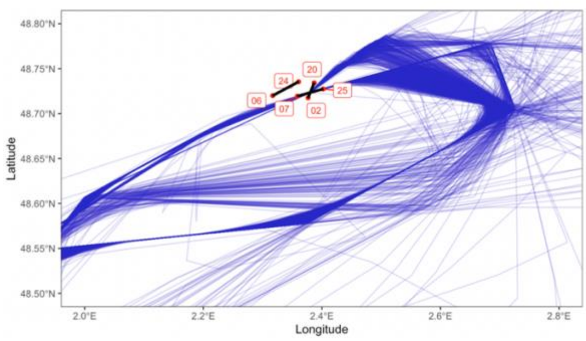


Fig 9. Lateral Profile on LFPO TMA from EUROCONTROL data source (Polishchuk, Lemetti, and Sáez 2019).

The difference between EUROCONTROL and OpenSky Network data is visible, as the second data source – ADS-B – records information in a shorter time interval, so there is greater precision in the information presented (Schäfer et al. 2018; Perrichon, Gendre, and Klein 2022). It is concluded, therefore, that the geometric framing to determine not only the accuracy of the data, but also to extract useful visualizations such as the ones seen above is very helpful for understanding the flow of air traffic, having applications in other areas such as Noise Abatement Procedures, homologation and development of arrival and departure procedures, optimization of fuel consumption and dynamization of the ATM, for example.

With regard to the relationship between the COVID-19 pandemic and the analysis of air traffic from the OSN, the paper “Assessing the Global COVID-19 Impact on Air Transport with Open Data” uses the retroactive data collected through the ADS-B system to assess the impact that the pandemic period had on air traffic flow, as presented in Figure 10. This period created the need to carry out studies and seek sources capable of determining the real size of the problem, especially in aviation, it was imperative to carry out studies and analyses about the recovery capacity of this industry (Koelle and Barbosa 2021; Strohmeier et al. 2021).

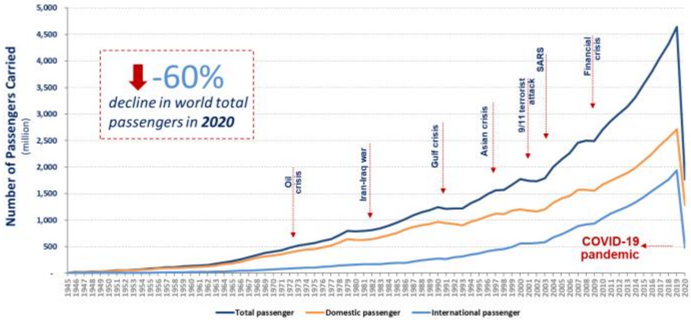


Fig 10. Passenger volume evolution worldwide (1945-2020) (ICAO 2021).

From this catastrophic scenario, initiatives such as the Open Sky Network became popular not only because they provided live data and users could see, at that moment, the reduction in air traffic, but because ADS-B messages were stored in Data Sets, which later served to meticulously determine the recovery process of operations in different locations (Koelle and Barbosa 2021; Schäfer et al. 2014). The storage capacity of the OSN shown also makes it possible to trace lines of investigation that are more local and related to indicators and socio-political decisions. Using the case of Brazil, for example, which has an approximate total of 696,000 deaths from COVID-19 (World Health Organization 2022) and has always been in the global spotlight due to the way it handled the pandemic period. It is possible to identify that the increase in traffic in Brazil increases as deaths increase but it suffers a strong reduction in departures in early 2021 due to an abrupt increase in the death rate, as shown in Figure 11.

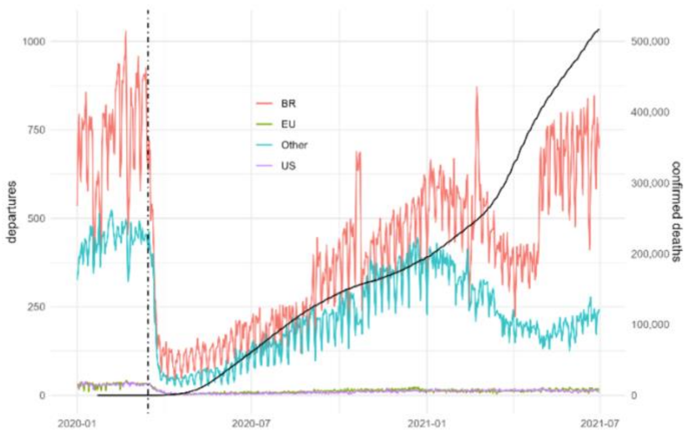


Fig 11. Regional Traffic and COVID-19 deaths in Brazil (Koelle and Barbosa 2021).

The third fundamental foundation that this work rests on is the R language and its programs, such as R Studio, R markdown and R Shiny. The R language has great relevance in this work, taking into account its great capacity to manage large amounts of data with a high volume of traffic (Big Data) in a way that the answers given by the program (graphs, tables, dashboards, for example) are highly customizable and adaptable (Kumari and Verma 2019). Therefore, using R Studio as a means to process OpenSky Network aeronautical data, the program's ability to shape and treat information with maximum flexibility and interactivity is imperative.

Within the context of the OpenSky Network and its great popularity after the pandemic period (due to the need to study the behaviour of air operations before, during and after COVID-19), an R package was created that transforms the OpenSky Network dataset into a package compatible with the R language, thus allowing research, studies and analyses to be carried out within the R universe. The package, entitled "openSkies", seeks to provide well-established and documented bases of air traffic obtained by OSN, with analysis capabilities statistics and data visualization (Ayala et al. 2021). In addition, the R language has been used in several academic and research segments, because most of these contributions are based on the concept of Reproducible Research. This consists of the ability for, for a given study, the results to be replicable, that is, another individual independent of the first will be able to find similar results with new data (Gandrud 2018). In this sense, the "openSkies" package is structured in a way that guarantees the replicability of the results, since it has a defined structure and allows the combination of different fields for the construction of the most varied analysis (Kumari and Verma 2019; Gandrud 2018; Ayala et al. 2021).

In addition to the works available and consolidated in the scientific community as sources of information and scientific bases for this project, it is important to highlight the existence of projects in development that may complement and assist the development of this project. This is the case of the thesis "Design of an ADS-B station in the OpenSky Network", which consists of the development and feeding of a database from the installation of a sensor capable of capturing the ADS-B messages transmitted by aircraft (Santos 2023). It would have the purpose similar to OpenSky Network, for example, but the development of this project – which also takes place within ISEC Lisboa – allows a greater adaptability and ease not only in obtaining the data, but also in the structuring and presentation of these, for future use not only in this project, as well as others within and beyond the institution.

In conclusion, this work is based on three main pillars: Automatic Dependent Surveillance Broadcast (ADS-B), Open Sky Network and the R language. The first is an innovative system that emerged as a replacement for RADAR, but which, for security reasons, ended up allying itself with the supposedly substituted and currently the ATM has redundancy of all information, thus reducing the error. Despite the difficulty of application in air operations due to the "ease" of suffering cyberattacks, it is a system that feeds the academic universe valuable information and that allowed the emergence of initiatives as, for example, the second pillar of this project. The Open Sky Network is a great source and a great object of study in the academic world due to its ability to provide live data as well as store them and generate a Data Set that allows carrying out different research on the behaviour of air traffic. Finally, the R language serves as the common thread of everything, as it has the ability to receive, process, shape and present the data that goes from ADS-B to the Open Sky Network and then to R Studio.

3. Modelling

The third chapter of this document aims to describe the processes that culminated in the fulfilment of the objectives initially presented, based on the concepts and studies discussed. In order to describe the whole modelling process of this project, the main spheres of Big Data concept will guide and serve as reference: Data Engineering, Data Science and Data Visualization, considering that all the processes, steps and procedures that occurred meet one of these niches, as outlined in Figure 12.

Fig 12. Layers of Big Data.

Firstly, in the field of Data Engineering, a series of choices were made in order to delimit the capabilities of the tool to be created, based on the stipulated objectives. Humberto Delgado Airport (LPPT) was determined as the focus of the study, given that it is immersed in the reality of ISEC Lisboa's students and faculty, in addition to being an airport that receives a large amount of traffic on a European and global scale. In addition to the airport, the year 2022 was determined to be analysed, given that it is the most recent completed year, and that it does not have major traffic contamination due to the COVID-19 pandemic, which ensures greater reliability to the analyses carried out.

Delimited the temporal and geographical characteristics, within the R Studio, it was imperative to a select a library capable of providing the program the necessary skills to access and manipulate the data stored by the OpenSky Network. In this sense, the library "openskyr" was chosen, considering its ability to obtain data from the OpenSky Network through four different functions: *"get_airport_data"*, *"get_flights_data"*, *"get_state_vectors"* and *"get_track_data"*. In the scope of this project, the first, *"get_airport_data"* and the last, *"get_track_data"*, were used: the first one connects to the OSN database, generates dataframes (tables) for arrivals or departures from a certain airport and a certain time period; as the second one provides aircraft routes variation, in latitude/longitude and altitude.

Considering Data Engineering processes for the first function, it is brought to the R environment and returns a dataframe structured in a default manner, so it is going to be named “Initial Structure”, considering it has columns that are of no use in terms of the objectives of this work, in addition to missing necessary columns, which will be added in the future through other support files. Even if it is a primary structure for presenting the data, it is already possible to identify the type of information and the library’s and OpenSky Network’s capabilities, as shown in Table 1.

TABLE 1. Type of data for Initial Structure of “get_airport_data” (OpenSky Network 2021)

Property	Type	Description
icao24	string	Unique ICAO 24-bit address of the transponder in hex string representation. All letters are lower case.
firstSeen	integer	Estimated time of departure for the flight as Unix time (seconds since epoch).
estDepartureAirport	string	ICAO code of the estimated departure airport. Can be null if the airport could not be identified.
lastSeen	integer	Estimated time of arrival for the flight as Unix time (seconds since epoch)
estArrivalAirport	string	ICAO code of the estimated arrival airport. Can be null if the airport could not be identified.
callsign	string	Callsign of the vehicle (8 chars). Can be null if no callsign has been received. If the vehicle transmits multiple callsigns during the flight, we take the one seen most frequently
estDepartureAirportHorizDistance	integer	Horizontal distance of the last received airborne position to the estimated departure airport in meters
estDepartureAirportVertDistance	integer	Vertical distance of the last received airborne position to the estimated departure airport in meters
estArrivalAirportHorizDistance	integer	Horizontal distance of the last received airborne position to the estimated arrival airport in meters
estArrivalAirportVertDistance	integer	Vertical distance of the last received airborne position to the estimated arrival airport in meters
departureAirportCandidatesCount	integer	Number of other possible departure airports. These are airports in short distance to estDepartureAirport.
arrivalAirportCandidatesCount	integer	Number of other possible departure airports. These are airports in short distance to estArrivalAirport.

From this initial scenario, it was stipulated that two dataframes would be created from this function, one for departures, named “LPPT_D”, and another for arrivals, named “LPPT_A”. Even though this division separates the flights’ volume, all analysis will be mirrored, using all available resources from both sources. After the creation of the two dataframes, which outlined the skeleton of the project, it was necessary to manipulate the Initial Structure in a way that met the requirements of the R language in order to fulfil the objectives: removal of columns, change of the data type of certain columns, creation of new columns and, finally, the addition of external files that support the information that initially exists.

After formatting the data coming from the Initial Structure, it was important to provide the two dataframes with higher quality and accuracy to the information that already existed. For example, the column "icao24" indicated the aircraft that operated the flight through the transponder code – unique for each aircraft – however, it was not possible to understand the type of aircraft operated (B737, A320, E195, for example). The same goes for the "callsign" column and the "estDepartureAirport" column (or "estArrivalAirport", depending on the dataframe: the information provided needs an external support to have greater value. Therefore, files external to the R and the "openskyr" library were uploaded so that it is possible to relate a transponder code ('icao24'), a 'callsign' or an aerodrome code, for example, to less technical and more accessible information, that is, the type of aircraft, the company that operates it and the location of the airport.

The first file added, entitled "Airline Codes", in .xlsx format (Excel), presented the code of a company ("Airline"), associated with the full name ("Airline Name") and the Country of Registration (|| TAP Air Portugal || TAP || Portugal ||, for example). In the R environment, it was possible to associate the 'callsign' column of the LPPT_A dataframes and "LPPT_D" to the 'Airline' column of the external file, through the "merge" function, which uses common terms of different data sources to join them. Thus, the arrivals (LPPT_A) and departures ("LPPT_D") dataframes began to present not only the company's three-lettered code, but also its respective name, in full.

In addition to "Airline Codes", the "Aircraft Database" file was also incorporated, in .xlsx format, which has data such as the code "icao24", "Aircraft Type" and "Wake Turbulence Category" (|| 3c70aa || A332 || Heavy ||, for example). From the association using the common term "icao24", it is possible to provide LPPT_A and "LPPT_D" dataframes the aircraft type information, in addition to the Wake Turbulence Category, a parameter of extreme importance in the context of traffic separation within Air Traffic Management.

Finally, the "Airport Codes" file, also in Excel format, was added in order to associate the column "ICAO Departure" (or "ICAO Arrival", depending on the dataframe) to the country where the airport is located through the name ("Destination Country" or "Origin Country"), as well as the latitude and longitude of the airport, for further geographical identification.

From all the processes described that modified the Initial Structure, we obtained the final dataframes "LPPT_A" and "LPPT_D", with 21 columns. Finally, Table 2 illustrates the data type for each column, for both "LPPT_D" and "LPPT_A".

TABLE 2. Type of data for Final Structure

Column	Type	Description
'ICAO Arrival'	string	Arrival airfield, in ICAO format.
'ICAO Departure'	string	Departure aerodrome, in ICAO format.
'Destination/Origin Country'	string	Country of origin or destination of the flight ("LPPT_D" and LPPT_A, respectively).
Longitude	numeric	Longitude of the airport of destination or origin ("LPPT_D" and LPPT_A, respectively).
Latitude	numeric	Latitude of the airport of destination or origin ("LPPT_D" and LPPT_A, respectively).
callsign	string	Flight identifier issued by the aircraft.
Airline	string	First three characters of the callsign, which indicate the airline.
'Airline Name'	string	Full name of the Airline.
Country	string	Airline's base country.
icao24	string	Aircraft transponder identifier.
'Aircraft Type'	string	Aircraft type identifier.
'Wake Turbulence Category'	string	Turbulence Mat Category.
'Date Hour'	date	Date (YYYY/MM/DD) and Time (HH:MM:SS) of flight departure or departure (LPPT_D and LPPT_A, respectively).
Year	string	Year (YYYY), obtained from the 'Date Hour' field.
Month	string	Month, in full, obtained from the 'Date Hour' field.
Day	string	Day, obtained from the 'Date Hour' field.
Hour	string	Time of day, obtained from the 'Date Hour' field.
Date	date	Date (YYYY/MM/DD), obtained from the 'Date Hour' field.
Time	string	Full time (HH:MM:SS), obtained from the 'Date Hour' field.
'Period of the Day'	string	Period of the day (Dawn, Morning, Afternoon, Night).
Weekday	string	Day of the week, in full.

The second function of the "openskyr" library, *"get_track_data"*, obtains the trajectory from the ADS-B messages transmitted of a given aircraft operating a certain flight. It is possible to notice that the return of the function gathers the positional data of the aircraft in 4 components: latitude, longitude, barometric altitude and heading. Therefore, *"get_track_data"* is able to compile, for each entry, the position, altitude and direction of the same aircraft and, as the messages transmitted by the ADS-B system have a high frequency, it is possible to trace a route both in the lateral and vertical profile very precisely, taking into account the time intervals shown in Table 3. The use of this function on a large scale, that is, automatically applied to a series of aircraft, allows the creation of a database capable of locating, through latitude and longitude, the trajectory of many aircraft simultaneously.

TABLE 3. Type of data for Initial Structure of "get_track_data"

Column	Type	Description
'icao24'	string	Aircraft transponder identifier.
'Callsign'	string	Flight identifier issued by the aircraft.
'time'	string	Date and Time, in Epoch Time, of the transmission of the message.
'longitude'	string	Longitude of the aircraft at the time of transmission.
'latitude'	string	Latitude of the aircraft at the time of transmission.
'baro_altitude'	string	Barometric altitude of the aircraft.
'true_track'	string	Direction of the aircraft at the time of transmission.
'on_ground'	string	Whether the system identifies whether the aircraft is on the ground.

Keeping in mind the above Structure, the LPPT_TRK_FLIGHTS dataframe was created, derived from the LPPT_A dataframe, to separate all the aircraft that will be subject to the "get_track_data" function to obtain trajectories for arrivals at Lisbon Airport. This filtering was also due to limitations of the function, which is not able to connect to the OpenSky Network API for all aircraft contained in the LPPT_A dataframe, which has 83543 records. Thus, only the flights that arrived at Lisbon Airport on September 15, during the afternoon, that is, between 12:00:00 and 19:00:00, were selected for this analysis. The generated dataframe had twenty-two inputs, that is, twenty-two flights landed on LPPT during the afternoon. Then, such dataframe was submitted to the "get_track_data" function through an iteration, that is, the function repeats for each identifier 'icao24' that appears in the "LPPT_TRK_FLIGHTS" dataframe. The result of this cycle generated the "LPPT_TRK" dataframe, containing the trajectories of each aircraft in the initial table. That is, for each one of the twenty-two flights presented in "LPPT_TRK_FLIGHTS", the new dataframe created contained the entire trajectory of the first of these flights, followed by the entire trajectory of the second, and so on. Finally, "LPPT_TRK" would be subdivided between "LPPT_TRK_VER", which contains the altitude variation information of the flights obtained (vertical profile) and "LPPT_TRK_LAT", containing the latitude and longitude variation data of the flights (lateral profile).

Taking into account an established data source capable of providing information and analysis and also meeting the objectives of the project, the modelling process moves on to the second plane of development, Data Science. With regard to this field inside Big Data, in the framework of volume analysis, we start from the Structures of the "LPPT_A","LPPT_D" and "LPPT_TRK" dataframes to perform measurements, that is, to transform large number of entries into information, valuable within the scope of this project.

Considering "LPPT_A e "LPPT_D" dataframes, it is possible to divide the data columns into four classifications: geographical, company, aircraft, and temporal. The geographical columns, in particular 'ICAO Arrival', 'ICAO Departure', 'Destination/Origin Country', 'Longitude' and 'Latitude', locate the flights that arrive or depart from Humberto Delgado Airport from the coordinates. The ability to geographically locate the airport of origin or destination allows a more direct view of the number of flights departing from or arriving at Lisbon Airport, on a map, for example.

The columns that give values for the airline, which are 'callsign', 'Airline' and 'Airline Name', allow the aggregation of flights by airline. Such grouping ensures a greater understanding of companies' behaviours according to the temporal and geographical metrics (most visited destination, most present time of day at the Airport, for example).

The aircraft columns – 'icao24', 'Aircraft Type' and 'Wake Turbulence Category' – are able to group arrivals and departures by aircraft type and Wake Turbulence Category. Understanding the pattern of arrivals and departures from aircraft has great associated value, taking in consideration that part of the operations within an airport are limited from the Wake Turbulence Category, which determines the separation between aircrafts, both at take-off and landing.

Finally, the temporal parameters, namely ‘Date Hour’, ‘Year’, ‘Month’, ‘Day’, ‘Hour’, ‘Date’, ‘Time’, ‘Period of the Day’, ‘Weekday’, favours the analysis of traffic patterns according to the different timescales obtained from the granularities presented by the OpenSky Network API.

Considering a route analysis framework, the LPPT_TRK dataframe allows us to generate a geographical visualization of the trajectories of the aircraft that arrive in Lisbon, both by the subdivision "LPPT_TRK_LAT" and by "LPPT_TRK_VER". The first form of visualization allows us to understand traffic patterns, holdings, shortenings and stretches of approach paths, that is, normal procedures in a high-traffic TMA. The second, on the other hand, allows to observe more or less accentuated descent profiles for approach, in addition to possible holding procedures in the decrease of altitude. Although it is a small sample of flights, this allows us to observe in greater detail some flight events such as those described above.

Thus, it is possible to identify the capabilities that the data provided by OpenSky Network have within the scope of this project. Such identification allows to generate forms of visualization of these data in different frames, but with an aspect consistent with its properties. In this context, the concepts of Data Visualization are applied, that is, the set of processes that culminates in a visualization of the data in a way that transmits a message, a message that would not have the same effect if presented through large raw databases.

In this sense, starting from LPPT_A and LPPT_D, the temporal parameters were used to construct graphs that would allow to represent a temporal granularity, that is, the volume of arrivals and departures, but in different time references. Using the bookstore "ggplot2", such visualization was developed through a bar graph that, in addition to the representation as a function of time, was also presented, for each column, the division of such volume in Wake Turbulence Category. Thus, it was possible to observe not only the traffic patterns as a function of the volume, but it was also possible to qualify the information, providing another axis of visualization that, as already discussed, has high value with regard to air traffic management in terminal areas, which is the WTC. Figure 13 represents the graph of volume of arrivals at Lisbon Airport as a function of the hours of the day and with subdivision of the volumes by WTC, exemplifying the process described above of the construction of such visualizations, which vary according to the temporal granularity (Month, Day of the Week, Period of the Day and Time of Day). The temporal granularity then allowed the creation of four graphs for the arrivals, as well as four graphs for the departures.

In addition to the visualization of the volumes through the graphs, from the LPPT_A and LPPT_D dataframes, it was also possible to geographically position the flights origin or destination airport or country from flights that passed through Lisbon Airport. For that purpose, the visualization from maps were chosen, both for airports and for countries. Using libraries available in R such as "ggplot2" and "rnaturalearth", the analysis of volumes by Airports was done for Europe/Northern Africa, while the framework by countries was also done for the region presented, but also on a global scale. In this sense, the maps were developed, both for departures, that is, the volume of traffic that receives the flights that leave from Lisbon Airport, and for the arrivals, that is, the volume of traffic that sends flights to Lisbon, both positioned geographically by countries or by airports. Figure 14 represents one of the visualizations generated from the processes described above.

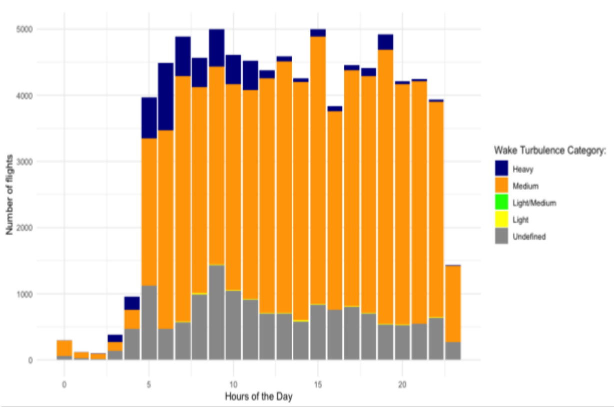


Fig 13. Number of Flights by Hour of the day graph.

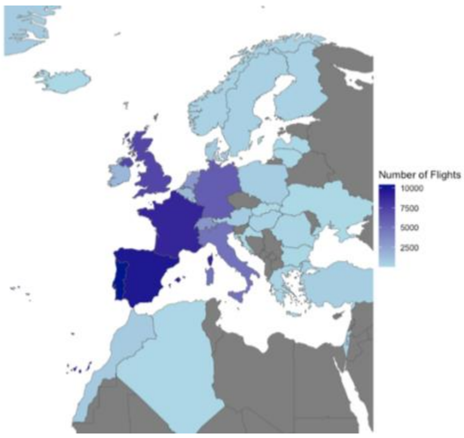


Fig 14. Number of Flights per country map.

Also, table visualizations were also developed from LPPT_A and LPPT_D that, even having the same format and presentation of large databases as the dataframes, aggregates data so that they acquire greater capacity to provide information. Thus, the constructed tables were also mirrored for arrivals and departures and indicated the volume of flights according to Aircraft Type ('Aircraft Type'), Airline ('Airline'), Airport of Origin or Destination ('ICAO Departure' or 'ICAO Arrival') and

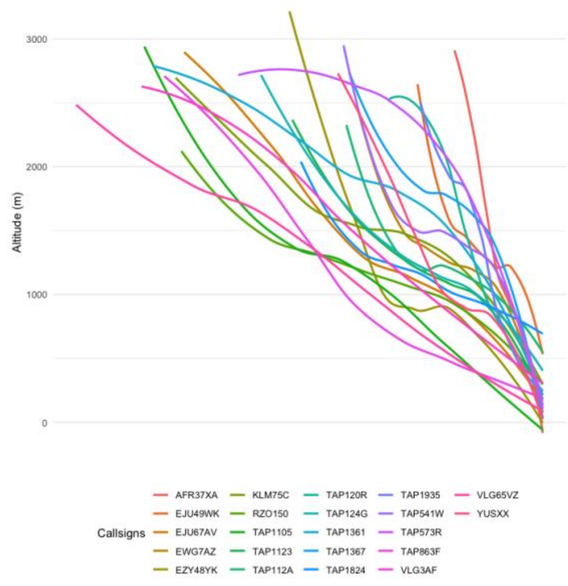
Country of Origin or Destination ('Origin Country' or 'Destination Country'). In addition to the indication of volumes, some other metrics were added to the tables, such as the most commonly operated aircraft, the most common country to be operated, the most common airline, as illustrated in Table 4, which represents volume of arriving flights per airport.

TABLE 4. Example of a table. Number of flights per Departure Airports

ICAO Departure	Number of Flights	Most common Airline	Most common operated Aircraft
LEMD	5097	TAP Air Portugal	A320
LFPO	3481	TAP Air Portugal	B738
LPMA	3461	TAP Air Portugal	A320

In addition to the LPPT_A and LPPT_D dataframes, which generated the visualizations described above, the LPPT_TRK dataframe, as previously discussed, has the structure to present arrivals at Lisbon Airport from the point of view of the trajectory of the aircraft. Considering that the analysis of the trajectory for landing at an airport can be carried out from a lateral profile point of view as well as the vertical profile point of view, LPPT_TRK was divided between LPPT_TRK_VER, containing the altitude variation information for the flights; and LPPT_TRK_LAT, which presents latitude and longitude variation data over time for the same flights.

Thus, for LPPT_TRK_VER, using the "ggplot2" library, a line graph was drawn that illustrates the descent profile of the flights, from the last 10,000 feet (3,000 m), separating each callsign by colour for better identification of the different flights, and allowing to visualize the different vertical trajectories, as shown in Figure 15. Secondly, for the LPPT_TRK_LAT dataframe, using the "leaflet" library, an approach similar to the LPPT_TRK_VER approach was used, in that the variation data of, in this case, latitude and longitude, would be responsible for tracing the routes of each flight arriving in Lisbon. The mentioned library is responsible for the use of maps overlaid with the desired data, that can be moved and zoomed by user's input. It was possible, therefore, to develop this form of visualization to represent the lateral trajectory for approach and landing at Lisbon Airport, as illustrated by Figure 16



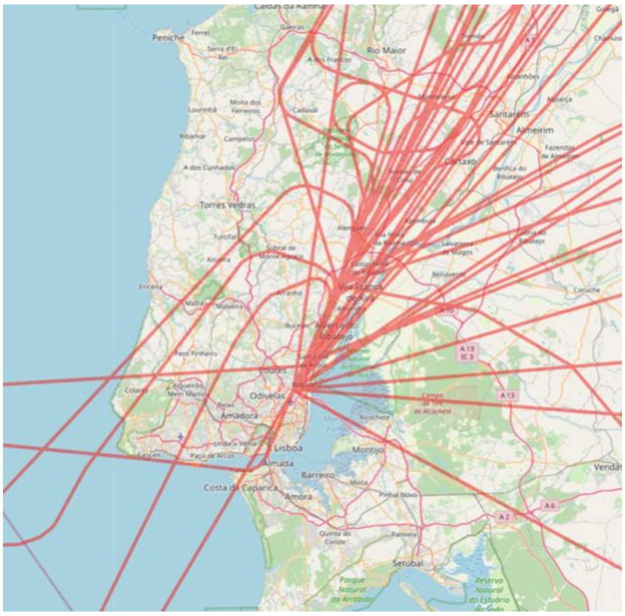


Fig 15. Vertical profile graph.

Fig 16. Lateral profile map.

After the completion of the development of the content responsible for filling the dashboard, it is time to structure and develop such interactive environment. The "rshiny" and "shinydashboard" libraries were used, which associate the R language with the ability to build dashboards. All visualizations (graphs, maps and tables) needed to be placed and presented in a way that was consistent with their respective properties and data type. Therefore, the dashboard was structured with a side panel that divided the contents by arrivals ("LPPT_A"), departures ("LPPT_D") and routes ("LPPT_TRK"), as well as a home page that welcomes the user to the dashboard, that is, the creation of subdivisions that allow a clearer navigation. Therefore, the side panel consists of the page "Home", "Departures Volume", "Arrivals Volume" and "Arrival's Routes". In addition, the pages "Departures Volume" and "Arrivals Volume" were divided into two more dub-pages, "In Time" and "Geographically", as illustrated in Figure 15.

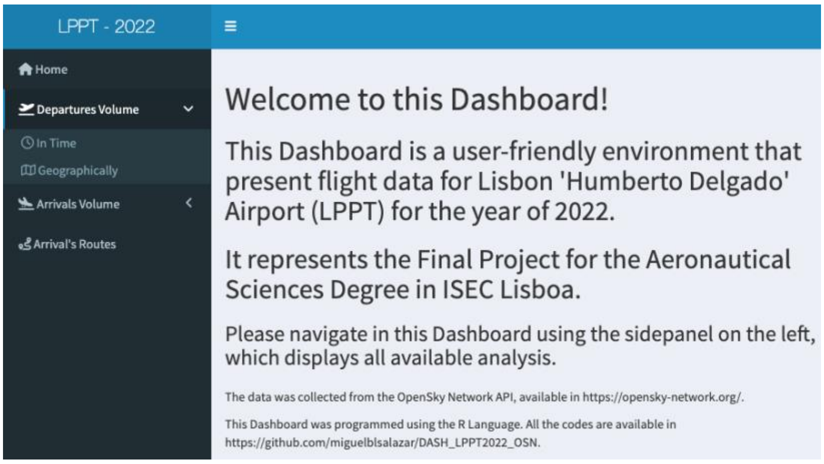


Fig 15. Dashboard Initial page and Side Panel.

Therefore, the development of this project, described throughout Chapter 3, was based on the three main concepts of Big Data: Data Engineering, Data Science and Data Visualization. The first, Data Engineering, consisted of obtaining the data from the OpenSky Network API, followed by adapting the data according to the objectives of this project, supporting them in auxiliary information sources that qualified the data. From this process, the "raw materials" of this project were obtained: the “LPPT_A”, “LPPT_D” and “LPPT_TRK” dataframes. The second concept, Data Science, consisted of the process of understanding the properties, characteristics, and capabilities of each piece of information that the original dataframes provided. Finally, Data Visualization, consisted of all the procedures that aimed to present the desired information so that visualizations were in accordance with what the selected data is able to offer. The aggregation of these 3 stages allowed, finally, the development of the data processing system from the OpenSky Network for the Lisbon region, in a dashboard format in R language.

4. Results

Chapter 4 of this project refers to the evaluation of everything that was obtained in the light of the stipulated objectives and the literature studied. It contextualizes, once again, the scenario to which this project is part of, placing the tool developed in this same scenario; It also evaluates the effectiveness and depth of the processes that culminated in the implementation of this tool, in addition to the effectiveness, performance and availability of the tool itself; The capacities and limitations of the dashboard in the light of the proposed objectives; Connects the project to the literature presented throughout Chapter 2, pointing out links and connections; Discusses key concepts that shaped the tool; It relates this project with others in development and already consolidated, in order to perpetuate the methodology implemented for a broader application.

Firstly, it is again referenced that this project is immersed within the context and concepts of Big Data. That is, the existence of a great capacity to collect relevant information in data format suggests, therefore, the need for similar capacity of treatment and processing the data, transforming large databases into valuable information. With regard to this project, which consists of the Development of a data processing system from the OpenSky Network for the Lisbon region, OSN has the ability to collect, position and identify aircraft around the globe. This project, on the other hand, aims to collect such data and qualify them, that is, give them the ability to transmit a message. Such messages are related to metrics and parameters that are directly connected to Air Traffic Management, such as the volume of landings and take-offs, as well as the displacements – vertical and lateral – in terminal zones. From the qualification of the data in information, it was carried out the presentation of such visualizations through the developed dashboard.

Secondly, it is important to evaluate the efficiency and step-by-step procedures that culminated in the creation of the dashboard, namely Data Engineering, Data Science and Data Visualization. The Data Engineering process, which allowed the obtaining of the dataframes that represented the raw material of this work – LPPT_A, LPPT_D and LPPT_TRK – was marked by the choice of the library "openskyr" and subsequent immersion of this in the R environment through the functions used (Gasco, 2020) get_airport_data for the first two dataframes and get_track_data, for the latter). The format offered by the author's pre-definition (Initial Structure, referenced in TABLE) allied to the following processes of modelling and associating data with external files allowed initial sources of information (source dataframes) assume the ideal format and aligned with the objectives of the project. When it comes to Data Science, this process made it possible to obtain greater clarity on the capacity of the data provided by the OSN associated with external sources of supplementary information. Consequently, the structuring of thinking focused on the visualizations that would be generated in the future became much more oriented and objective. Finally, the Data Visualization field was nothing more than the materialization of the previous processes, a kind of crowning of all the work done in the format of graphs, maps and tables that were, of course, in accordance with the properties of each piece of information and aligned with the objectives and assumptions of this tool.

Thus, we can perceive not only the importance of these elements within the concepts of Big Data, but also realize that despite FIGURE CANVA illustrates a linear process, the construction of this project has always shown itself as a kind of foresight of what one would like to visualize in function of the properties obtained, and in function of what was built as a foundation, initially. That is, the linearity shown is not verified, whereas, for example, much of what was accomplished in the phase of Data Engineering He also had in mind the "final product", that is, visualization, and the same goes for the scientific process in function of engineering as well as visualization.

The result of the processes described above materialized in the format of the dashboard, illustrated in the FIGURE DASHBOARD. This interactive environment, already available for navigation, study and use, meets the objectives proposed by this project as it provides volume analysis (arrivals and departures) as well as a geographical analysis of routes to Lisbon Airport in 2022. The choice of the year 2022 – being a "completed" year and with little influence of COVID-19 – allowed the realization of realistic analyses throughout an entire year, allowing the understanding of traffic patterns for an airport of great importance in the European and global scenario. It is, therefore, an interactive tool, but also with a real-time feature, that is, the extrapolation of this study to another period or another airport automatically changes the views presented on the dashboard. In order to guarantee complete access to the information that supports this project, a repository was created on the Github platform, which stores and presents the codes used and allows its use for new studies. The project is available through the link https://github.com/pffmachado/flightir (Salazar and Filipe Faria Machado 2023).

From this context related to the dashboard, it is possible to evaluate its capabilities and limitations. It is an interactive environment that allows the user to act as a selector of the parameters to change the visualization of the data (temporal scales, geographies, company parameters, aircraft, for example), thus allowing different ways of observing the same data and even different ways of merging data from different visualizations; the visualizations presented are able to value the data presented, to the extent that they are consistent with the message they want to convey. The different ways of presenting the volume of take-offs and landings for Lisbon Airport, as well as a third approach focused on trajectories, allow a more holistic view of operations. Using as an example the Volume graph as a function of time scales allows to measure traffic patterns in the most varied readings of time (by month, to understand seasonality; by day of the week, to understand weekly flows; and by period and hour of day, to understand the operations with greater granularity and precision).

On the other hand, it is possible to identify limitations in the developed tool, such as the considerable discrepancy between the arrivals recorded by the dataframe LPPT_A and departures recorded by dataframe LPPT_D. The discrepancy between values may be associated with the way OpenSky Network registers the Airports of Origin and Destination: by the presence of other aerodromes in the vicinity of Lisbon Airport, the database can associate the destination as any aerodrome close to LPPT (LPCS – Cascais Aerodrome, LPAR – Alverca Aerodrome, LPST – Sintra Aerodrome, for example). In the same way that flights departing from Lisbon may have an ICAO code of destination airport not consistent with the actual aerodrome of destination. Despite this discrepancy, the analysis performed does not deviate from the expected and a coherence of the data is maintained. In addition, the OSN API does not provide a direct way to allow an analysis of delays, that is, there is no comparison between the scheduled time and the actual time of arrival or departure. Therefore, it can be concluded that the developed tool has great processing capabilities and presentation of large databases but is subject to some irregularities and discrepancies that are related to sources external to this project.

Bearing in mind the characteristics and properties of such a tool, the focus of this analysis and discussion is directed to the connection between this project and all the literature presented and developed throughout the Chapter **Error! Reference source not found.**. Despite all the collection of documents, reports, papers and books dialogue directly or indirectly, there are four that have a direct connection with the development of this project. "Assessing the Global COVID-19 Impact on Air Transport with Open Data" is a study that, similar to this project, makes use of the environment promoted by the R language and its libraries to assemble analysis and visualizations coming from the OpenSky Network (Koelle and Barbosa 2021). The search for understanding the impacts of the COVID-19 pandemic on air transport directed the project to the OSN database, where it was able to search for volumes of flights in different regions of the Globe in order to establish comparisons between countries, with regard to political, economic and social strategies. The document "Assessing the Global COVID-19 Impact on Air Transport with Open Data", therefore, relates directly to this project in view of the methodology through the R environment and relates more specifically to the development of dataframes LPPT_A and LPPT_D, responsible for providing volumes of arrivals and departures to Lisbon Airport.

On the other hand, more focused on the analysis of trajectories that, in this project, was approached through the dataframe LPPT_TRK, it is possible to highlight two documents responsible for providing references and guidance for the implementation of the section of this project. "A Geometric Approach to Study Aircraft Trajectories: The Benefits of OpenSky Network ADS-B Data" is a study that, despite having a more focused framework on the geometry of trajectories and comparison between the data sources of EUROCONTROL and ADS-B (Perrichon, Gendre, and Klein 2022), provides examples that dialogue with the objectives of this project. For example, the visualization of the arrivals of flights arriving from LFBO (Toulouse-Blagnac) in LFPO (Paris-Orly), relate directly to the analysis presented in the dashboard regarding the Lateral Profile of arrivals at Lisbon Airport, as they present flight trajectories with a high degree of accuracy, providing the possibility of viewing the trajectories in high-traffic TMAs.

In addition to the analysis of trajectories in a vertical profile, the document "Evaluation of Flight Efficiency for Stockholm Arlanda Airport using Open Sky Network Data" provides a vertical profile analysis of a given aircraft – also comparing data sources (OpenSky, EUROCONTROL, for example) – in order to better understand fuel consumption, delays and waits in TMA and lack of ATM optimization for Stockholm Arlanda Airport (Polishchuk, Lemetti, and Sáez 2019). The FIGURE VERTI. represents the vertical profile for the approach of flight SAS 410 and, from mathematical models capable of understanding the greater or lesser consumption of fuel, it is possible to study the efficiency of such Airport, including its traffic management. This representation is one of the bases of support of the analysis carried out in the dashboard for the vertical profile, from the flights that arrived in Lisbon during the afternoon of September 15, 2022.

Addressing a more general theme of this work, which is the structuring of codes, files and visualizations in function of the concept of Reproducible Research, this can be seen as one of the great foundations within the pillar of R Studio, in view of its great relevance with regard to propagating this study to other airports, other time periods and other lenses. The book "Reproducible Research with R and RStudio – Second Edition" seeks to present this concept, which is nothing more than the ability of a research, study or analysis, to be independently reproduced by other individuals, so that the results are similar or coherent (Gandrud 2018). In view of this project, both the writing of the codes – in a cohesive and coherent way – with explanatory comments, combined with a well-consolidated and publicly available file structure on "Github" platform, it can be said that this project meets the concepts explored by Gandrud. That is, the possibility of creating a dashboard for the airport of Porto (LPPR), Frankfurt (EDDF) or Rio de Janeiro (SBGL), without the main structure being changed, being possible to analyse volumes of arrivals and departures, lateral and vertical trajectories.

Therefore, this project was able to consolidate and formalize the relationship between the OpenSky Network API, using a dashboard for Lisbon Airport as a materialization of Big Data concepts, in the light of previous studies that, directly or indirectly, dialogue with the central themes developed here. In addition, the structuring of this project also involves the possibility of reproducing it in other scenarios and models, maintaining a base structure.

5. Conclusions

In the fifth and final Chapter, the initial premises established will be resumed and recommendations and possible projections of this project will be presented. Therefore, this consisted in the development of a data processing system from the OpenSky Network for the Lisbon region. This ideal materialized in a dashboard responsible for presenting the volume of arrivals, volume of departures and analysis of trajectories for Lisbon Airport in 2022. Among the objectives established in addition to the creation of this tool, was the search to remedy the lack of an intermediary between the user and large databases such as OpenSky Network, FlightRadar24, for example, in addition to providing ISEC Lisboa and the aeronautical community as a whole, an environment capable of presenting information on air traffic patterns to Lisbon.

From such initial conditions, were searched in the academic environment sources that were based on the same themes and concepts addressed in this project and that, consequently, helped in its constitution of the scientific bases, ensuring not only credibility, but also a guidance focused on concepts established in the scientific community. The sources ensured the premises of this project, justifying the need for the creation of a data processing system, an interactive visualization environment and the construction of this entire process due to the reproducibility of research.

Thus, the development of this project was supported by the community and unfolded from three concepts, or layers, within Big Data: Data Engineering, Data Science and Data Visualization. The first facilitated the creation of the raw materials of this project: the LPPT_A dataframes, regarding arrivals at Lisbon Airport; LPPT_D, which shows departures from Lisbon Airport; and LPPT_TRK, which presents the lateral and vertical trajectories of a clipping of flights that arrived in Lisbon. This process consisted mainly in the use of the R language and association, in the programming environment, between the "openskyr" library, materialized in the functions “get_airport_data” and “get_track_data”, and external files of aeronautical bodies responsible for giving more precision to the data presented. Data Science, in turn, was responsible for organizing the raw materials acquired and understanding what kind of information and message these can offer, that is, understanding the temporal and geographical component and the parameters aimed at the Aircraft and the Airline. This process was also already aimed at sketching and experimenting with the combination of different components and metrics, ensuring a coherent visualization with high informational value. Finally, Data Visualization consisted of the use of the R environment to generate graphs, maps and tables that would later be presented in an interactive environment. For this, the bookstores "rshiny" and "shinydashboard" were used to concretize this tool in the R Studio environment. In the end, the result was a dashboard that presents the volume of arrivals and departures, as well as a presentation of trajectories for flights that passed through Lisbon Airport in the year 2022.

As part of the perpetuation of this project for other contexts in scenarios, the publication of the visualization codes as well as the publication of the dashboard code on the Github platform, free of charge, assists in obtaining such information and propagating this approach to the currently available databases. Considering the collaborative character of the Big Data universe, it should be noted that different sources of information can – and should – be used by this interactive platform. Scaling this tool can be of great use for similar studies in other scopes.

Still within the concept of collaboration and taking into account the development of the "Project of an ADS-B station in the OpenSky Network" within ISEC Lisbon itself, the dialogue between these two projects can generate a more precise tool that receives information in a way that is according to demand, without the need for conversions or external files. The development of this database allows, therefore, greater malleability with regard to the structuring of the data that feed this project and other future ones (Santos 2023). This collaboration within ISEC Lisboa itself allows the concepts explored in this and other complementary projects to go beyond the walls of the institution, meeting one of the main objectives not only of this, but of part of academic projects: the possibility of validation and reproduction in another niche.

6. References

Abdulaziz, Abdulrazaq, Surajo Namadi, Abdulmalik Yaro, and Ashraf Ahmad. 2019. “Surveillance Radar System Limitations and the Advent of the Automatic Dependent Broadcast System for Aircraft Monitoring.” www.atbuftejoste.com.

Ayala, Rafael, Daniel Ayala, Lara Sellés Vidal, and David Ruiz. 2021. “OpenSkies - Integration of Aviation Data into the R Ecosystem.” *The R Journal* 13 (2): 590–99.

Baek, Kwang-Yul, and Hyo-Choong Bang. 2012. “ADS-B Based Trajectory Prediction and Conflict Detection for Air Traffic Management.” *International Journal of Aeronautical and Space Sciences* 13 (3): 377–85. https://doi.org/10.5139/ijass.2012.13.3.377.

Gandrud, Christopher. 2018. *Reproducible Research with R and RStudio*. Edited by Chapman and Hall/CRC. 2nd ed.

Gasco, Luis. 2020. “Luis Gasco - Openskyr.” 2020. https://github.com/luisgasco/openskyr.

Gilbert, Glen A. 1973. “Historical Development of the Air Traffic Control System.” *IEEE Transactions on Communications* 21 (5): 364–75. https://doi.org/10.1109/TCOM.1973.1091699.

ICAO. 2021. “2020 Passenger Totals Drop 60 Percent as COVID-19 Assault on International Mobility Continues .” 2021.

ICAO Asia. 2007. “Guidance Material on Surveillance Technology Comparison.”

Koelle, Rainer, and Fabio Lourenco Carneiro Barbosa. 2021. “Assessing the Global COVID-19 Impact on Air Transport with Open Data.” In *AIAA/IEEE Digital Avionics Systems Conference - Proceedings*. Vol. 2021-October. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/DASC52595.2021.9594337.

Kumari, Anita, and Neeraj Verma. 2019. “AN ANALYTICAL REVIEW STUDY ON BIG DATA ANALYSIS USING R STUDIO.” *International Journal of Engineering Technologies and Management Research* 6 (6): 116–22. https://doi.org/10.5281/zenodo.3266146.

Miranda, Miguel. 2022a. “Primary Radar.” *Air Dream College*.

———. 2022b. “SSR & Transponder.” *Air Dream College*.

Open Sky Network. 2022. “About the OpenSky Network.” 2022. https://opensky-network.org/.

OpenSky Network. 2021. “OpenSky Rest API.” 2021. https://openskynetwork.github.io/opensky-api/rest.html#arrivals-by-airport.

Perrichon, Rémi, Xavier Gendre, and Thierry Klein. 2022. “A Geometric Approach to Study Aircraft Trajectories: The Benefits of OpenSky Network ADS-B Data.” In *Engineering Proceedings*. Basel Switzerland: MDPI. https://doi.org/10.3390/engproc2022028006.

Polishchuk, Tatiana, Anastasia Lemetti, and Raúl Sáez. 2019. “Evaluation of Flight Efficiency for Stockholm Arlanda Airport Using OpenSky Network Data.” *Proceedings of the OpenSky Workshop 2019: EPiC Series in Computing* 67: 13–24.

Rodrigues, Claudia Vanessa Coutinho. 2010. “ADS-B - Automatic Dependent Surveillance Broadcast. Estudo Do Impacto Em Portugal.”

Salazar, Miguel, and Paulo Filipe Faria Machado. 2023. “Flighir - Github Repository.” 2023. https://github.com/pffmachado/flightir.

Santos, João. 2023. “Projeto de Uma Estação ADS-B Na Rede OpenSky Network.” ISEC Lisboa.

Schäfer, Matthias, Vincent Lenders, and Ivan Martinovic. 2013. “Experimental Analysis of Attacks on Next Generation Air Traffic Communication.” In *International Conference on Applied Cryptography and Network Security*.

Schäfer, Matthias, Martin Strohmeier, Vincent Lenders, Ivan Martinovic, Matthias Wilhelm, and T U Kaiserslautern. 2014. "Bringing up OpenSky: A Large-Scale ADS-B Sensor Network for Research." In *IPSN-14 Proceedings of the 13th International Symposium on Information Processing in Sensor Networks*, 83–94. http://tinyurl.com/cwsyr2r.

Schäfer, Matthias, Martin Strohmeier, Matthew Smith, Markus Fuchs, Vincent Lenders, Ivan Martinovic, OpenSky Network, and Switzerland lastname. 2018. "OpenSky Report 2018: Assessing the Integrity of Crowdsourced Mode S and ADS-B Data." http://opensky-network.org.

Schultz, Michael, Judith Rosenow, and Xavier Olive. 2022. "Data-Driven Airport Management Enabled by Operational Milestones Derived from ADS-B Messages." *Journal of Air Transport Management* 99: 102164. https://doi.org/10.1016/j.jairtraman.2021.102164.

Seymour, William. 2018. "Optimising the OpenSky Network."

Strohmeier, Martin, Xavier Olive, Jannis Lübbe, Matthias Schäfer, and Vincent Lenders. 2021. "Crowdsourced Air Traffic Data from the OpenSky Network 2019-2020." *Earth System Science Data* 13 (2): 357–66. https://doi.org/10.5194/essd-13-357-2021.

Strohmeier, Martin, Matthias Schäfer, Vincent Lenders, and Ivan Martinovic. 2014. "Realities and Challenges of NextGen Air Traffic Management: The Case of ADS-B." *IEEE Communications Magazine*, 2014.

Sun, Junzi, Luis Basora, Xavier Olive, Martin Strohmeier, Matthias Schäfer, Ivan Martinovic, and Vincent Lenders. 2022. "OpenSky Report 2022: Evaluating Aviation Emissions Using Crowdsourced Open Flight Data." https://opensky-network.org.

Wesson, Kyle D, Todd E Humphreys, and Brian L Evans. 2014. "Can Cryptography Secure Next Generation Air Traffic Surveillance?" In *IEEE SECURITY & PRIVACY*.

World Health Organization. 2022. "WHO Coronavirus (COVID-19) Dashboard." 2022.