Compilation and Applications of an Open-Source Dataset on Global Air Traffic Flows and Carbon Emissions

Salgas, Antoine; Sun, Junzi; Delbecq, Scott; Planès, Thomas; Lafforgue, Gilles

DOI
10.59490/joas.2024.7365

Publication date
2024

Document Version
Final published version

Published in
Journal of Open Aviation Science

Journal of Open Aviation Science (2024), Vol.**2**
doi:10.59490/joas.2024.7365

JOAS

DATA

Compilation and Applications of an Open-Source Dataset on Global Air Traffic Flows and Carbon Emissions

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

[1]Fédération ENAC ISAE-SUPAERO ONERA, Université de Toulouse, France
[2]Faculty of Aerospace Engineering, Delft University of Technology, the Netherlands
[3]TBS Business School, France

*Corresponding author: antoine.salgas@isae-supaero.fr

(Received: 19 January 2024; Revised: 21 May 2024; Accepted: 1 July 2024; Published: 9 July 2024)
(Editor: Michael Schultz; Reviewers: Luis Delgado, Tamara Pejovic, Duc-Thinh Pham)

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 study bridges this gap by presenting an innovative open-source dataset detailing 2019's global air traffic flows and associated CO_2 emissions. A comprehensive approach that compiles diverse flight data sources is presented. The remaining data gaps are addressed by constructing a route network through systematic Wikipedia parsing and by estimating the related traffic using socio-economic data. Then, an aircraft performance model to estimate CO_2 emissions is implemented. This methodology promises reinforced reproducibility and broader data accessibility in aviation environmental research. Several reference datasets are used to evaluate the accuracy of the open-source dataset. Despite various levels of accuracy for individual routes, major traffic flows are well estimated at the country and continental levels, albeit with room for refinement to ensure consistent data reliability. To facilitate the exploration of the dataset, the AeroSCOPE tool has been developed. To initiate research prospects, use cases of this dataset are proposed, concerning the network potential of electric and hydrogen-powered aircraft and inequalities in air transport.

Keywords: Open Data; Air Traffic; Aviation Emissions; Scientific Software

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 such as 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 approximately 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, both from 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 achieved by both industrial stakeholders [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 coverage 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. To do so, different open-source datasets with limited geographical coverage are compiled based on their extent and characteristics. Since their combined scope remains incomplete, an innovative method based on the parsing of Wikipedia airport pages is used to build a global route network. Their related traffic is estimated by training regression models using statistical and socio-economic data. This is then used to complete the various open-source datasets compiled in the first place. The developed dataset is used for performing several applications in terms of air transport decarbonisation and inequalities. This paper is also part of the development of AeroMAPS, a dedicated open-source prospective scenario simulator [19], because it especially aims at developing its regionalised assessment capacity in the future. Note that this paper is an extended version of a conference paper [20] and completes it with Section 4. The idea behind this extended version is to provide a basis for reflection on the potential applications of an open-source air traffic dataset, justifying the work undertaken in the core of this article, which nevertheless remains the compilation of the dataset.

To this end, the paper is organised as follows. 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 in Section 3. Several applications are provided in Section 4, also introducing AeroSCOPE, a dedicated tool for exploring the dataset. Finally, concluding remarks and perspectives are given in Section 5.

2. Method

2.1 Data collection and aggregation methodology

The methodology for compiling an open-source database is comprehensively detailed in this section, with 2019 selected as the baseline year due to the significant disruptions in air traffic patterns caused by the COVID-19 pandemic in subsequent years [21]. 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.

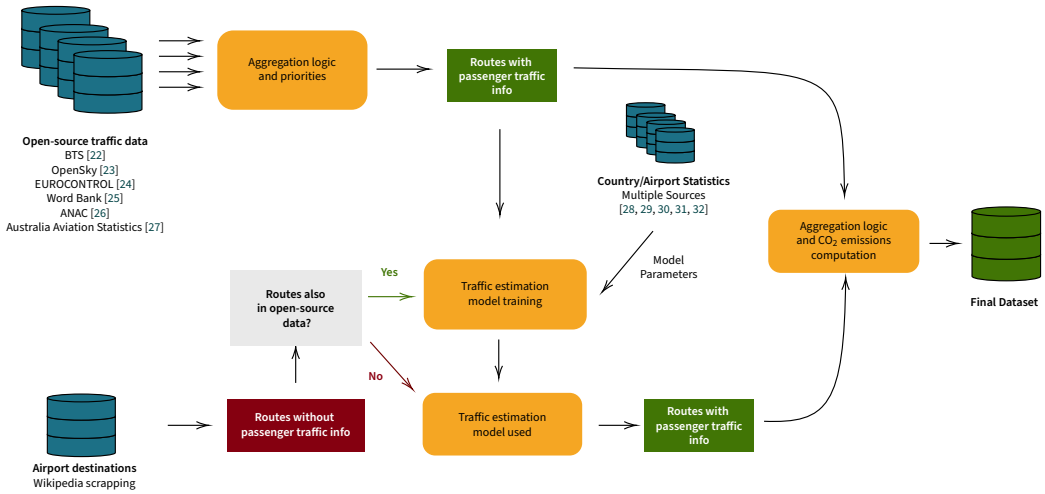

Figure 1. Traffic dataset creation flowchart.

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 [33]. 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. 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 [34] 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 scope and 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 [22] 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 [25], 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 [26] and Australian [27] 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 [23, 24], 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.	✓	✓	✓	✓	[22]
OpenSky	Global (partial)	Radar	✓	✓	✓	-	[23]
Eurocontrol	To/from EU	Radar	✓	✓	✓	-	[24]
World Bank	International	Adm.	✓	-	-	✓	[25]
ANAC	Brasil	Adm.	✓	-	✓	✓	[26]
AUS Stats	Australia	Adm.	✓	-	-	✓	[27]

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 [35]. 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 Section).

The choice to use Wikipedia as a source for filling data gaps is predicated on the comprehensiveness

[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

of its airport destination listings. However, we acknowledge the dynamic nature of Wikipedia pages, which are constantly updated by the community, given Wikipedia's open editing model. Different issues arise from this. First, 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. The network could therefore differ from the actual network of 2019, even more given that the COVID-19 crisis happened in between. Another important point to note when working with Wikipedia is that, despite the community's proofreading efforts, errors may persist in some airports. Some authors could use low-quality sources to redact the airport page, and there is no systemic way to check this in the parsing process. To mitigate potential inaccuracies resulting from this, a database validation process is given in Section 3 using a reference commercial dataset. A last issue are airports that would not be listed in the original set of airports [35] that is parsed to collect the associated destinations. To address this problem, any airport identified during the destination parsing process that is not already in the original set of URLs is subsequently added to this set. Therefore, the only way an airport can be missed is if it is neither listed among the Wikipedia airport pages nor served by any flight originating from an airport in the original list. These limitations should be put in the context of this work: route level accuracy is not the most important requirement, as long as large countries or regional flows are well estimated.

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 airport are also collected on the Wikipedia pages of the airport. A Wikidata item is also linked to each airport page[4]. Wikidata is a structured database made to provide data storage for Wikipedia among others. Thus, similar data (for example, annual passenger traffic, ICAO and IATA codes) can be found on Wikipedia and Wikidata. However, an advantage of Wikidata is that the fields are dated, compared to Wikipedia "last-year available" information. Not every airport has all the features and the airport list is filtered to retain only airports with an IATA code, thus offering commercial services. Airport geographical coordinates, countries and continental codes are added by merging an airport database [36] 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 [31] 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 assessed.

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 [37], 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 [30]. The inequality structure is captured by the Gini coefficient of incomes [30] and the Inequality-adjusted Human Development Index (IHDI) [32]. 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 [30]. Their surface and their insularity are also used because one can think that the size of the country affects the number of domestic flights [29, 30].

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.

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

This final group of features deals with bilateral relations between airports on the different routes. These include the bilateral trade flows [28], the number of airlines serving each route (established during the Wikipedia list parsing) and the great circle distance between the two airports.

To summarise the feature collection, the available data for each route is therefore:

- Number of airlines on the route (partitioned on Regular/Charter/Seasonnal)
- Great circle distance of the route
- Trade value between the two countries
- Domestic/international status of the route
- Annual number of passengers of both airports
- Population living within 30, 70 and 150 km of both airports
- Number of concurrent airports (and their traffic) within 30, 70 and 150 km of both airports
- Gini index and human development index of both countries
- Gross domestic product per power purchase parity of both countries
- Number of incoming and outbound tourists of both countries
- Share of tourism in exports of both countries
- Surface and insularity of both countries
- Airport data (airport and country codes, geographical coordinates)

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 their 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 into a number of seats offered per route with average aircraft capacities (as explained in Section 2.2), which reduces the data accuracy. Moreover, in the case of [23], 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 the 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.

The priority order retained is thus administrative data from the BTS [22], completed by data from the Word Bank [25], and from Brasilian [26] and Australian [27] authorities. Radar data from Eurocontrol [24] are added and finally those from the OpenSky network [23]. This is achieved because only the dependent variable is of interest at this point and not the other details provided by the dataset.

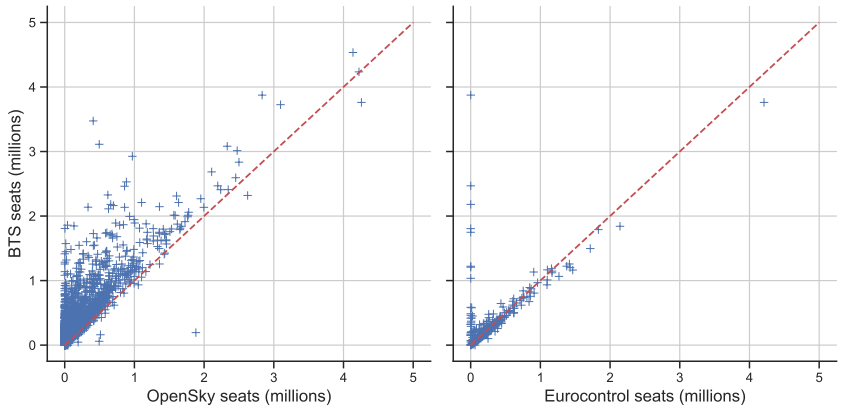

Figure 2. Number of seats per route: radar data [23, 24] comparison with BTS [22]. Contains only data in both datasets.

Therefore the impact of missing capacity in OpenSky data mentioned above is limited: adding this data last in the priority order only increases the number of traffic-determined routes by 2%. In these 2%, there is still a risk of mistakenly characterising traffic on a road as known, when in fact it is only partially covered by radar data and should be estimated by the model rather than used to train it.. To limit this, only flights with a capacity of over 1,000 seats per year are kept. This threshold has been chosen arbitrarily.

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 remaining 59% traffic-determined routes. They are divided into a typical 80%-20% train-test data split. It consists of separating the data into two groups; the first one to train the regressor and the second to evaluate its performance on unseen, but known data.

Methods for traffic forecasting in the literature are based on time series or causal regressions. Only the second case is applicable here as the problem is to determine current unknown flows, for which no past data is available either. The most frequent causal method is to use gravity models [38], but much more advanced methods based on neural networks are also being developed [39]. Different regression methods are tested in this section, but no external models from the literature are used, the intention being to be able to base the work on the open-source features collected internally in the previous section.

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. Using a linear regression requires a few processing steps to the data. Features need to be numerical, therefore information such as the origin/destination country/continent is dropped. An alternative to this would be to encode them as binary variables but it would greatly increase the number of features. Data entries with missing features are also dropped (or imputed). The results are inadequate despite an acceptable R^2 (which quantifies how well the variance of the dependant variable is explained by the model) 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 regression) and manual feature selection 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. All the features but those mentioned here are dropped to this end. This method is also insufficient as illustrated by the *Log-Linear* graph of Figure 3, where all high-demand routes are severely underestimated. The log regression gives more weight to low-traffic training points. 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. Alternatively, training a gravity model on a homogeneous dataset (based on a single source and routes of the same kind) could be another way of improving its performance.

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

If the two previous approaches are not able to capture enough of the data variability, they could be relevant to predict country-level or industry average flows, as they are completely interpretable. They could also be more suited to regional analyses with more similar flows, or complemented by other features such as the ticket price for instance.

Instead, more sophisticated machine-learning methods based on regression trees are tested. Regression trees are a machine-learning technique used to estimate a continuous dependent variable by making successive decisions based on predictor variables at each node of the tree, ultimately leading to predictions at the leaves. To reduce their tendency to overfit and increase robustness, multiple regression trees can be combined, and their predictions can be averaged through an ensemble learning approach. Random forests [40] are an example of combining multiple regression trees while sampling subsets of the features and the data in a bootstrapping process. Therefore, it combines several weak random regression trees to produce a reliable estimator of the dependent variable less prone to overfitting. The random forest regressor does not handle missing values (NaN), like the previous linear regression. Therefore, data entries with missing features must be removed from the dataset, or the value of the missing feature can be imputed arbitrarily. It can be set to zero, or the mean value of the feature for instance. Here they are set to the value of the 1000-quantile (i.e. 0.1% bottom of all the values of the feature), following the idea that missing features are more likely to be found on small airports. Training random forest is long (around 20 minutes with 1000 trees per forest), which complicates the tuning of the hyperparameters of the model. Nevertheless, the R^2 is improved to values around 0.7 (depending on the train-test split). Compared to the linear regression, there are no more negative estimates, as the tree is trained on only positive dependant variables.

Other regression algorithms also handle missing values, and they could be used to avoid this intervention on the dataset. It is the case of XGBoost [41], a tree-based gradient-boosted regression method. XGBoost operates by iteratively building an ensemble of decision trees, each tree correcting the errors of the previous ones following the gradient of the loss function. XGBoost provides much faster 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.

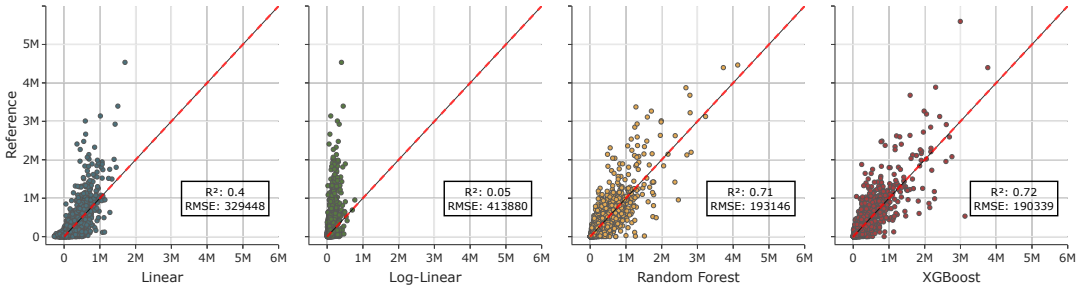

Figure 3. Seat capacity estimation: regression techniques tested.

Due to its ease of use and to avoid missing feature imputation or deletion, 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 Figure 18 in Appendix 1. The most important feature is the distance of the relation. Several explanations can be given: longer flights are more expensive, and require more travel time and time available at a longer destination. It is logical that longer flights are a priori associated with less traffic; conversely, very short flights can also have low traffic because they serve smaller cities and that there are alternatives. Decision trees can be used to represent these two assumptions at the same time. The second most important (pair of) feature is the total traffic of both arrival and departure airport, with an arithmetical link between the size of an airport and the traffic flow on its routes. The number of regular airlines on the route is also one of the most important feature, as the market on large routes is sufficient for several competitors. The number of seasonal or charter competitors is way less important. Population around the airports is the next pair of most important features. However, 3 different radius are considered, which might diminish the importance of each of these features. The traffic in competing airports comes next, and is more important than the number of competitors itself. Excepting the bilateral trade value, all the socio-economic features (GDP, Gini, IHDI,...) are used way less by the model. 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. An interesting prospect that was not investigated would be to look at the product of origin-destination features (product of origin-destination airport traffic for instance).

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 (and BTS) 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 when possible. 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. Only one source is used per route to ensure there is no double counting. From a practical point of view, the aggregation logic consists therefore starting with the BTS [22], on which OpenSky [23] and then Eurocontrol [24] extra routes are concatenated. It is then completed by the three other administrative sources, from the Brazilian authorities [26], the World Bank [25] and the Australian government [27].

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.

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

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

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

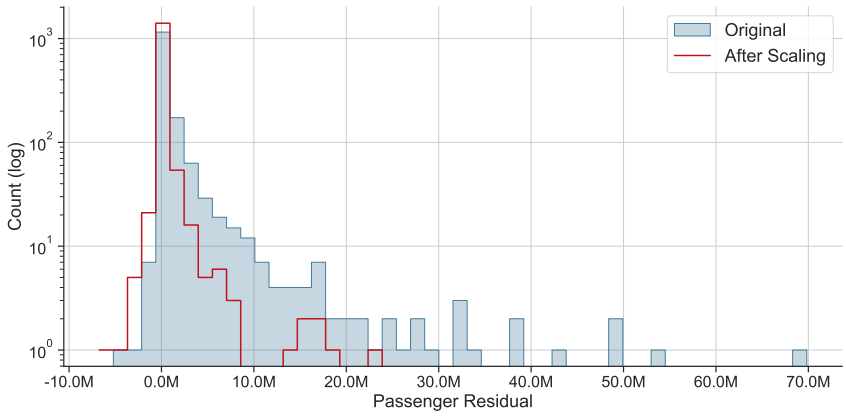

Figure 4. Estimated flights scaling effect on airport passenger traffic residual.

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. An important point about the process of aggregating data is that, because of the inherent differences in their nature, there are differences in quality between sources. For example, some sources use an average number of seats per aircraft to estimate the number of seats available on each route, while others provide this information directly.

This is even more the case for routes where capacity is estimated; accuracy at route level is relative, as will be discussed in more detail in Section 3. This limitation must be taken into account in any subsequent analysis.

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 sections 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 [42] 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 [34], is used. It requires only the knowledge of the aircraft type and the distance of the flight to estimate the fuel burn. It builds, for 133 different aircraft types, quadratic regressions on flights simulated with a higher order model based on Eurocontrol's BADA. The goodness of fit of the regression is very high with an average coefficient of determination of 0.997. When evaluated on other missions than those used to fit the models, the average approximation error is of -0.09% (95% confidence interval: -4.08 to 2.70%), with short flights consumption being underestimated. The authors report that the error reduces once several flights are aggregated, making the model suitable for global or aggregated scale fuel consumption estimation. However, when compared to real-world fuel consumption, such as those reported by United-States-based airlines, the aggregated fuel burn consumption is underestimated by 4.8%, because of simplified operational assumptions in the high-order model. FEAT was used for its ease of use, but there are other references in the literature that could be used, such as a surrogate model of OpenAP [42] or analytical models from [43].

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 [22, 23, 24]. 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 to 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 [22] 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).

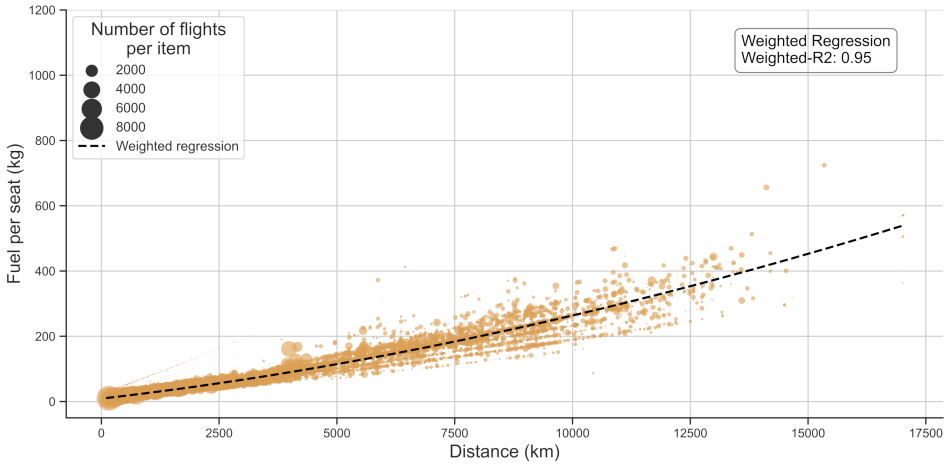

Figure 5. Fuel burn regression on FEAT-computed data points.

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

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 Section) 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 [37], 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 [44], there are 5570 million seats available in the open-source dataset of this work. It means a difference of around +1% compared to IATA.

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% [44]. It's a relatively rough method that ignores geographic disparities, but the differences in ASK growth were relatively limited from 2018 to 2019 [44]. 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 per country of departure 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.

It is interesting to draw a parallel between the share of data that is estimated (Figure 7) and the quality of the estimate (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 [22]; and errors of less than 3% in the majority of Western European countries, where [24] data is abundant. Australia and Brazil, for which specific datasets are used, are also very well estimated. As it can be seen in Figure 7, the share of estimated data is very low for all of these regions. In the rest of the world, the error reaches reasonable levels in major countries such as China, with a +6% error, and India, with a -9% error, despite a large share of estimated data (62% and 41% respectively). Similar comments can be made for Mexico or Iran for instance. The most disappointing major market is Indonesia, with 24% of the traffic missing, for the same reasons as described before, at the airport level. The traffic is not well captured for some African and Middle Eastern countries. In the case of Yemen, there is a direct link with the share of estimation as a data

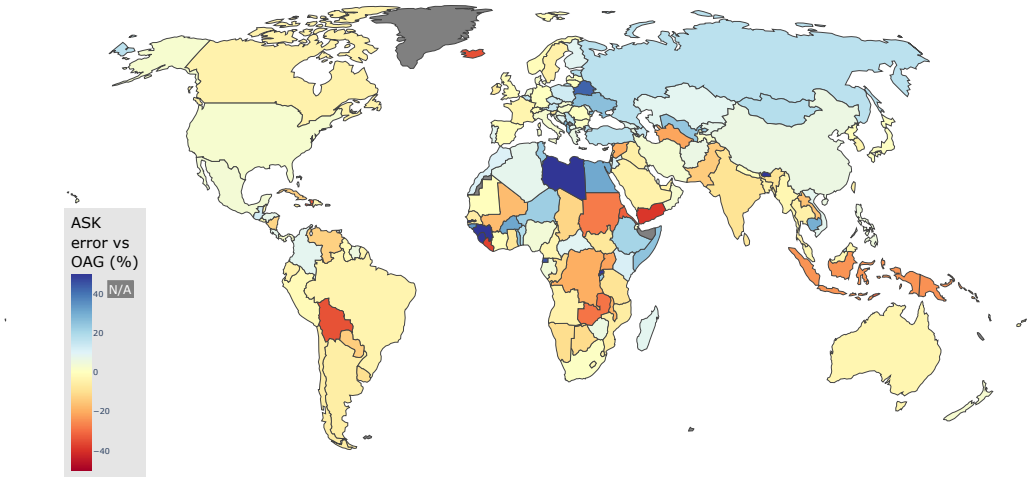

Figure 6. Relative ASK error per country of departure compared to reference data from OAG.

source, but that can not be said with Chad or Sudan where most of the data comes from the World Bank dataset. The supporting information around it is limited, specifically about the maintenance of the dataset. Further research is needed to determine whether this is an error in the World Bank dataset or whether the reference is incorrect.

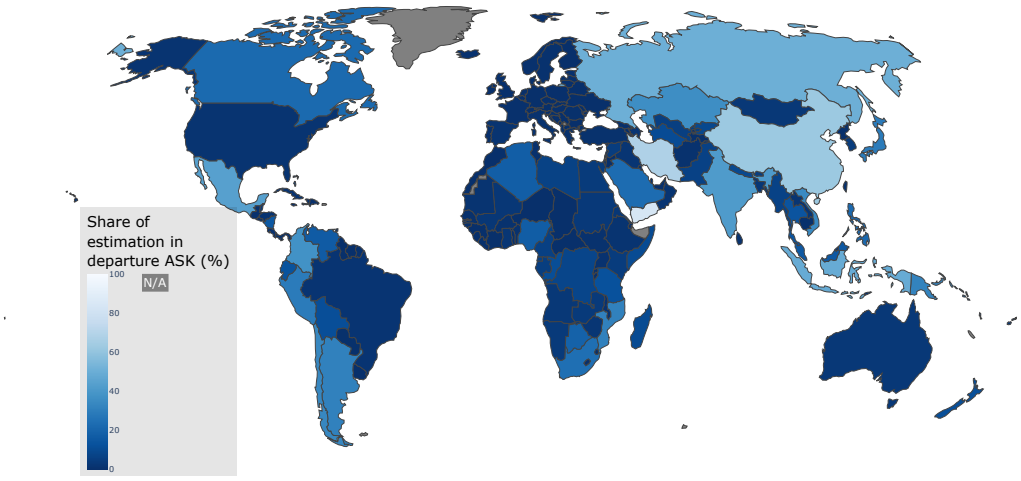

Figure 7. Share of estimation as a source in total departing ASK.

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 [23, 25] do not, resulting in this large error.

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

Bn ASK (error)	**Africa**	**Asia**	**Europe**	**N.America**	**Oceania**	**S.America**
Africa	99 (-8.0%)					
Asia	165 (5.4%)	2,826 (-1.2%)				
Europe	402 (10.9%)	1,215 (7.4%)	1,239 (0.8%)			
N.America	34 (44.7%)	628 (3.7%)	938 (3.6%)	1,811 (1.2%)		
Oceania	4 (-5.5%)	352 (-6.4%)	14 (627.2%)	179 (12.3%)	148 (-5.2%)	
S.America	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 indicate a trend for the surrogate fuel-burn model to be optimistic on shorter routes.

To sum up, the dataset is suitable for the use case for which it is intended, i.e. it can provide an unbiased estimate of traffic at fairly high levels of aggregation (continental or by country, in some cases). It is advisable to include a sensitivity study in the potential uses of the dataset, using as uncertainties the error levels reported in Tables 3 or 4 or Figure 6, for example. However, it cannot

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*

be used without reliability checks in areas that are poorly covered by existing datasets, or if the application requires accuracy at road level. This could be the case for network optimisation.

4. Applications and use cases

In this section, a dedicated tool for exploring the dataset is introduced. Moreover, several applications relying on the dataset are performed concerning air transport decarbonisation and inequalities. This section highlights its potential to give insights into air transport decarbonisation while suggesting areas for further research. The analyses presented in this section were performed in an interactive notebook, which is referenced in the Reproducibility section.

4.1 AeroSCOPE: a dedicated tool for exploring the dataset

In order to provide user-friendly access for exploring the open-source dataset, the AeroSCOPE tool has been developed and is available as an open-access and open-source web application (see Reproducibility Section). It does not aim at providing a comprehensive data analysis module: for more complex or custom purposes, using a Jupyter Notebook instead is recommended. The user interface is developed using Seaborn and Plotly graph libraries and the interaction is handled using ipyvuetify, a Jupyter Widgets implementation of Vuetify UI components. A screenshot of the user interface is given in Figure 19 in Appendix 1.

Three aggregation options are offered: continental, country, and more detailled levels. At each of these levels, it is possible to interactively filter the dataset to focus on a specific region, route, or airline. Note that international flight emissions were accounted for using a departure country takes-all convention. This is not an issue since all routes were made non-directional during the estimation process, before being symmetrically directionalised at the end of the process (i.e. there are as many passengers from B to A as from A to B). Three metrics can be observed: CO_2 emissions, ASK and seats.

An example of a continental-level analysis is given in Figure 8. This figure is a multi-level treemap, in which the top level is the world, the second level is the origin continent, and the third level is the destination continent. Each area is proportional to the CO_2 emissions of the corresponding origin-destination flow. The remarks mentioned during the validation process still hold: the values at the continental level are reliable, but more caution is required at sub-levels. That is particularly true for countries outside Europe with a significant fraction of domestic flights, which may lead

to inaccuracy at the route and country levels. The data is presented in a disaggregated manner at the route level but the estimation, and fuel burn models are not meant to compare for instance the performance of two aircraft types on a given route. Similarly, comparing airlines operating on regions covered differently could lead to incorrect conclusions. All these modes are instead intended to explore trends, such as flight distance distribution.

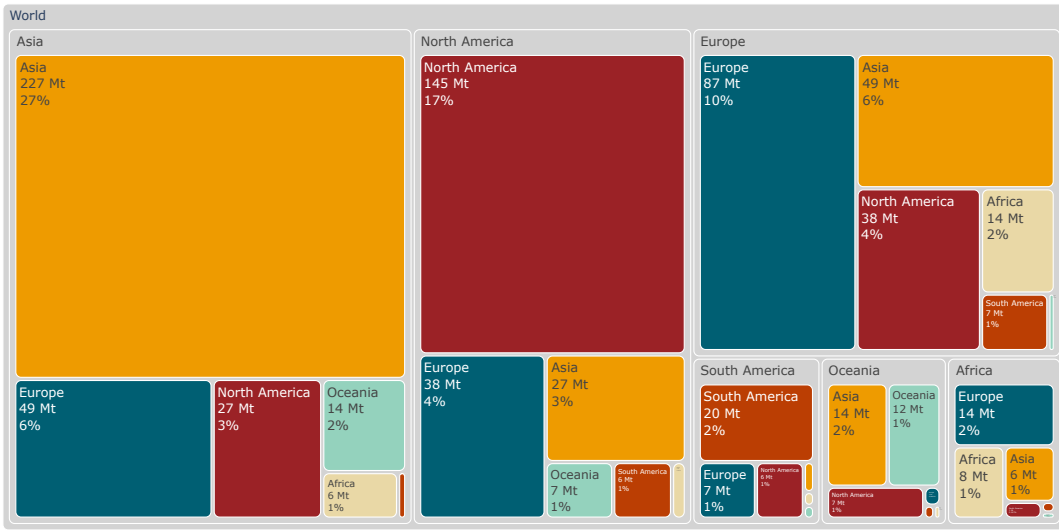

Figure 8. Continental flows treemap for CO_2 emissions.

4.2 Network potential of electric and hydrogen-powered aircraft

A first application is the evaluation of future electric and hydrogen aircraft in terms of traffic and emission coverage. The aggregation level of the dataset allows for isolating the routes on which a specific aircraft would be able to fly. Such an evaluation capability is relevant in the frame of a transition towards battery-electric and hydrogen-powered aircraft (using either fuel cells or hydrogen gas turbines), whose performance will differ from that of conventional aircraft.

In the case of battery-electric flights, the low energy density of the batteries confines these aircraft to very short range missions: under the optimistic [8] hypothesis of 800 Wh/kg batteries, [45] evaluates the potential of such an aircraft to 500 NM missions. In the case of hydrogen-powered aircraft, despite a very interesting energy density, hydrogen occupies much more volume than kerosene, besides requiring cryogenic storage. It led aircraft manufacturers to target a short-to-medium range application for this technology. For instance, [46] declares studying both 1000-NM and 2000-NM concepts.

This dataset allows the evaluation of the relevance of such aircraft on the 2019 network. Note that their life cycle CO_2 emissions are not considered in this paper, and the following results should only be evaluated in terms of market potential. Doing a full life cycle assessment seems to be a prerequisite for further impact assessment, especially for hydrogen and electric aircraft. The country in which they operate is also important, as their total emissions will depend largely on the electricity emission factor. Such prospective scenarios can be simulated and evaluated using [19] for instance.

The cumulative sums for the three metrics available in AeroSCOPE as a function of flight distance are given in Figure 9. Vertical lines represent the maximal range of the three aircraft aforementioned. The figure shows that the cumulative sum of available seats grows faster than either ASK or CO_2.

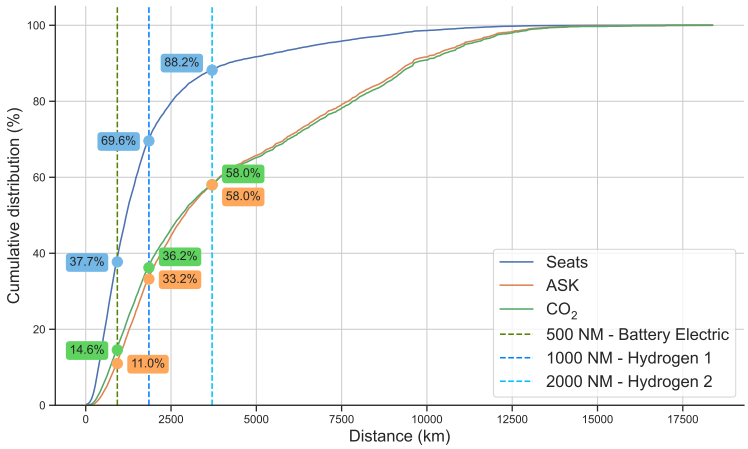

Figure 9. Cumulative distribution functions according to the flight distance on the total dataset.

This is intuitive, as shorter flights are more frequent and carry many passengers, generating less ASK and CO_2 than longer flights. As a consequence, one can see that most seats are offered on short haul routes: 69.6% of them are below 1852 km (1000 NM), and 37.7% on very short haul routes are below 926 km (500 NM). That is an interesting prospect for the deployment of those alternative aircraft architectures, but this will not contribute to the same extent to the decarbonisation of air transport.

The battery-electric aircraft would address a market representing 14.6% of the air transport emissions, while a conservative 1000-NM hydrogen aircraft would address up 36.2% of the emissions. The more ambitious 2000-NM hydrogen aircraft would address up to 88.2% of the seats and 58.4% of the emissions. It is worth noting that this analysis is applied to the 2019 network, and there is no guarantee that the distributions will remain constant over the years. In particular, the relative weight of the different regions is likely to change [3]. Moreover, one can also imagine a network reconfiguration that would foster more direct routes, or on the contrary, more centralised networks.

Since local governments can support one or another decarbonisation option (by subsidising or regulating for instance), it is relevant to look at the decarbonisation potentials of these aircraft on a national scale as well. Large differences are observed between countries as it can be seen on Figure 10, on which the cumulative distribution of CO_2 emissions is represented for the world, and for Norway and the Netherlands.

Norway is a long country, on which land transportation is made difficult by numerous fjords and mountains. Therefore, short-haul flights are over-represented compared to the previous global figure. Flights potentially operated by the fictional battery-electric aircraft represent 39.2% of their total air transport CO_2 emissions (on an all-departures convention), while flights below 2000-NM represent 81.8% of the same figure. It means that operational considerations set apart, the Norwegian government could heavily rely on such solutions. One can note that the very low emission factor of electricity in Norway (due to large hydroelectricity availability) makes the solution a priori attractive, before a full assessment.

In contrast, the Netherlands are a country almost 10 times smaller and more compact, where domestic operations are almost non-existent. The main Dutch airport (Amsterdam-Schipol) is a major hub for KLM Airlines, and 82M seats were offered to and from the airport in 2019. Although a dense short-haul European network exists, which starts from the Netherlands, it is outweighed by the nu-

merous long-haul and carbon-intensive routes. Therefore, the market represented by an ambitious hydrogen-powered aircraft would cover only 30.8% of the CO_2 emissions. The electricity emission factor was also much worse in the country. These factors could lead to a short-term policy in support of other solutions, such as biomass-based drop-in SAF which nevertheless requires additional considerations on biomass availability. It is important not to over-interpret this comparison. The potential for deploying alternative aircraft in these two countries is indeed comparable in terms of volume covered; however, the Netherlands must also explore other solutions.

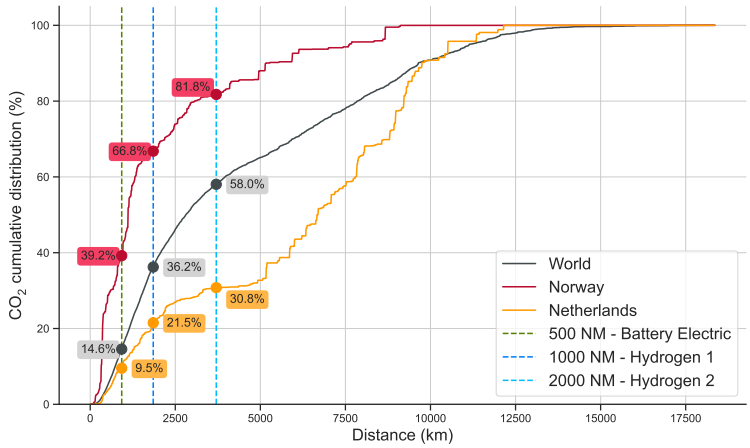

Figure 10. Cumulative CO_2 distribution functions for Norway, the Netherlands and the total dataset.

Given that the transition to another energy carrier requires infrastructural changes at airports, this case study can be extended by looking at which airports need to be equipped first to have a rapid impact on decarbonisation. This is a fairly complex optimisation problem to solve, given the interconnections in the network. As this is not the subject of this work, the question is approached with a very simplified approach in the case of a hypothetical 500-NM electric aircraft.

A first approach would be to rank the various airports in order of the gross CO_2 emissions represented by flights below 500 NM, and then to select the top n airports that one wants to equip with adequate charging infrastructure. Finally, the emissions covered by routes between two airports of this set are measured as a fraction of theoretical maximal potential where all airports would be equipped. A route is considered as valid if and only if it links two selected airports. As it can be seen on the left graph of Figure 11, the fraction covered increases sharply with the number of airports equipped. The routes between two members of the first $n = 500$ airports cover 69.6% of the total battery-electric aircraft potential (vertical dashed blue line).

Since there is no guarantee that this process is optimal (for instance if those airports were linked only to smaller airports not in the top n airports), another approach is tested to select $n = 500$ airports. To do so, the largest airports are still selected, but including their largest neighbours to ensure the electric aircraft will be able to operate from this hub on a small network. The selection stops when the total number of airports selected reaches $n = 500$. The effect of this approach is shown on the right graph of Figure 11. A simple case is plotted in orange where the i^{th} largest neighbours of each major airport are selected without criterion on their overall ranking. The first point on the left selects 0 neighbour for each airport and thus corresponds to the previous approach, which covers therefore again 69.6% of the potential. As neighbours are selected (from the left to the right of the graph), the orange line decreases progressively, with large potential airports being traded for smaller potential airports albeit neighbours of the largest. To mitigate this effect, restriction is made

on the overall rank of the considered neighbours. When only the top-500 airports are eligible for neighbour selection, it corresponds to the first approach, hence a constant potential no matter how many neighbours are considered for each major airport. Two intermediate cases are investigated, with 800 and 1000 largest airports being considered as eligible for neighbour selection. However, it only slows down the reduction of the potential covered. Thus, no "network effect" can be observed with this simplified approach.

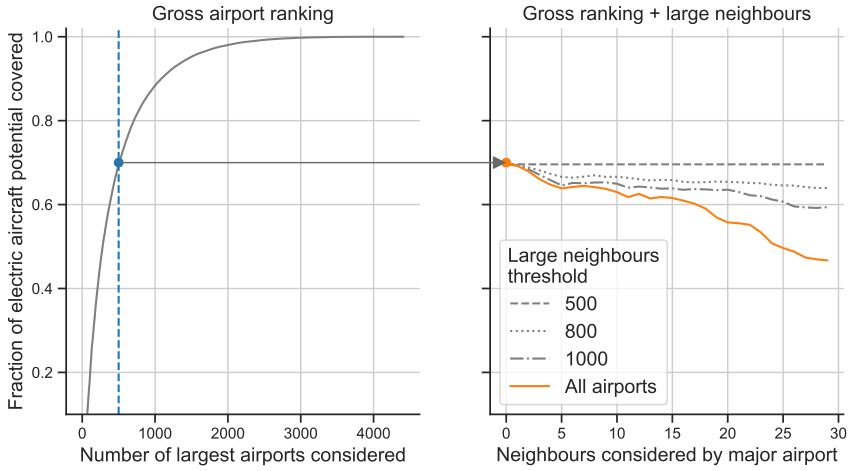

Figure 11. Selection of 500 airports for electric aircraft operations: decision on largest potential alone (left) and including largest neighbours (right).

This raises the need to consider more sophisticated methods to improve on the simplified solution of selecting the airports with the biggest gross potential. Optimisation based on random airport selection is complicated by the number of possible combinations. The problem is equivalent to the challenge of choosing 500 airports in the 4400 available in this electric-aircraft routes dataset. This yields around 10^{675} possibilities, making any brute force exploration impossible. By considering the network as a graph, it would be possible, for example, to consider the problem as the selection of a variable number of sub-graphs, under the constraint of respecting $n = 500$ nodes.

To better align with the objectives of this paper, a simpler method based on that presented in the previous paragraph is used. It consists of going through the list of the largest airports, selecting them or not, as well as a variable number of their neighbours until $n = 500$ airports are selected. The selection process is described by pseudo-code in Appendix 2. The output of the algorithm is maximised using a genetic algorithm [47]. It explores the possible combinations between the top airports selected and a specific number of neighbours for each airport. Note that this approach does not guarantee an optimal solution. It however succeeds in slightly improving the selection of $n = 500$ airports by increasing the share of electric aircraft potential covered from 69.6 (obtained with the first approach) to 70.6%. It can be seen in Figure 12 that the airports selected by the two approaches are mostly the same, with less than 10% different airports between the two sets. However, despite very close overall results, the optimisation seems to concentrate the network over the same areas, as expected given how the function is built, with each major airport being a "seed" for small networks. A lot of high-potential airports are thus removed from the dataset since they are geographically far from any top-end high-potential airport.

The difference in performance between naive selection and optimisation increases sharply as the size of the set of airports to be selected decreases. Figure 13 illustrates this effect. For a very small set of airports equipped, the gross ranking selection is not efficient: largest-potential airports are not

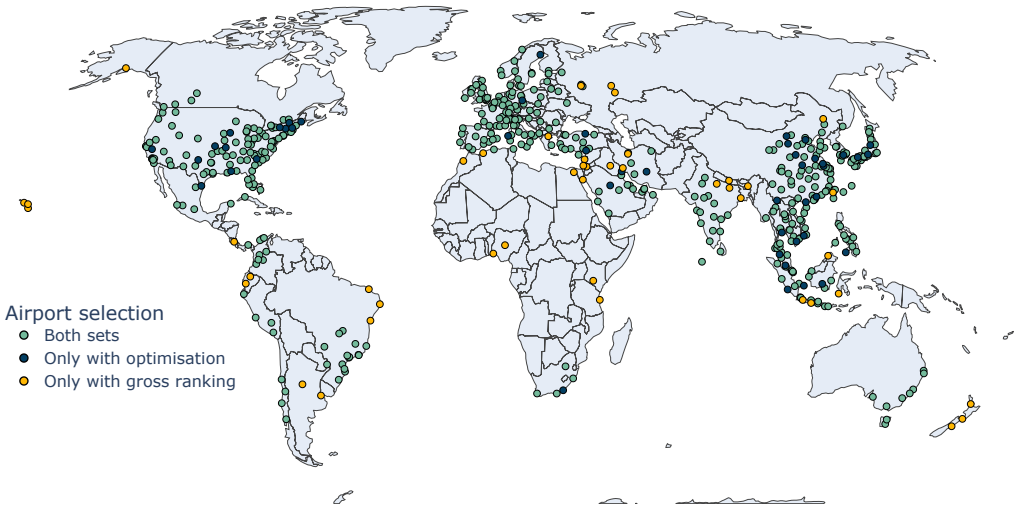

Figure 12. Comparison of naive and optimisation approaches: airports selected for $n = 500$ airports.

necessarily geographically close and are not likely to create a network of electric-aircraft compatible routes. Therefore, some will be isolated thus reducing the effective potential. The optimisation approach is much more robust to this phenomenon and will select some local networks. This effect is reduced as more airports are selected. For instance, optimisation approach improves the potential covered by 25 airports by more than 200%, but the potential covered by 1000 airports by 0.4% only. This trend could be reduced for longer-range aircraft, as the largest airports are more likely to be within flying distance of each other, but this was not investigated.

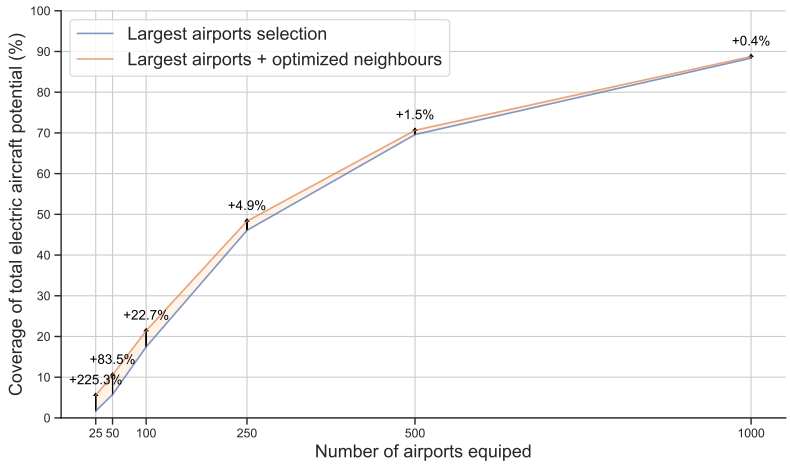

Figure 13. Comparison of naive and optimisation approaches: potential covered and relative improvement.

To summarise this section, the present dataset allows the evaluation of the potential of future aircraft architectures on the current network. This potential is certain, but unequal in different countries. Simple approaches can be used to select the airports in which to prioritise the infrastructure investment to accommodate such aircraft. Detailed graph-theory-based approaches should be pursued to

find a rigorous and optimal solution to the problem in question, but these approaches seem to be more relevant if the set of airports to be equipped is small; as naive gross ranking selection seems efficient for larger sets.

4.3 Country-level air transport inequalities

In this second application case, the aggregated dataset collected in this paper is used to evaluate the country-level inequalities in the use of air transport. The "country-level" refers to the fact that internal inequalities in the use of aircraft for each country are not addressed here by lack of data. An overview of passengers per country inhabitant is illustrated in Figure 14.

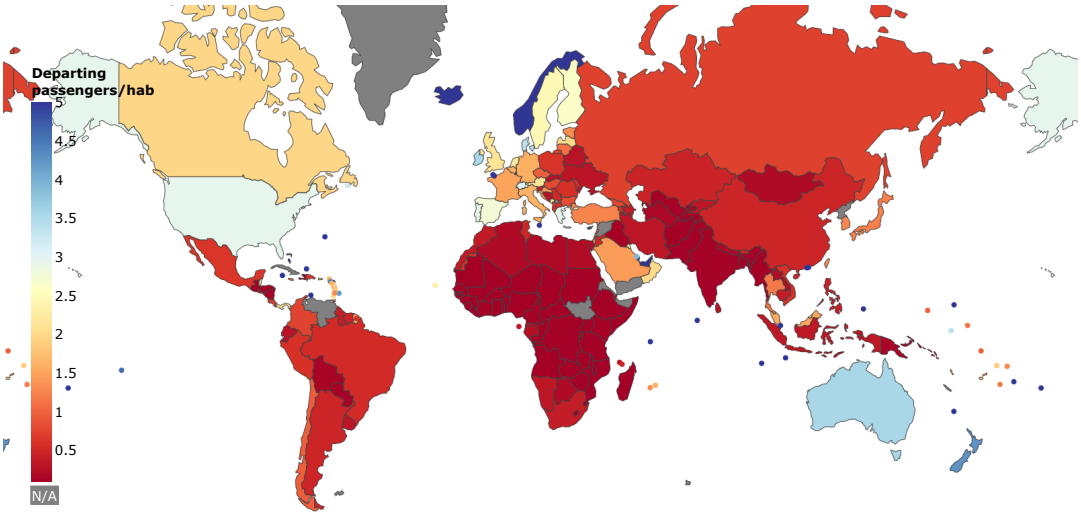

Figure 14. Departing passengers per country inhabitant. Note that the colour scale has been cut (see text).

This figure shows the number of passenger departures per inhabitant, country by country in 2019. The number of available seats was converted to a number of departures by using a load factor of 83% [44]. Note that this metric does not capture the origin of the travellers: it only represents the departures per inhabitant and encompasses trips by locals as well as tourists visiting the country. It is therefore an incomplete measure of the way in which populations use air transport. In view of the wide disparities in values (from 0.01 to 16 departures/inhabitant/year), the colour code has been truncated to 0.1/5. For all countries with more than 5 departures per inhabitant, the colour is always blue and red for all those below 0.1. The world average is at 0.6 departures per inhabitant. We can see that many islands or archipelagos are characterised by a very high value: 10.3 departure for Iceland, 7.3 for the Bahamas, or 7.8 for the Seychelles. It is the case also for Norway (6.1 departures) but also for city-states such as Singapore (6.2 departures), Hong Kong (4.9 departures) and Macao (7 departures). This is also the case for certain Gulf countries such as Qatar (8.2 departures) and the United Arab Emirates (7.3 departures). These two last categories are often characterised by the presence of major airline hubs (Singapore Airlines, Cathay Pacific, Qatar Airways, Emirates and Etihad, for example). This necessarily increases the metric calculated, as it takes into account connecting passengers.

To find out more about the disparities in aviation use, Figure 15 represents the Lorenz curves associated with ASK, seats, and CO_2 inequalities. This type of curve is used to associate the cumulative share of a total quantity (here the ASK, seats, or CO_2) with the corresponding cumulative share of the population. The equality case would see both quantities evolve at a similar rate and the curve

would be a diagonal, represented by the purple dashed line on the figure. As there is no information on the intra-country inequalities in the present dataset, the curves assume that each country is a homogeneous and strictly egalitarian cohort.

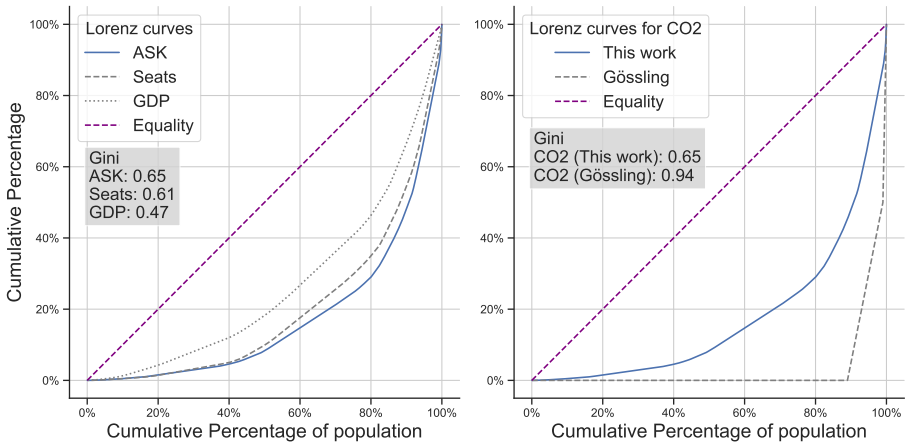

Figure 15. Country-scale aviation inequalities and various comparisons

The Gini coefficient of the distribution is computed to have an inequality metric. A Gini coefficient of 1 represents a perfectly unequal distribution, while 0 would be a perfectly equal one. It is defined by Eq. (7), in which the data is partitioned in n countries, X_k being the cumulative share of the population of the k first countries, and Y_k the corresponding cumulative share of the considered metric (here the ASK, seats, or CO_2).

$$G = 1 - \sum_{k=0}^{n-1} (X_{k+1} - X_k)\,(Y_{k+1} + Y_k) \tag{7}$$

The left graph of Figure 15 shows that the Gini coefficients for both metrics are very close (0.65 for ASK, 0.61 for seats). The curve was not represented for CO_2 as it was similar to ASK. In an attempt to compensate for the methodological shortcomings outlined above and to provide some insights for analysis, comparisons are proposed.

First, GDP inequalities are measured at the global level using the same methodology, which does not take into account inequalities within countries (see the left graph of Figure 15). It is then possible to compare aviation inequality to this more generic metric. The Gini coefficient is 0.47 for the latter, indicating that GDP is more evenly distributed between countries than the use of air transport.

Second, a comparison is performed based on the aviation use inequalities values given by [37]. Their main result is that 11% of the population flew in 2018, with 1% of the total population being responsible for 50% of the emissions. It makes it possible to plot a coarse curve (see the right graph of Figure 15) and to compute an associated Gini coefficient of 0.94. This value is much higher than the Gini coefficient calculated for CO_2 in this paper (for 2019 data) which equals 0.65. It suggests that besides being unequally distributed between countries, air transport is also unequally distributed within countries. A thorough study of inequalities within countries, involving a sociological approach, would give more weight to the analyses that are carried out in this section.

Finally, it is possible to compare the Gini calculated from [37] to more rigorous inequality metrics. Even tough it is difficult to talk about the distribution of GDP (value of goods and services produced

on a territory) within a population, some approaches measure the distribution of the income (earned by the citizens of a territory regardless of their location). For the reference, the associated world-level Gini coefficient was 0.67 in 2019 [48].

To highlight country-level air transport inequalities, various illustrative emissions distributions are proposed, and the results are provided in Figure 16 with a comparison to the actual distribution. For instance, it shows the changes in CO_2 emissions that would occur if these were distributed evenly between countries according to their respective populations. It also shows what would happen if all countries had the same air transport emissions per inhabitant as the United States. In this case, when combined, the world air transport CO_2 emissions would increase by 420%. If France is used as a reference instead of the United Sates, worldwide air transport CO_2 emissions would increase it by 175%. If everyone flew like in China or India, air transport CO_2 emissions would fall by 35% or by 86% respectively. The same conclusions can be observed if the exercise were carried out for ASK instead of CO_2.Large percentages of population in these countries have not flown on airplanes.

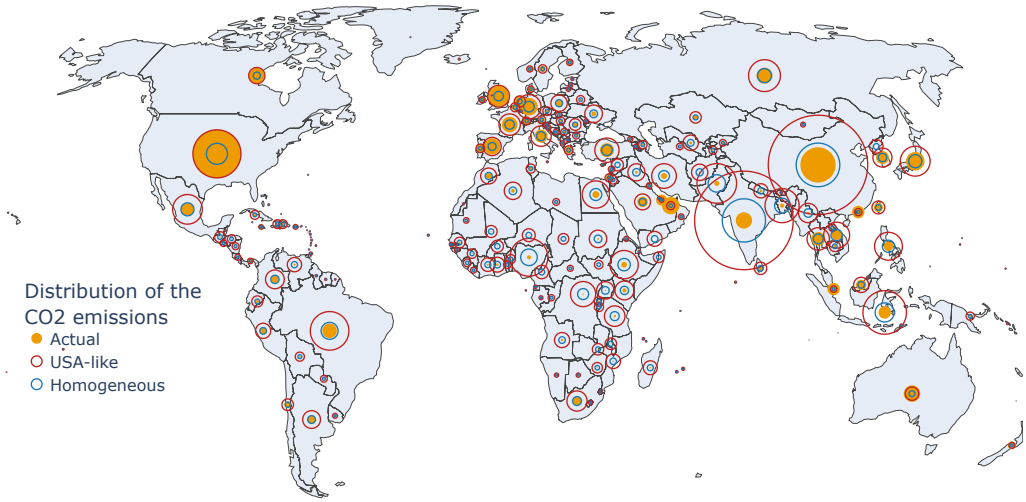

Figure 16. Departing passengers per country: actual, egalitarian and worldwide USA value. Areas proportional to CO_2 emissions

Lastly, another approach to study country-level air transport inequalities is to examine at potential determinants of the demand. Figure 17 represents the evolution of departures per inhabitant versus the relative GDP per capita (PPP). It is possible to fit a regression model to this data. The R^2 is low (0.35) when each country is considered equally important for the regression, but it increases to 0.84 when the regression is weighted according to their population. This could be explained by the numerous low-population countries with a high number of departures in the top left corner of the figure. A second-order polynomial model given in Eq. (8) is therefore proposed to give a rough estimate of the number of departing trips per inhabitant D_h as function of the GDP per capita in dollars, noted GDP_c.

$$D_h = -0.021 + 2.5 \cdot 10^{-5} \cdot GDP_c + 3.7 \cdot 10^{-10} \cdot GDP_c^2 \tag{8}$$

To improve this model, adding the share of exports attributed to tourism to the regression slightly increases the model accuracy (R^2=0.87) but it remains a minor improvement as this metric is more important for low-populated countries which are given low importance in the weighted regression.

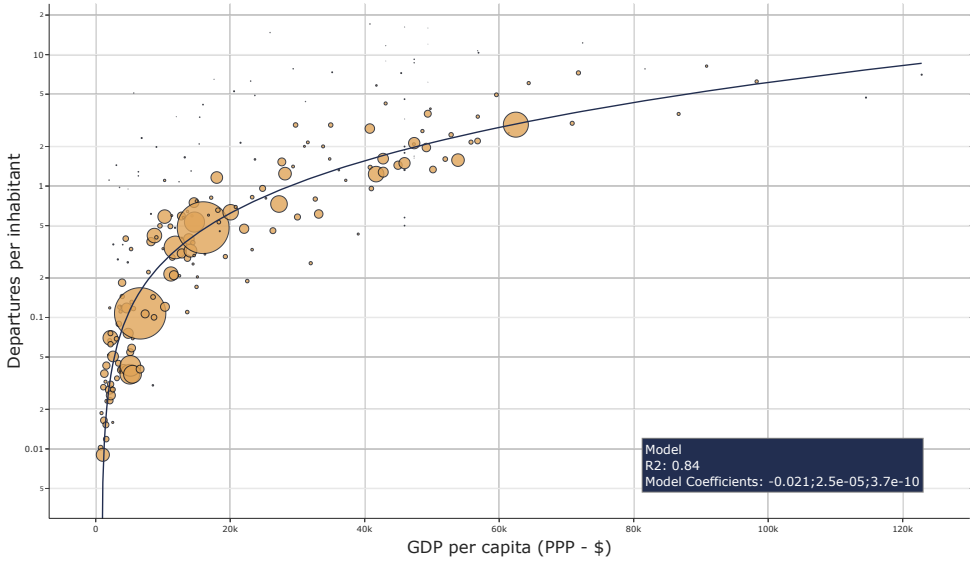

Figure 17. Departures per capita function of GDP per capita. Marker size is proportional to the population.

When the weights are removed, considering this second regression metric improves the R^2 from 0.41 to 0.57 (constant country set). This additional investigation is detailed in the supplementary material (see Reproducibility Section).

5. Conclusion

This manuscript presents a resource for aviation environmental research: a customizable traffic and CO_2 emissions dataset that allows to conduct thorough analyses of the aviation sector's carbon footprint. It fills a problematic gap in this research field: no open-source dataset with global coverage was available, making any resulting analysis reproducibility difficult.

To overcome this data limitation, open-source datasets (with limited coverage) were collected and compiled. However, this still does not provide satisfactory geographical coverage, particularly for domestic flights outside Europe and North America. To overcome this limitation, Wikipedia data on airports, 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 Available Seat Kilometres (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: kerosene consumption from the International Energy Agency, traffic volumes from OAG and the International Air Transport Association and CO_2 inventories from the International Council on Clean Transportation. In terms of seating capacity, the dataset 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 errors can remain in smaller aviation countries with an important share of domestic flights. 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. Thus, potential users of the dataset should ensure that their use case is not too sensitive to the traffic levels on individual routes.

To facilitate the exploration of the dataset, a dedicated graphical user interface was developed: AeroSCOPE. It allows interactive filtering of the dataset at continental, country, and route level. Two application cases were also described in this paper to study the potential of aircraft with new energy sources and current inequality in air travel.

On the one hand, it was found that a 500-NM battery-electric aircraft would cover 37.7% of the total seats offered, and 14.6% of the CO_2 emissions. Equipping 500 airports with the dedicated charging infrastructure would enable addressing 69.6% of this potential. Differences exists between countries: for instance, a 2000-NM hydrogen aircraft based in Norway could operate on routes representing 81.8% of their air transport emissions, but the same aircraft based in the Netherlands would address only 30.8% of their emissions. It highlights the need for local governments and airlines to pursue decarbonisation options adapted to their characteristics.

On the other hand, country-level air transport inequalities were studied by estimating the number of departing passengers per capita worldwide. A world average of 0.6 departures/inhabitant in 2019 was found, with an unequal distribution between countries (Gini coefficients above 0.6, not considering intra-countries inequalities). The inability of the approach of this paper to take into account the intra-country inequalities highlights a crucial area for future methodological improvements. A model to estimate the number of departures per capita as a function of GDP per capita (PPP) was also proposed.

The accuracy of the dataset 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 [49] 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. In addition, other relevant features could be added to enrich the model, especially if they are route specific. Some may also be substituted if more relevant data is found. Classification techniques could also be used to infer the aircraft types, to improve the fuel burn estimation. The airlines may also be inferred, refining the estimation by ensuring their network is consistent with the potential of their fleets. Graph Neural Networks (GNN) could be used 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.

This research was particularly made to provide a traffic and emissions dataset to enhance AeroMAPS, an open-source 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 scopes, thus opening new research avenues. Therefore, 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, it could serve various purposes. For instance, it can be used in sociological studies on the relationships between income and propensity of air travel and related topics. Finally, this work was performed for the year 2019. Maintaining the database over time raises however the question of process automation, as most of the data collection and processing work was made manually using notebooks. The compilation work relies

on the continued availability of the sources used, which is not guaranteed. Nevertheless, meeting the challenge of 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 18 is a representation of the relative importance of the various features used in the XGBoost regressor. The F-score represents the number of times the feature was selected to split the tree during the training process. Each bar is related to one of the features listed in Section 2.3.2. The most prevalent features are the distance between the airports, the passenger traffic of airports (named *consolidated_pax* in the figure) and the number of airlines (named by the category of the relation: *Regular* but also *Seasonal*, *Seasonal charter*). The population surrounding each airport (named *pop_X_Y*, *X* being replaced by the distance in km and *Y* by the departure/arrival status of the airport in the figure) is also important, although the co-linearity of the various population metrics, in a radius of 30, 70 and 150 km around the airports reduces 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_Y and airport_X_Y*). Although the trade flows are an important feature, other socio-economic features are less used by the training process.

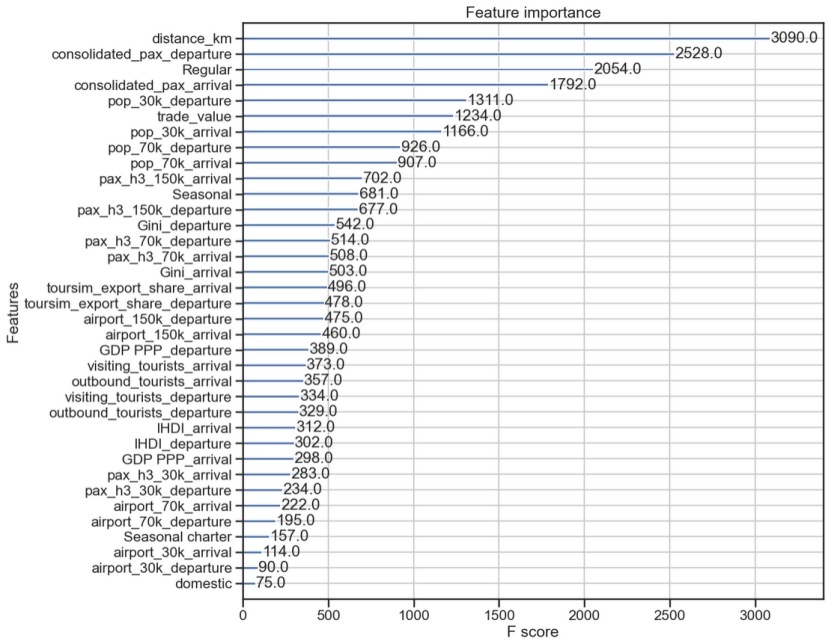

Figure 18. XGBoost feature importance.

Figure 19 is a screenshot of the AeroSCOPE tool. As explained in the article, it features 3 different aggregation levels of the dataset, each corresponding to a tab of the user interface. In the case of Figure 19, the dataset is aggregated at the country level, and departures are filtered for Finland. A map displays the emissions associated to each destination country and two additional plots below analyse the distribution of CO_2 emissions according to flight distance and to aircraft type.

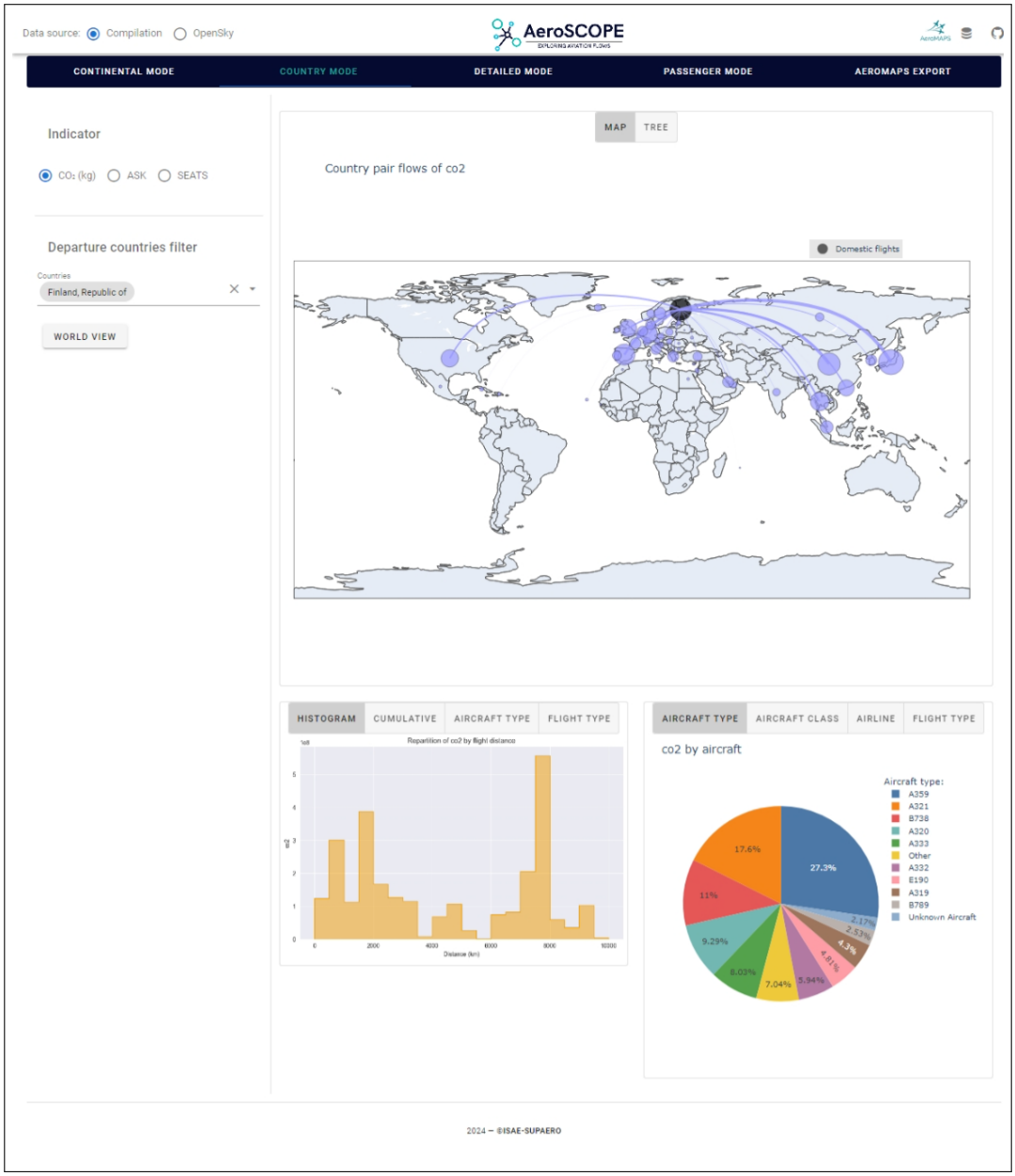

Figure 19. AeroSCOPE user interface.

Appendix 2. Airport selection algorithm

Algorithm 1 find_score(top_potential_airports, top_neighbours_lists, nmax, n_neighbours)

Input:
top_potential_airports: Ordered list of the top potential airports
top_neighbours_lists: Dictionary associating an ordered list of top-potential neighbours to each airport
nmax: Number of airports to select
n_neighbours: **Optimisation parameters:** List of the number of neighbours to select for each airport.

```
selected_airports ← []
i ← 0
count ← 0
while count < nmax and i < nmax do
    if n_neighbours[i] > −1 then                          ▷ Condition for the top-airport to be selected
        ap ← top_potential_airports[i]
        if airport not in top_list then                   ▷ Add the airport to the selection
            elected_airports.append(airport)
            count ← count + 1
        candidate_neighbours ← top_neighbours_lists[airport] [:n_neighbours[i]]   ▷ Getting top neighbours
        for el in candidate_neighbours do
            if count < nmax and el not in selected_airports then   ▷ Add the neighbour to the selection
                selected_airports.append(el)
                count ← count + 1
    i ← i + 1
score_i ← compute_potential(selected_airports)   ▷ Computing share of potential between airports of the set
if count < nmax then   ▷ Penalisation of cases where less than 500 airports are selected (exit on i)
    score_i ← score_i/2
return score_i
```

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. Note that external inputs, too large for online data storage, require user action. When applicable, these actions are described in the same file.

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 main 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, the notebooks for the applications are stored in *04_applications*. The AeroSCOPE web application is accessible at https://aeromaps.eu/aeroscope and the related code is stored at https://github.com/AeroMAPS/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: *Journal of Open Aviation Science* 1.1 (Dec. 2023). DOI: 10.59490/joas.2023.7147. (Visited on 01/18/2024).

[20] Antoine Salgas, Junzi Sun, Scott Delbecq, Thomas Planès, and Gilles Lafforgue. "Compilation of an open-source traffic and CO2 emissions dataset for commercial aviation". In: *Proceedings of the 11th OpenSky Symposium*. Journal of Open Aviation Science, Oct. 2023. DOI: 10.59490/joas.2023.7201. (Visited on 01/19/2024).

[21] International Air Transport Association. *Air Passenger Market Analysis*. Tech. rep. Aug. 2023. URL: https://www.iata.org/en/iata-repository/publications/economic-reports/air-passenger-market-analysis---august-2023/.

[22] 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).

[23] 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.

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

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

[26] 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).

[27] 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).

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

[29] 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).

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

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

[32] 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).

[33] 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).

[34] 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.

[35] 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).

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

[37] 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.

[38] ICAO. *Manual on Air Traffic Forecasting*. Tech. rep. 2006. URL: https://www.icao.int/MID/Documents/2014/Aviation%20Data%20Analyses%20Seminar/8991_Forecasting_en.pdf.

[39] Andreas M. Tillmann, Imke Joormann, and Sabrina C.L. Ammann. "Reproducible air passenger demand estimation". In: *Journal of Air Transport Management* 112 (Sept. 2023), p. 102462. ISSN: 0969-6997. DOI: 10.1016/j.jairtraman.2023.102462. URL: http://dx.doi.org/10.1016/j.jairtraman.2023.102462.

[40] 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.

[41] 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.

[42] 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.

[43] Adeline Montlaur, Luis Delgado, and César Trapote-Barreira. "Analytical Models for CO2 Emissions and Travel Time for Short-to-Medium-Haul Flights Considering Available Seats". In: *Sustainability* 13.18 (Sept. 2021), p. 10401. ISSN: 2071-1050. DOI: 10.3390/su131810401. URL: http://dx.doi.org/10.3390/su131810401.

[44] 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).

[45] Albert R. Gnadt, Raymond L. Speth, Jayant S. Sabnis, and Steven R. H. Barrett. "Technical and environmental assessment of all-electric 180-passenger commercial aircraft". In: *Progress in Aerospace Sciences* 105 (Feb. 2019), pp. 1–30. DOI: 10.1016/j.paerosci.2018.11.002.

[46] Airbus. *ZEROe - Low carbon aviation - Airbus*. Section: Innovation. June 2021. URL: https://www.airbus.com/en/innovation/low-carbon-aviation/hydrogen/zeroe (visited on 11/20/2023).

[47] Nikolaus Hansen, Yoshihikoueno, ARF1, Gabriela Kadlecová, Kento Nozawa, Luca Rolshoven, Matthew Chan, Youhei Akimoto, Brieglhostis, and Dimo Brockhoff. *CMA-ES/pycma: r3.3.0*. 2023. DOI: 10.5281/ZENODO.7573532. URL: https://zenodo.org/record/7573532.

[48] L. Chancel, T. Piketty, E Saez, and G. Zucman. *World Inequality Report 2022*. Tech. rep. World Inequality Lab, 2022.

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