Journal of Open Aviation Science (2023), Vol.**1**
doi:10.59490/joas.2023.7201

JOAS

PROCEEDINGS | *The 11th OpenSky Symposium*

Compilation of an open-source traffic and CO_2 emissions dataset for commercial aviation

Antoine Salgas,[*,1] Junzi Sun,[2] Scott Delbecq,[1] Thomas Planès,[1] and Gilles Lafforgue[3]

[1]ISAE-SUPAERO, Université de Toulouse, France
[2]Faculty of Aerospace Engineering, Delft University of Technology, the Netherlands
[3]TBS Business School, Université de Toulouse, France
[*]Corresponding author: antoine.salgas@isae-supaero.fr

(Received: 24 October 2023; Revised: 18 December 2023; Accepted: 17 January 2024; Published: 20 January 2024)
(Editor: Martin Strohmeier; Reviewers: Antonio Franco and Max Li)

Abstract
The study of the environmental transition of the aviation sector calls for prospective traffic scenarios. Detailed traffic and emissions inventories are often needed to refine the available analyses and to enable the simulation of regionalised scenarios. In the past studies, these are generally based on commercial, proprietary traffic data, making their dissemination problematic and reducing the reproducibility of the science produced. Open-source alternatives do exist, but with limited geographical coverage. This paper presents a method to aggregate different sources of flight information, in order to obtain an open-source air traffic dataset for 2019. Then, missing flight information is identified and completed using an airline route database built from Wikipedia parsing and related socio-economic data. After that, several reference datasets are used to evaluate the accuracy of the extended open-source dataset. Despite varying accuracy for different routes, major traffic flows are reasonably well estimated at the country and continental levels. Finally, the CO_2 emissions are obtained using an existing aircraft performance surrogate model, and the accuracies are examined compared to the results from previous studies.

Keywords: Open-data; Emissions; Air Traffic

Abbreviations: ASK: Available Seat Kilometres, IATA: International Air Transport Association, ICAO: Internation Civil Aviation Organisation

1. Introduction

In the context of climate change, it is necessary to quickly reduce greenhouse gas emissions to limit global warming consequences, and it requires the implementation of mitigation strategies across all business sectors [1]. Although commercial aviation currently contributes to only 2.6% of those emissions, the trend is for this proportion to increase [2]. This is due to the air traffic expected growth [3, 4] and the limited technological options to increase aircraft efficiency [5]. Achieving deeper decarbonisation requires in particular the use of Sustainable Aviation Fuels (SAF) such as biofuels, electrofuels or hydrogen if produced using low-carbon energies [6, 7]. However, this raises other issues, such as the consumption of resources like biomass and electricity, which other sectors are also looking to as means of decarbonisation [8].

Some of these mitigation measures are likely to be more expensive for the air transport industry. For example, the widespread use of sustainable aviation fuels could increase the operating expenses of

airlines by around 40% by 2050 [9], meaning that public policies such as taxes or subsidies will be necessary to foster their adoption. Different options exist to allow the implementation of such policies, such as blending mandates or aircraft efficiency regulations. The various low-carbon energies considered are not all equivalent, from both a sustainability and an economic point of view [9]. Choosing the adequate option could lead to substantial improvements in the energy transition efficiency [10]. This highlights the need for a detailed multidisciplinary evaluation of different prospective energy transition scenarios. Such work is done by both industrial actors [6, 7] and academia [11, 12, 13, 14, 15, 16].

A common requirement for prospective scenarios is to have emissions inventories in a base year from which trends can be projected to estimate future emissions. Such inventories could be based on detailed commercial flight schedule databases that are not open-source [17] or on the total fuel consumption of the sector, for instance from International Energy Agency (IEA) [18]. This solution, despite being open-source and allowing free dissemination of data, is not detailed enough to capture the geographical disparities of air transport and the associated different growth perspectives [3, 4]. Similarly, regional analysis capacities could be relevant when it comes to biomass or electricity characteristics or to better replicate the various coverages of existing legislative measures.

This work presents a methodology for estimating air traffic flows for a given year (2019) with an acceptable level of accuracy, based exclusively on open-source data. This paper is also part of the development of AeroMAPS, a dedicated open-source prospective scenario simulator used for instance in [19], because it especially aims at developing its regionalised assessment capacity in the future.

First, the data sources used are introduced and the data pipeline used is presented step by step in Section 2. The resulting dataset is then evaluated and validated in Section 3, before concluding remarks and perspectives are given in Section 4.

2. Method

2.1 Data collection and aggregation methodology

The overall process for obtaining an open-source database is described in this section. 2019 has been chosen as the reference year for building this database, as the following 3 years are largely disrupted by the consequences of COVID-19 [20]. The main objective of the process is to obtain, for each air route, the associated traffic volume, and if possible, the aircraft used to ultimately estimate the associated CO_2 emissions. The traffic metric used is the number of seats available on each route, rather than the number of passengers because the former is a better proxy of the number of flights. This section summarises the overall process represented on Figure 1. The different steps are briefly described in the following paragraphs, and in more detail in dedicated sections.

There are numerous open-source datasets available, but none of them provide global coverage, which is only available from commercial sources. They are described in Section 2.2. As a first step, the chosen approach is to combine those datasets in order to achieve the greatest possible spatial coverage. In order to address overlapping sources, a prioritisation logic based on source characteristics is introduced.

Although the completeness of the combined dataset compared to individual sources is improved, it remains incomplete. To fill this gap, a specific method is proposed in Section 2.3. A comprehensive but disaggregated data source is used: the collaborative encyclopaedia, Wikipedia. In fact, there is a recommended design pattern for airport pages which includes a section that lists all the destinations served from the airport, along with the airlines [32]. This information is easily accessible to the author of an airport's Wikipedia article, as it is often available on the website of the airport.

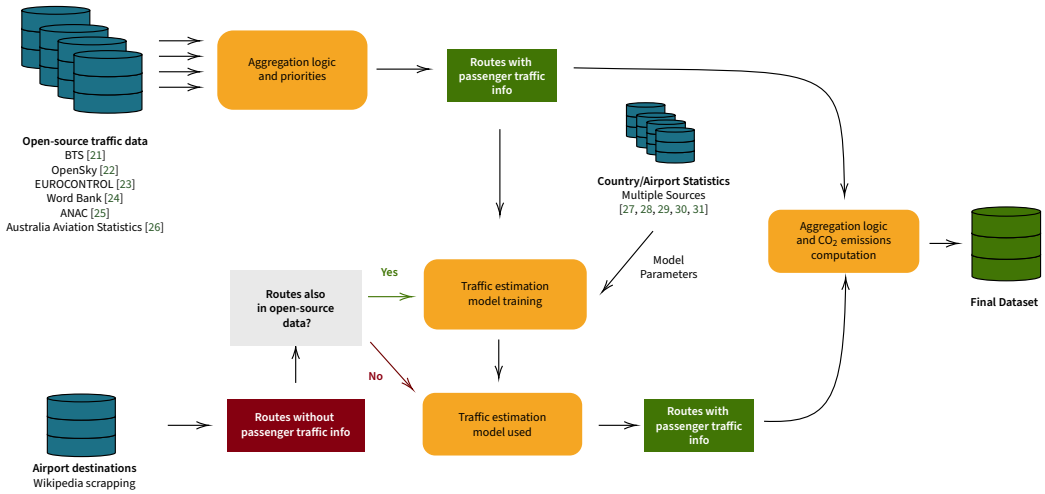

Figure 1. Traffic dataset creation flowchart.

Automated retrieval of these lists is relatively easy, provided that the list of Wikipedia URLs associated with commercial airports is available. This process provides a much more comprehensive list of routes than was previously available. This is described in Section 2.3.1. However, there is no information on the seating capacity associated with each route or the frequency of flights. It must therefore be estimated. To do this, the open-source data mentioned above is used again, this time to train a regression model. It uses economic, geographical and statistical data associated with the airport and/or country of origin and destination of each flight. The sources and models used are described in Section 2.3.2 and 2.3.3. Once the training is complete, it is possible to determine the traffic on each route found previously.

Concerning traffic data, the final step is to aggregate this estimation with open-source data, prioritising the latter where available. The method is described in Section 2.4.

Lastly, CO_2 emissions are computed as explained in Section 2.5. If the aircraft model is known, a surrogate aircraft performance model [33] is used, and otherwise, the fleet average consumption per seat at the corresponding distance is used.

2.2 The starting point: available open-source data

Various open-source flight data are available online, as shown in Table 1, but each one has its own limitations. The estimation of air transport CO_2 emissions often do not require detailed trajectory of individual flights. Instead, a relatively aggregated level of data is usually sufficient.

The first category of sources comes from administrations (Adm.). The format offered by [21] is particularly adapted, with each data item compiling all the monthly flights of a given airline, with a given aircraft type, on a given route and with the associated payload. However, the extent of the database is limited to flights going to and from the United States of America (USA). A relatively similar dataset is made available by the World Bank [24], and includes all the international flights. The drawback in this case is that the database does not provide information on the airlines and aircraft used. It is not a maintained database, with only a single edition, with the most recent data items corresponding to 2019, which corresponds to the studied year in this paper. Brazilian [25] and Australian [26] civil aviation authorities also provide such information with various levels of information as it can be seen in Table 1.

The other category of sources comes from radar or ADS-B monitoring of flights. This is offered by [22, 23], the latter being open-access for academia but not fully open-source. The former is completely open-source and based on a collaborative ADS-B collection network but still lacks coverage[1]. Those radar sources do not feature payload information but they provide information on the aircraft used and its operator contrary to most administrative sources. In this case, the payload can be retrieved using the average seating capacity of each aircraft type (using an aircraft database made available for academia[2]) and average load factors.

Table 1. Various characteristics of the open-source references considered.

Source	Coverage	Collection	Route	A/C Type	Airline	Payload	Ref.
BTS T-100	To/from USA	Adm.	✓	✓	✓	✓	[21]
OpenSky	Global (partial)	Radar	✓	✓	✓	-	[22]
Eurocontrol	To/from EU	Radar	✓	✓	✓	-	[23]
World Bank	International	Adm.	✓	-	-	✓	[24]
ANAC	Brasil	Adm.	✓	-	✓	✓	[25]
AUS Stats	Australia	Adm.	✓	-	-	✓	[26]

2.3 Filling the gaps: estimating traffic on unreferenced routes

2.3.1 Creating a route database

As explained in the introduction of this section, the Wikipedia pages of airports are used as a source to establish a complete route network, without however knowing the traffic on each route. The first step is to reference all airports served by commercial airlines available on Wikipedia. A community-based list of airports served is available for each continent [34]. The related URLs of airport Wikipedia pages are retrieved by parsing the list using an HTML analysis library[3]. For each airport found, another Python script opens the URL and analyses the HTML content of the page to find the "Airline and Destinations" section. This section contains a table with the Wikipedia URLs of all the airports served by each airline from the explored airport. This data is stored and, after iterating over all the airports, a complete route database can be established. The associated code and further explanations on this data parsing step are given in the associated Jupyter Notebooks (see Reproducibility statement).

Note that an important limitation of the process lies in the dynamic nature of Wikipedia pages, which are constantly updated by the community. No viable option was found to extract the data at a given point in the past, and therefore the parsed route database is used as it was in April 2023. Another important point to note when working with Wikipedia is that, despite the community's proofreading efforts, errors may still persist in some airports. Others may be missing from the list of airports used in the first place, which was subsequently merged with the list of destinations to reduce errors.

2.3.2 Route database feature enrichment

The previously established list of airports is enriched by adding relevant features to build a regression model, using routes included in the open-source data to train a model.

The first relevant set of features is directly related to each airport. Besides collecting airport's destinations in the previous step, the passenger traffic, aircraft movements and airport codes of each

[1] The actual coverage can be seen on https://opensky-network.org/network/facts
[2] https://www.planespotters.net
[3] https://pypi.org/project/beautifulsoup4

airport are also collected on the Wikipedia pages of the airport. A Wikidata item is also linked to each airport page[4]. Similar data (for example, annual passenger traffic, ICAO and IATA codes) can be found, while an advantage of Wikidata is that the fields are dated, compared to Wikipedia "last-year available" information. Not every airport has all these features and the list is filtered to retain only airports with an IATA code. Airport geographical coordinates, countries and continental codes are added by merging an airport database [35] with the airport list. When it comes to estimating the traffic on a given route, besides the airport traffic itself, the neighbouring population of airports could be relevant, especially when the airport traffic information is not available from Wikidata or Wikipedia sources. A global population dataset [30] is used to determine the population in the vicinity of airports at three levels (in hexagons inscribed in 30, 70 and 150 km radius circles). Similarly, the number of competing airports in the same vicinity regions is calculated.

The second group of features used refers more to the countries themselves. With a correlation between the number of kilometres flown per inhabitant and the country's wealth [36], some socio-economic metrics are used. The Gross Domestic Product (GDP) per capita (Power Purchase Parity) is used to capture the raw wealth of each country [29]. The inequality structure is captured by the Gini coefficient of incomes [29] and the Inequality-adjusted Human Development Index (IHDI) [31]. Since tourist countries are more likely to be served by more flights, the number of inbound and outbound tourists, as well as the share of tourism in the exports of each country is also used as a feature [29]. Their surface and their insularity are also used because one can think that the size of the country affects the number of domestic flights [28, 29].

Those airport/country-related features are added to the previous airport dataset. The completed airport database is merged into the route database and route-related features are added to the dataset. This final group of features deals with bilateral relations between airports on the different routes. These include the bilateral trade flows [27], the number of airlines serving each route (established during the Wikipedia list parsing) and the great circle distance between the two airports.

2.3.3 Traffic estimation model

The route database is completed by a dependent variable: in this case, the number of seats available on each route. To do so, the open-source datasets described in Section 2.2 are used. They are merged into the Wikipedia-parsed route database. Note that trying to infer the aircraft type used or its operator was not achieved for the sake of simplicity.

For the values taken by the dependent variable, it is more suited to use the various administrative sources. Indeed, using radar sources requires converting a number of flights in a number seats offered per route with average aircraft capacities and load factors, which reduces the data accuracy. Moreover, in the case of [22], a general trend towards traffic underestimation was found when the concerned relation was also included in another dataset. Figure 2 demonstrates this trend comparing radar data with BTS data, and the same trend is observed when comparing with the World Bank dataset. It could be explained by the fact that only a partial number of flights on a given route were detected by the ADS-B collection network. Indeed, either the origin/destination or aircraft could be unknown (or mismatched) for some flights on a route, resulting in a capacity underestimation once data is aggregated and compared to an administrative complete source. Note that OpenSky coverage has surely been improved since 2019. The same phenomenon can be seen for Eurocontrol data although with a different pattern: either the data is well correlated, or not at all, suggesting that some individual flights were included in the Eurocontrol dataset without being in its nominal coverage zone. Note that the joint coverage of BTS and Eurocontrol datasets is limited to only the Europe-USA flights, explaining the relative scarcity of data points in the comparison of Figure 2.

[4]For instance, Toulouse-Blagnac item can be found at: https://www.wikidata.org/wiki/Q372615

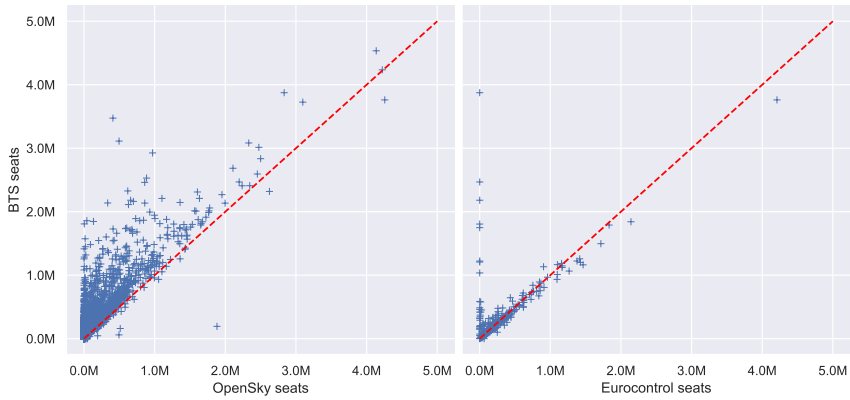

Figure 2. Number of seats per route: radar data [22, 23] comparison with BTS [21]. Contains only data in both datasets.

The priority order retained is thus [21, 24, 25, 26, 23, 22], using first the administrative although less detailed sources and then the radar sources. This is achieved because only the dependent variable is of interest at this point and not the other details provided by the dataset.

After merging those sources, 41% of the Wikipedia-scrapped routes remain traffic-undetermined and will be estimated. To do so, several regression methods are tested using the 59% traffic-determined routes and a typical 80%-20% train-test data split.

First, a linear regression is performed using Eq. (1), where S_{AB} are the seats offered between cities A and B, α_i the regression coefficient relative to feature F_i. The results are inadequate despite an acceptable r^2 for the purpose (between 0.3 and 0.5, depending on the train/test split). Indeed, a simple linear regression induces some negative estimates. It means those values should be forced to zero afterwards not to include "negative" seats available. Moreover, very highly frequented routes are highly underestimated as it can be seen in Figure 3. Since no particular caution was taken in selecting non-colinear features, regularisation techniques (lasso regularisation) are tested without improving the metrics. In fact, without regularisation, many coefficients are already null. The relationship between the dependent variable and the predictors therefore appears to be non-linear.

$$S_{AB,lin} = \alpha_0 + \alpha_1 F_1 + ... + \alpha_n F_n \tag{1}$$

Then, a reduced gravity model, given in Eq. (2) where P_K is the population in city K, I_K the income per habitant, D_{AB} the distance between two cities and x_P, x_I, x_D the relative log-linear regression coefficients, is tested to simply account for potential non-linearities, but it is also insufficient (Figure 3). A very low r^2=0.05 is found. More features could have been added to improve this, with however large restrictions on the data entries used (no zeros allowed). This path was not investigated.

$$S_{AB,log-lin} = \frac{(P_A \times P_B)^{x_P} \times (I_A \times I_B)^{x_I}}{D_{AB}^{x_D}} \tag{2}$$

Instead, more sophisticated machine learning methods are tested, starting with a random forest[37]. This combines several weak random regression trees to produce a reliable estimator of the dependent variable little prone to overfitting. The r^2 is improved to 0.7 (depending on the train-test split). Compared to the linear regression, there are no more negative estimates. However, the random forest regressor does not handle missing values (NaN), just like previous regressors. Therefore,

data entries with missing features must be removed from the dataset, or the missing feature can be imputed arbitrarily (using for instance a defined quantile: here the first 1000-quantile was used, following the idea that missing features are more likely to be found on small airports).

Some regression algorithms that also handle missing values could be used to avoid this intervention on the dataset. It is the case of XGBoost [38], a tree-based gradient-boosted regression method. XGBoost was tested and provided fast and slightly improved results over the random forest. The fast training speed allowed testing on several random train-test splits and the r^2 is between 0.65 and 0.75 on most tests. A specific loss function (following a Tweedie distribution) was chosen to prevent negative estimates and to allow large amounts of routes to have low traffic.

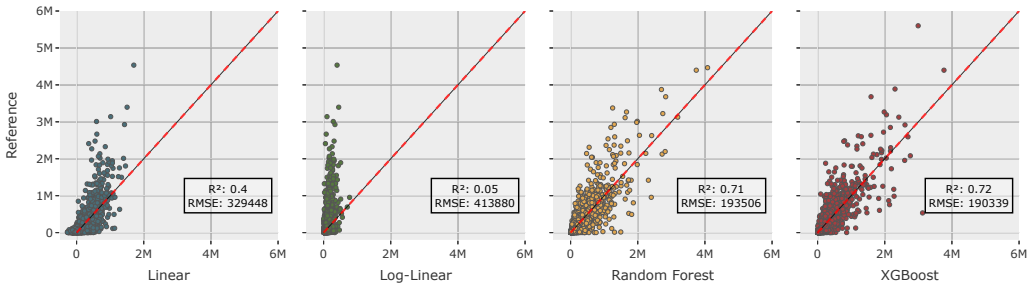

Figure 3. Seat capacity estimation: regression techniques test.

Due to its ease of use and to avoid missing feature imputation (not having a feature is information in itself), the XGBoost regressor is selected to estimate the number of seats on unknown routes. It is trained a last time on the whole known dataset, before its final usage on traffic-undetermined routes. The relative importance of each feature used by the regression model is given in Appendix 1.

2.4 Final dataset aggregation

This section presents the final dataset aggregation method which consists of two stages.

In a first step, the open-source databases mentioned in Section 2.2 are used again, but this time as a way to construct an aggregated database combining the different sources rather than to complete the Wikipedia-built route database. The aggregation logic differs from the one given in Section 2.3.3. In this previous case, the accuracy of the dependent variable (number of seats available per route) was of interest rather than the exact model of aircraft flying each route. However, if one is interested in CO_2 emissions as a primary goal, it is more relevant to have access to the aircraft type, rather than the exact payload. Indeed, the surrogate fuel burn model used, which is the one preferred for accurate estimations (see Section 2.5), requires the aircraft type and the flight distance, not the payload. Moreover, having as much aircraft and airline information as possible could be of interest in performing airline network analysis for instance. This means that the radar sources should be favoured over the administrative sources. To avoid the underestimation phenomenon described in Section 2.3.3 while keeping aircraft information, radar sources are prioritised, but with an administrative source as a validation backup. This backup is used to decide if the radar data is chosen or not for each route in the aggregated dataset. If the gap with the backup is too high, the administrative source is used, losing the aircraft-type information. The priority order is [21, 22, 23, 25, 24, 26].

In a second step, the estimated data from Section 2.3.3 is used to complete the route database with the additional routes missing in the open-source databases. Since the scope of this work aims at having a relatively reliable estimate of air traffic at the regional level rather than the route level, an

extra scaling step is performed on these estimated data. It consists in using once again the passenger traffic information of each airport PAX_{AP_i} parsed on Wikipedia. Once all flights $PAX_{FL_{AP_i-AP_j}}$ of open-source data going to or from this airport are accounted for, there could be a residual traffic $\Delta_{PAX_{A_i}}$ as shown in Eq. (3). This value should correspond to the estimated data. In this case, γ_{PAX,AP_i} of Eq. (4) would be equal to 1. If this is not the case, in an iterative process, each route capacity is multiplied by its origin and destination airport scaling factor as shown in Eq. (5). Bounds are specified to restrict this relatively rough process. It converges very quickly (8 rounds) to a minimal residual three times lower than originally. The effect on airport residuals is shown in Figure 4. Note that the process could degrade the route-level accuracy by altering the results obtained with the estimator.

$$\Delta_{PAX_{AP_i}} = PAX_{AP_i} - \sum_{Open-source,AP_j} PAX_{FL_{AP_i-AP_j}} \tag{3}$$

$$\gamma_{PAX,AP_i} = \Delta_{PAX_{AP_i}} / \sum_{Estimated,AP_j} PAX_{FL_{AP_i-AP_j}} \tag{4}$$

$$PAX^*_{FL_{AP_i-AP_j}} = PAX_{FL_{AP_i-AP_j}} \cdot \gamma_{PAX,AP_i} \cdot \gamma_{PAX,AP_j} \tag{5}$$

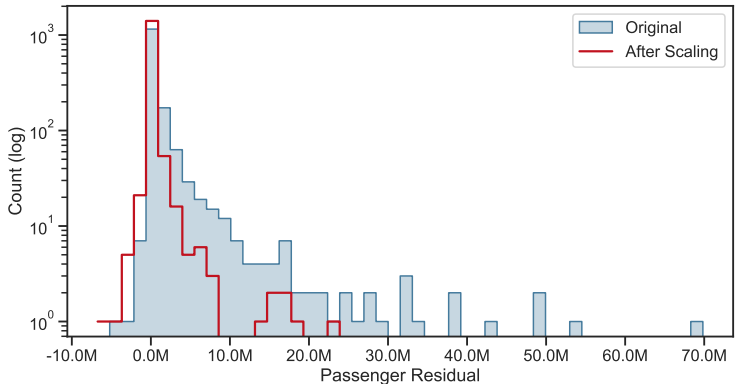

Figure 4. Estimated flights scaling effect on airport passenger traffic residual.

Finally, the corrected estimated data is aggregated on top of the open-source data. The number of seats attributed to each source is given in Table 2.

Table 2. Final source distribution in the compiled dataset.

Source	BTS	OpenSky	Eurocontrol	W.Bank	ANAC	AUS.	Estimation	Total
Mn Seats (%)	1295 (23.2)	207 (3.7)	1346 (24.1)	862 (15.5)	133 (2.4)	69 (1.2)	1657 (16.5)	5570
Bn ASK (%)	3027 (28.4)	551 (5.2)	2738 (25.7)	2338 (21.9)	172 (1.6)	79 (0.7)	1758 (16.5)	10664

It is also interesting to compute the Available Seat Kilometres (ASK), which is a widely used traffic metric. It is obtained by multiplying the number of seats available on each route by the corresponding great-circle distance.

2.5 From traffic to fuel burn to CO_2 emissions

The previous work focused on estimating traffic data on each route. The process of estimating the associated fuel burn requires using an aircraft fuel burn model. Several models are available at different levels of fidelity. For instance, OpenAP [39] is a detailed open-source model, but requires the real flight path to determine the aircraft fuel burn. This level of information is not available in the current dataset. Therefore, a very simplified surrogate model, FEAT [33], is used. It requires only the knowledge of the aircraft type and the distance of the flight to estimate the fuel burn. The aggregated error at the fleet level is reported to be below 5%.

Two cases are present in the aggregated dataset of this paper. Some of the collected sources have an aircraft model associated with each data entry: it is the case for [21, 22, 23]. As illustrated in Table 2, these entries represent 51% of the total seats offered and 59.3% of the ASK. FEAT can be applied directly to the data items to compute the associated fuel burn. However, in the case of the other sources and of the estimated data (representing 49 and 40.7% of the seats and ASK), the aircraft type information is not known. Therefore, a regression is performed on the aforementioned FEAT-computed data points to derive a fuel burn per seat function of the flight distance (Figure 5). To account for the fact that the data points represent a variable number of flights, this regression is weighted according ot this variable. It gives a satisfactory level of fidelity for the use case considered, with a high weighted r^2 of 0.95, but it should be reminded that the regression is based on a surrogate model and not the actual fuel burn. Some outliers can be seen on Figure 5, and especially at lower ranges with a group of minor data entries whose fuel burn increases quickly compared to the trend. They are related to the quality of [21] data, in which the seats offered are specific to each item. It therefore includes VIP charter flights with very low cabin density for which the fuel burn per seat increases rapidly. This effect cannot be seen for the other sources, in which an average value per aircraft type is considered. The fuel burn corresponding to each remaining route is then computed using this regression. The associated equation is given by Eq. (6).

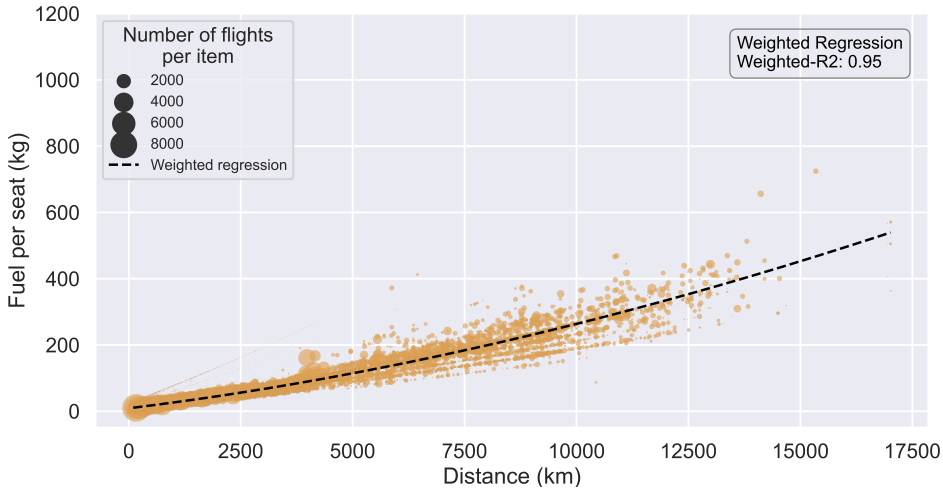

Figure 5. Fuel burn regression on FEAT-computed data points.

$$F_{seat}(d) = 9.07 + 1.65 \cdot 10^{-2}\ d + 9.43 \cdot 10^{-7}\ d^2 - 4.76 \cdot 10^{-12}\ d^3 \tag{6}$$

Lastly, the CO_2 emissions are immediately computed from the fuel burn using an emission factor of 3.16 $kgCO_2/kg_{fuel}$.

3. Final dataset accuracy and validation

The quality of the final dataset is evaluated at several aggregation levels and using different comparison references. A dedicated testing notebook is provided (see Reproducibility statement) in order to explore in detail the accuracy of the final dataset, specifically at the country and airport levels. It includes interactive figures giving access to details of the analyses mentioned below.

The first solution for evaluating the final dataset is to use the IEA World Energy Statistics dataset [18]. It gives the actual kerosene consumption for domestic and international aviation for each country. However, general aviation, all-cargo and military aircraft consumption is often included in the data. At the global level, the fuel consumption of the final dataset represents 80.7% of the IEA data. According to [36], general and military aviation accounts for 12% of the inventory and cargo 17%. This figure includes all freighter flights, not included in the dataset of this work, and those flights account for approximately half of the cargo emissions [17]. It means the emissions covered by our dataset would be around 80%, consistent with the ratio found above. On the individual country level, it is more difficult to reach a conclusion given that the breakdown between the different types of aviation mentioned above is disparate from one country to another. For example, the coverage rate for kerosene consumption in the USA is 74%, even though the data is of good quality (BTS). Military and general aviation are very present in the country. In contrast, the coverage of a country like Japan is 97%.

The second solution is to compare the results with air traffic data from the International Air Transport Association (IATA). While IATA reports 4543 million seats sold or 5506 million seats available with an 82.5% load factor [40], there are 5570 million seats available in the open-source dataset, i.e. a difference of around 1%.

The third solution is to use OAG data. OAG is a commercial data provider which offers global and detailed flight schedules, making it a useful dataset to assess the accuracy of the global dataset, despite using 2018 data instead of 2019 due to cost constraints. Seat capacities are converted to 2019-like values using a uniform growth rate for ASK of 3.4% [40]. It's a relatively rough method that ignores geographic disparities, but the differences in ASK growth were relatively limited from 2018 to 2019 [40]. The analysis shows a high correlation between the open-source data and OAG data despite the estimated data being more dispersed but apparently unbiased. For OpenSky data, a bias towards underestimation is present and was previously discussed. The r^2 between this work and OAG data is 0.74 for the seating capacity and 0.91 for the ASK. Overall, the seating capacity of the open-source dataset totals 96% of the one of OAG, while the ASK totals 101.6%. Two types of error can be distinguished at this global level. The first case is when the route is recorded by both datasets, but with different volumes: around 387 Mn seats are "lost" in this case. The second case is when routes are in either dataset but not in the other: 280 Mn seats are on routes not referenced in OAG data, and 122 Mn in the opposite case.

Airport and regional accuracy levels are also explored with OAG data. The traffic of most major airports is reasonably well computed, which is not surprising with the scaling process presented in Section 2.4. There are some notable exceptions, such as Liège airport (LGG) in Belgium, in which the ASK are overestimated by 2000%. In fact, Liège is a major cargo hub while a minor passenger airport and some cargo flights are erroneously remaining in the dataset. Thus, some cargo flights were not correctly referenced as such in the Eurocontrol dataset and considered as passenger flights. There are also errors probably driven by the estimation itself. For example, the final data set is missing around 50% of the ASK for Calcutta (CCU) or Surabaya (SUB) airports. It could be linked to two combined phenomenons: there is a significant domestic traffic in these countries meaning estimation is used a lot, and these countries are less covered by the training dataset (in which Europe and USA are over-represented, hence a bias). Going beyond these limits would require an airport-by-

airport investigation, which would be time-consuming and beyond the scope of this work, but it does show that there is still room for improvement at this level of granularity. To get closer to the dataset creation objectives (providing a coherent basis for the regionalisation of AeroMAPS), it is more appropriate to focus on accuracy at the country level. The error in terms of ASK is represented in Figure 6, showing that the open-source dataset quality is variable. However, this map does not allow visualizing the associated traffic volumes. Indeed, looking at the raw data shows that most poorly estimated countries have moderate traffic. Most European, North American and Eastern Asian countries are well accounted for. There is an overestimation in some eastern European countries, while they are normally well covered by Eurocontrol data. The possibility of an error on the OAG dataset was not investigated but as mentioned above, the OAG dataset used was a 2018 dataset, converted to 2019 values using a worldwide growth rate. It could be the reason for country-level errors.

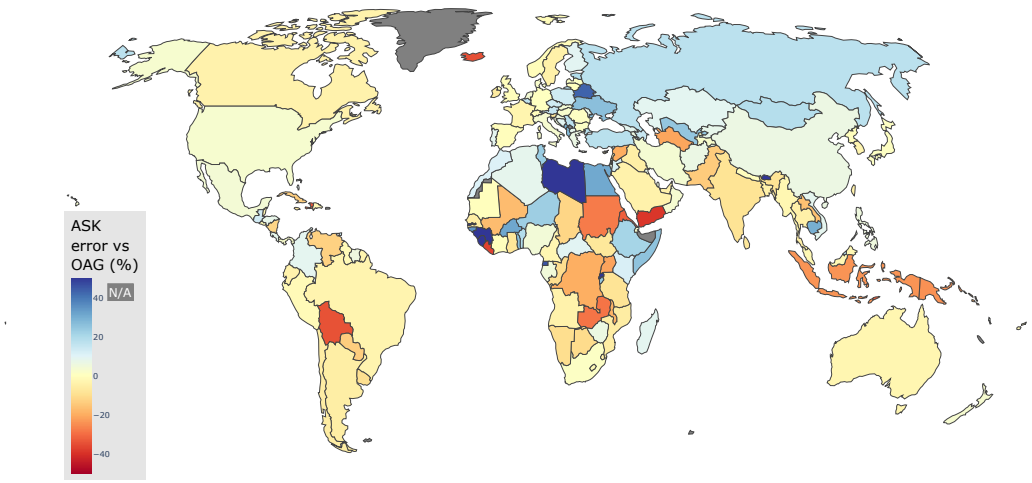

Figure 6. Relative ASK error compared to reference data from OAG.

It is interesting to draw a parallel between the source used and the quality of the estimate in Figure 6. As mentioned before, the regions for which open-source data is available are relatively well estimated. For example, we have a 2.9% error in the United States, where most of the data is from [21]; and errors of less than 3% in the majority of Western European countries, where [23] data is abundant. Australia and Brazil, for which specific datasets are used, are also very well estimated. In the rest of the world, the estimate reaches reasonable levels in major countries such as China, with a +6% error, and India, with a -9% error. The most disappointing major market is Indonesia, with 24% of the traffic missing, for the same reasons as described before, at the airport level.

This discussion raises awareness of the importance of available open data for reliable projections at the country level. The process of merging sources presented in this paper makes it possible to meet the initial objective of creating a baseline for AeroMAPS scenarios, but it would be dangerous to rely on it to make decarbonisation forecasts for some of the data-poor countries. Similarly, it should be remembered that merging multiple sources is a long and complex process and that monitoring emissions in this way year after year is tedious. A greater number of open sources, even at a fairly high level of granularity, but complete, would make it possible to provide an even more solid basis for decarbonisation forecasts without compromising data confidentiality. It would also make it possible to improve comparability between different regions of the world, which suffers here from a different data collection method, with the introduction of potential biases.

OAG data are finally used for investigating the continental flows (Table 3). Top 15 largest flows are all within 12% of OAG reference for ASK, with much better results in most of the cases (top 3 flows are within 1.2%). Minor flows are less accurate and especially Europe/Oceania. Investigation was done to find error reasons and in this case, it was found to be related to the specific nature of Europe-Oceania routes. Indeed, the very long distances involved mean that there is most often a technical stop such as Singapore for London-Sydney or Los Angeles for Paris-Papeete. Some datasets consider the two legs of the flight as separate (OAG) while some others [22, 24] do not, resulting in this large error.

Table 3. Continental ASK flows and relative error compared to OAG data.

Bn ASK (error)	AF	AS	EU	NA	OC	SA
AF	99 (-8.0%)					
AS	165 (5.4%)	2,826 (-1.2%)				
EU	402 (10.9%)	1,215 (7.4%)	1,239 (0.8%)			
NA	34 (44.7%)	628 (3.7%)	938 (3.6%)	1,811 (1.2%)		
OC	4 (-5.5%)	352 (-6.4%)	14 (627.2%)	179 (12.3%)	148 (-5.2%)	
SA	8 (19.2%)	15 (38.9%)	178 (4.6%)	157 (-6.6%)	6 (6.3%)	236 (-6.3%)

The last solution for assessing the dataset accuracy is to use an aviation emissions inventory from the International Council on Clean Transportation (ICCT) [17], available as open-source at the country level. Despite the fact that they sourced their flight data at OAG, emissions were computed using their own model (GACA), making it relevant to assess the CO_2 emissions accuracy of the work presented in this paper. Emissions are allocated to each country according to the country of departure of each flight. [17] separates cargo and passenger transport emissions, while in the present work, all passenger flights were accounted for without attributing a share of their emissions to freight transported in aircraft's holds (often referenced as "belly cargo"). To correct this perimeter difference, ICCT passenger emissions were increased homogeneously by 8.8% to cover those emissions. Table 4 provides ASK and CO_2 emission comparison at the continental level. It is interesting to note that the world level CO_2 estimation is highly accurate, although this aggregated value hides a moderate over-estimation on international flights and a more important underestimation on domestic flights. Although the trend is similar on ASK, the conversion to CO_2 emissions increases the error on domestic flights, suggesting that the fuel burn model might be optimistic on shorter flights. The accuracy levels on major flows remain however in an acceptable range for the general context of prospective scenarios. Looking at the country level, once again domestic flights in Indonesia are the poorest estimated major market (-25% CO_2). It is also worth mentioning that US-domestic emissions are underestimated by 7.5% while the ASK are underestimated by only 1.4%, which could confirm the trend of the surrogate fuel-burn model to be optimistic on shorter routes.

4. Conclusion

The work presented in this paper aims at providing a customisable fuel burn or CO_2 emissions baseline for AeroMAPS, a prospective scenario simulator for the decarbonisation of air transport. The target is to be able to generate such scenarios for regional entities or user-defined perimeters. It requires a bottom-up city pair flight dataset, which was available with a global coverage only from commercial, non-open-source providers, fiercely restricting AeroMAPS research reproducibility and diffusion. To overcome this limitation, open-source datasets (with limited coverage) were collected and compiled. However, it still does not allow to achieve satisfactory geographical coverage, partic-

Table 4. ASK and CO_2 emissions inventory at the continental level and comparison with ICCT values.

	Total (error)		International (error)		Domestic (error)	
	Bn ASK	$MtCO_2$	Bn ASK	$MtCO_2$	Bn ASK	$MtCO_2$
Asia	3857 (0.86 %)	306 (-0.3 %)	2458 (5.3 %)	189 (4.3 %)	1399 (-6.1 %)	117 (-6.9 %)
N. America	2781 (0.5 %)	224 (-3.5 %)	1176 (3.5 %)	95 (3.0 %)	1604 (-1.5 %)	129 (-7.9 %)
Europe	2729 (0.3 %)	211 (0.0 %)	2355 (0.3 %)	179 (0.8 %)	374 (0.3 %)	32 (-4.1 %)
S. America	411 (-2.1 %)	34 (-2.7 %)	228 (-0.7 %)	18 (3.0 %)	183 (-3.8 %)	16 (-8.5 %)
Oceania	337 (3.2 %)	27 (4.4 %)	236 (3.5 %)	19 (8.2 %)	100 (2.3 %)	8 (-3.9 %)
Africa	281 (15.6 %)	22 (4.6 %)	253 (19.1 %)	19 (10.3 %)	28 (-9 %)	2.7 (-22.9 %)
RoW*	269 (22 %)	21 (3.9 %)	265 (23.1 %)	20 (7.5 %)	4 (-27.3 %)	0.6 (-53.1 %)
World	10664 (1.4 %)	844 (-0.9 %)	6971 (4 %)	538 (3.3 %)	3693 (-3.3 %)	306 (-7 %)

** Generic Rest of the World (RoW) category for unspecified countries in ICCT inventory*

ularly for domestic flights outside Europe and North America. Therefore, Wikipedia airport pages, and in particular the list of destinations associated with each airport, were used to concatenate a global database of air routes. Using various features related to airports, countries and routes, an XGBoost regression model was trained on the already known routes and then used to estimate the available seats on each route of the Wikipedia-based dataset. The resulting estimated dataset was concatenated with the other open-source datasets to create a global, open-source, air traffic dataset with a capacity in terms of seats offered associated with each route. The ASK are immediately obtained, while the associated fuel burn and CO_2 emissions are estimated using a surrogate aircraft performance model. The accuracy of the dataset was evaluated using four external sources. In terms of seating capacity, it is 4% lower than the OAG total. However, it surpasses the OAG total by 1.6% in terms of ASK. In terms of CO_2 emissions, at constant perimeters (commercial passenger flights and belly cargo), the value found is only 0.9% below the one of ICCT. The results are more contrasted at lower aggregation levels. Most continental flows are correctly accounted for, as well as most major markets, while high estimation errors can remain in small domestic aviation countries. At the route level, despite a good trend compared to OAG data (r^2=0.91 for ASK), there is some dispersion and it should be remembered that this aggregation level is not the use case the dataset is designed for.

This accuracy could be improved in multiple ways. First, the dataset aggregation logic was designed to be able to accommodate more open-source data when such are made available. It could be the case with minor aviation countries whose authorities make this data available and that were not retrieved during the research part of this work. For instance, including Indian records [41] would improve both the raw sources and likely the estimation model capacities by reducing its bias towards occidental countries. Its particular format requires an important preprocessing and was therefore not used in this work. Besides that, features were selected without exhaustivity. Some could be replaced by more relevant route-specific features if there are any. The regressor has been used and parameterised with an approach that favours simplicity of use over a detailed understanding of the logic involved in regression trees. More investigation on this could improve the estimation accuracy. Classification techniques could also be used to infer the aircraft types, to improve the fuel burn estimation. The airlines could also be inferred, refining the estimation by ensuring their network is consistent with the potential of their fleets. More sophisticated approaches could also be based on the large amount of data collected for this project. Graph Neural Networks (GNN) could be used for example as a way to represent the route networks, with airports as nodes and flights as edges. The idea would be to estimate each unknown edge using both airport capacity and other features as well as known edges. That is similar to what was achieved in this paper, but using the structure of the data instead

of estimating each flight independently one from the other.

More broadly, the next step of this work is to adapt AeroMAPS architecture to handle this new customisable dataset. Besides this main use case of the dataset, this research could serve various purposes. For instance, one could imagine a sociological study on the relationships between income and propensity to travel or related topics. Finally, this work was performed for the year 2019 but extending the database to other historical years would be of interest, for instance, to monitor the progress of aviation decarbonisation. This could also help towards predictions of traffic evolution in prospective scenarios.

Appendix 1. Supplementary figures

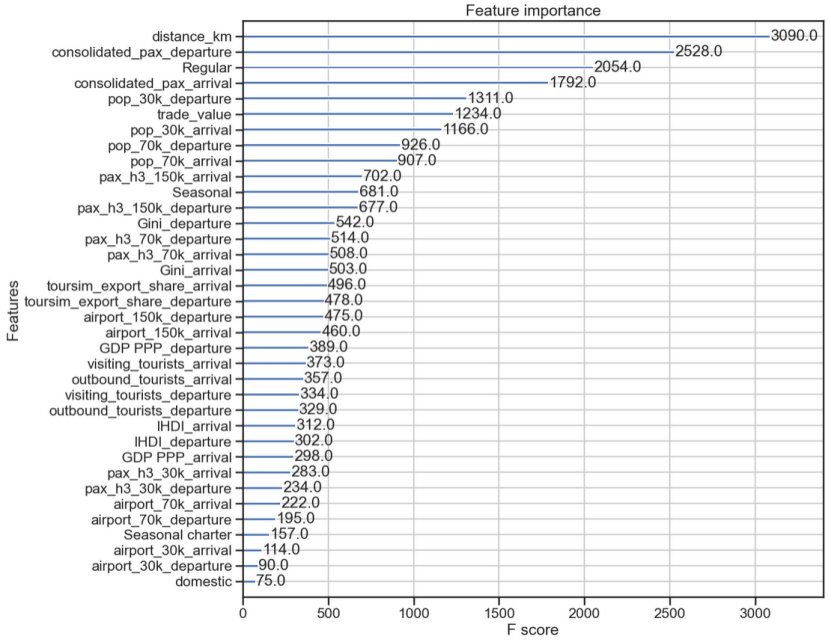

Figure 7. XGBoost feature importance.

The F-score represents the number of times the feature was selected to split the tree during the training process. The most prevalent features are the distance between the airports, the passenger traffic of airports (*consolidated_pax*) and the number of airlines (*Regular but also Seasonal, Seasonal charter*). The population surrounding each airport (*pop_X_X*) is also important, although the co-linearity of the various population metrics, in a radius of 30, 70 and 150 km around the airports could reduce the relative importance of each feature individually. The same remark is true for the total passenger traffic in the area and the number of concurrent airports (*pax_X_X and airport_X_X*). Although the trade flows are an important feature, other socio-economic features are less used by the training process. This is not surprising, given that metrics such as passenger traffic at each airport or the number of airlines are much more direct assessments of the number of passengers on each route.

Acknowledgement

The work presented in this paper was made during an international thesis mobility at TU Delft in the Netherlands. Thus, the author would like to thank ISAE-SUPAERO, EDAA, Erasmus+ as well as the Franco-Dutch network EOLE for the scholarships awarded to this end and all the members of the faculty of Aerospace Engineering of TU Delft for their hospitality.

Author contributions

- Antoine Salgas: Conceptualization, Methodology, Data Curation, Validation, Visualization, Software, Writing–Original draft
- Junzi Sun: Supervision, Resources, Validation, Visualization, Writing–Review and Editing
- Scott Delbecq: Supervision, Writing–Review and Editing
- Thomas Planès: Supervision, Writing–Review and Editing
- Gilles Lafforgue: Supervision, Writing–Review and Editing

Open data statement

All the data used for this work are included in the associated GitHub mentioned in the following section and described in depth in the README file. Not that external inputs, too large for online data storage, require user action. When applicable, these actions are described in the same file.

The dataset is archived at: https://zenodo.org/doi/10.5281/zenodo.10125898

Reproducibility statement

The source code associated with this project is stored on a public GitHub repository available at https://github.com/AeroMAPS/AeroSCOPE_dataset. It consists of four folders.

- *01_wikipedia_parser* stores notebooks used to parser Wikipedia pages.
- *02_airport_features* stores notebooks used for feature collection to prepare estimation.
- *03_routes_schedule* is where the notebooks, used for open-source data compilation, regressor training and estimation, final compilation and testing, are stored.
- Finally, a web application, developed for exploring the dataset but not detailed in this paper, is stored in another repository available at https://github.com/AeroMAPS/AeroSCOPE, and is accessible at https://aeromaps.eu/aeroscope.

References

[1] Intergovernmental Panel on Climate Change. *Climate Change 2022: Mitigation of Climate Change. Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press. 2022. URL: https://www.ipcc.ch/report/ar6/wg3/downloads/report/IPCC_AR6_WGIII_SummaryForPolicymakers.pdf (visited on 05/30/2022).

[2] Martin Cames, Jakob Graichen, Anne Siemons, and Vanessa Cook. *Emission reduction targets for international aviation and shipping*. Tech. rep. European Parliament's Committee on Environment, Public Health and Food Safety, 2015. URL: https://www.europarl.europa.eu/RegData/etudes/STUD/2015/569964/IPOL_STU(2015)569964_EN.pdf (visited on 10/17/2023).

[3] Airbus. *Airbus Global Market Forecast 2023-2042*. 2023. URL: https://www.airbus.com/sites/g/files/jlcbta136/files/2023-06/Airbus%20Global%20Market%20Forecast%202023-2042%20Presentation.pdf (visited on 07/24/2023).

[4] Boeing. *Commercial Market Outlook 2023-2042*. 2023. URL: https://www.boeing.com/commercial/market/commercial-market-outlook/index.page (visited on 08/14/2023).

[5] Scott Delbecq, Jérôme Fontane, Nicolas Gourdain, Thomas Planès, and Florian Simatos. "Sustainable aviation in the context of the Paris Agreement: A review of prospective scenarios and their technological mitigation levers". In: *Progress in Aerospace Sciences*. Special Issue on Green Aviation 141 (Aug. 2023), p. 100920. DOI: 10.1016/j.paerosci.2023.100920.

[6] Air Transport Action Group. *Waypoint 2050*. Tech. rep. 2021. URL: https://aviationbenefits.org/media/167417/w2050_v2021_27sept_full.pdf (visited on 05/30/2022).

[7] NLR, SEO Amsterdam Economics, A4E, ACI-Europe, ASD, CANSO, and ERA. *Destination 2050, A route to net zero european aviation*. Tech. rep. 2021. URL: https://www.destination2050.eu/wp-content/uploads/2021/03/Destination2050_Report.pdf (visited on 05/30/2022).

[8] Phillip J. Ansell. "Review of sustainable energy carriers for aviation: Benefits, challenges, and future viability". In: *Progress in Aerospace Sciences*. Special Issue on Green Aviation 141 (Aug. 2023), p. 100919. DOI: 10.1016/j.paerosci.2023.100919.

[9] Antoine Salgas, Thomas Planès, Scott Delbecq, Florian Simatos, and Gilles Lafforgue. "Cost estimation of the use of low-carbon fuels in prospective scenarios for air transport". In: *AIAA SCITECH 2023 Forum*. American Institute of Aeronautics and Astronautics, 2023. DOI: 10.2514/6.2023-2328.

[10] Antoine Salgas, Thomas Planès, Scott Delbecq, Gilles Lafforgue, and Joël Jézégou. "Modelling and simulation of new regulatory measures in prospective scenarios for air transport". In: Lausanne, July 2023. DOI: 10.13009/EUCASS2023-593.

[11] Volker Grewe et al. "Evaluating the climate impact of aviation emission scenarios towards the Paris agreement including COVID-19 effects". In: *Nature Communications* 12.1 (June 2021), p. 3841. DOI: 10.1038/s41467-021-24091-y.

[12] Thomas Planès, Scott Delbecq, Valérie Pommier-Budinger, and Emmanuel Bénard. "Simulation and evaluation of sustainable climate trajectories for aviation". In: *Journal of Environmental Management* 295 (Oct. 2021), p. 113079. DOI: 10.1016/j.jenvman.2021.113079.

[13] Milan Klöwer, MR Allen, DS Lee, SR Proud, Leo Gallagher, and Agnieszka Skowron. "Quantifying aviation's contribution to global warming". In: *Environmental Research Letters* 16.10 (Oct. 2021), p. 104027. DOI: 10.1088/1748-9326/ac286e.

[14] Lynnette Dray, Andreas W Schäfer, Carla Grobler, Christoph Falter, Florian Allroggen, Marc EJ Stettler, and Steven RH Barrett. "Cost and emissions pathways towards net-zero climate impacts in aviation". In: *Nature Climate Change* 12.10 (2022), pp. 956–962. DOI: 10.1038/s41558-022-01485-4.

[15] Candelaria Bergero, Greer Gosnell, Dolf Gielen, Seungwoo Kang, Morgan Bazilian, and Steven J Davis. "Pathways to net-zero emissions from aviation". In: *Nature Sustainability* 6.4 (2023), pp. 404–414. DOI: 10.1038/s41893-022-01046-9.

[16] Romain Sacchi, Viola Becattini, Paolo Gabrielli, Brian Cox, Alois Dirnaichner, Christian Bauer, and Marco Mazzotti. "How to make climate-neutral aviation fly". In: *Nature Communications* 14.1 (2023), p. 3989. DOI: 10.1038/s41467-023-39749-y.

[17] Brandon Graver, Dan Rutherford, and Sola Zheng. *CO2 emissions from commercial aviation: 2013, 2018, and 2019 | International Council on Clean Transportation*. Tech. rep. 2020. URL: https://theicct.org/publications/co2-emissions-commercial-aviation-2020 (visited on 01/12/2022).

[18] International Energy Agency. *Energy Statistics Data Browser*. 2023. URL: https://www.iea.org/data-and-statistics/data-tools/energy-statistics-data-browser (visited on 08/14/2023).

[19] Thomas Planès, Scott Delbecq, and Antoine Salgas. "AeroMAPS: a framework for performing multidisciplinary assessment of prospective scenarios for air transport". In: *Preprint submitted to Journal of Open Aviation Science* (2023).

[20] International Air Transport Association. *Air Passenger Market Analysis*. Tech. rep. Aug. 2023.

[21] United States Bureau Of Transportation Statistics. *Bureau of Transportation Statistics: T-100 segment database*. 2022. URL: https://transtats.bts.gov/Fields.asp?gnoyr_VQ=FMG (visited on 05/25/2022).

[22] Matthias Schäfer, Martin Strohmeier, Vincent Lenders, Ivan Martinovic, and Matthias Wilhelm. "Bringing up OpenSky: A large-scale ADS-B sensor network for research". In: Apr. 2014, pp. 83–94. DOI: 10.1109/IPSN.2014.6846743.

[23] Eurocontrol. *Aviation Data for Research*. 2023. URL: https://www.eurocontrol.int/dashboard/rnd-data-archive (visited on 05/25/2022).

[24] The World Bank. *Global Airports Flows*. Jan. 2023. URL: https://datacatalog.worldbank.org/search/dataset/0038117/Global-Airports (visited on 08/14/2023).

[25] Agência Nacional de Aviação Civil (ANAC). *Dados Estatísticos*. 2023. URL: https://www.gov.br/anac/pt-br/assuntos/dados-e-estatisticas/dados-estatisticos (visited on 08/14/2023).

[26] Australian Government. *Australian Domestic Airline Activity—time series*. July 2023. URL: https://www.bitre.gov.au/publications/ongoing/domestic_airline_activity-time_series (visited on 08/14/2023).

[27] United nations Trade Statistics. *UN Comtrade - Trade Data*. URL: https://comtradeplus.un.org/TradeFlow (visited on 08/14/2023).

[28] Wikipedia (community). *List of island countries*. Aug. 2023. URL: https://en.wikipedia.org/w/index.php?title=List_of_island_countries&oldid=1168790788 (visited on 08/14/2023).

[29] The World Bank. *World Bank Open Data Portal*. 2023. URL: https://data.worldbank.org (visited on 08/14/2023).

[30] Kontur. *Global Population Density - Humanitarian Data Exchange*. 2023. URL: https://data.humdata.org/dataset/kontur-population-dataset (visited on 08/14/2023).

[31] United United Nations Devlopment Programm. *Inequality-adjusted Human Development Index*. 2023. URL: https://hdr.undp.org/inequality-adjusted-human-development-index (visited on 08/14/2023).

[32] Wikipedia (community). *Wikipedia:WikiProject Airports*. Aug. 2009. URL: https://en.wikipedia.org/w/index.php?title=Wikipedia:WikiProject_Airports/page_content&oldid=308204634 (visited on 10/10/2023).

[33] Kyle Seymour, Maximilian Held, Gil Georges, and Konstantinos Boulouchos. "Fuel Estimation in Air Transportation: Modeling global fuel consumption for commercial aviation". In: *Transportation Research Part D: Transport and Environment* 88 (Nov. 2020), p. 102528. DOI: 10.1016/j.trd.2020.102528.

[34] Wikipedia (community). *Wikipedia:WikiProject Aviation/Airline destination lists*. Jan. 2015. URL: https://en.wikipedia.org/w/index.php?title=Wikipedia:WikiProject_Aviation/Airline_destination_lists&oldid=643250904 (visited on 10/11/2023).

[35] OurAirports. *Airport dataset*. 2023. URL: https://ourairports.com/data/ (visited on 10/11/2023).

[36] Stefan Gössling and Andreas Humpe. "The global scale, distribution and growth of aviation: Implications for climate change". In: *Global Environmental Change* 65 (Nov. 2020), p. 102194. DOI: 10.1016/j.gloenvcha.2020.102194.

[37] Leo Breiman. "Random Forests". In: *Machine Learning* 45.1 (Oct. 2001), pp. 5–32. ISSN: 1573-0565. DOI: 10.1023/A:1010933404324. URL: https://doi.org/10.1023/A:1010933404324.

[38] Tianqi Chen and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: ACM, 2016, pp. 785–794. DOI: 10.1145/2939672.2939785.

[39] Junzi Sun, Jacco M Hoekstra, and Joost Ellerbroek. "OpenAP: An open-source aircraft performance model for air transportation studies and simulations". In: *Aerospace* 7.8 (2020), p. 104. DOI: 10.3390/aerospace7080104.

[40] International Air Transport Association. *Airline Industry Economic Performance - June 2022*. Tech. rep. 2022. URL: https://www.iata.org/en/iata-repository/publications/economic-reports/airline-industry-economic-performance---june-2022---data-tables/ (visited on 06/30/2022).

[41] India Ministry of Civil Aviation. *Indian Flight Schedules*. URL: https://www.kaggle.com/datasets/nikhilkhetan/indian-flight-schedules (visited on 10/10/2023).