Travel Impact Model (TIM)
ADVISORY COMMITTEE

TECHNICAL BRIEF

Methodology for validating fuel burn model changes

18 November 2024

EXECUTIVE SUMMARY

This technical brief presents the validation methodology adopted for the Travel Impact Model (TIM) to assess how close the TIM fuel burn estimates are to real-world fuel burn values. The methodology was developed by the Engineering team of Google in partnership with the International Council on Clean Transportation (ICCT), and discussed in two Task Group (TG) meetings with experts that are part of or delegated by the TIM Advisory Committee (AC). The framework development and discussions took place in 2023. In January 2024, the AC agreed to incorporate the validation methodology into the TIM workflow. Since then, this framework is applied every time a model change that affects fuel burn is proposed, to verify if the change improves the model.

Incorporating this validation framework into the TIM aims at increasing the model's accuracy and consistency. Also, it contributes to TIM's transparency as it is a reproducible methodology. In this document, we present the validation framework and an example of its application to analyze the effectiveness of a model change.

As the TIM estimates emissions at the flight level, ideally, model validation would be applied considering actual fuel burn data also at the flight level, and if possible, from several markets. Preferably, the TIM validation would be developed using public datasets to freely test any model change and disclose results. However, public historical flight data at the needed granularity is scarce. The only public dataset we identified that contains fuel burn data at the flight level is from Brazil. A limitation of this dataset is that fuel burn data is only provided for Brazilian airlines, which reduces market and aircraft coverage, especially for long-haul flights.

To deal with this lack of public data, Google has partnered with a growing number of airlines that provide actual operational data, including fuel burn. However, this data is private, which limits validation transparency. Given these data limitations, we combine public data from Brazil with the anonymized version of the private airline data to represent the validation sample. The combined sample considers operations from 2019 in Brazil and 2019, 2021, and 2022 for the partner airlines. The aircraft models represented in the dataset account for 76.2% of the global flights in 2019.

The proposed validation framework is divided into three steps: i) clean the sample to remove flights that are not relevant, ii) aggregate fuel burn data to make it comparable with estimated emissions, and iii) apply a set of metrics to evaluate the model. Step ii) aims to statistically define a representative fuel burn value for a given route-aircraft-airline combination to be compared with the model estimates. Establishing a representative fuel burn value is needed as the dynamic and stochastic behavior of operational and weather conditions create variations in the fuel burned for a specific aircraft operated by a single airline on a unique route. In step iii), we consider as error the difference between the estimated fuel burn and the representative real fuel burn for each route-aircraft-airline group. The metrics analyzed include: general error metrics (median absolute error, under/overestimation trends), distribution of errors, distance-based error metrics, and distance- and aircraft-based error metrics. For the distance-based and distance- and aircraft-based metrics, we calculate weighted averages considering the global flight distribution to scale the results of the sample to the global context.

We provide a practical example of the validation framework application, analyzing the implementation of the distance correction factor, a model change that was under discussion when the validation methodology was being developed. We compared the real-world fuel burn with the TIM estimates, considering two versions of the model: baseline (TIM 1.8.0) and model with the distance correction application. We observe that the error metrics individually may have conflicting results, and that the impact of a model change may be unevenly distributed, improving some portions of the flights, but potentially worsening others. The decision about whether a change should be implemented or not needs to consider the aggregated set of error metrics in a careful analysis, considering what each metric represents.

The recommended approach to approve model changes is the following:

1. The validation analysts (Secretariat or Google) run the baseline model and the alternative model, which is the model with the application of the proposed change, through the validation process. The analysts summarize the validation results, including the error metrics and a discussion of the findings.

2. The group investigating the change (TG or AC) and the validation analysts meet and discuss the validation results. If discussed at the TG level, the group should reach a consensus on a recommendation to the AC.

3. The AC decides taking into account the validation results and the recommendations of the TG.

In addition to characterizing the impact of a model change, the application of the validation framework helps identify possible modeling issues, which can motivate future improvements. The validation methodology will keep improving and new metrics may be added in the future, as model changes are tested, and as more real-world data is incorporated into the sample to increase the market coverage.

CONTENTS

1. OBJECTIVE

The objective of the model validation work developed by the International Council on Clean Transportation (ICCT) and Google was to establish a transparent and reproducible validation methodology for the Travel Impact Model (TIM), increasing its accuracy and consistency. The validation methodology will also be used as a key decision input to analyze if proposed model changes are actually improving the model by providing estimates that are closer to real-world flights. The validation methodology helps inform the AC decision-making process on whether to adopt particular changes to the model. The final goal is to publish the validation methodology and results, giving external parties the tools to reproduce and scrutinize our validation claims.

In addition to presenting the validation methodology, this document discusses if the proposed methodology effectively analyzes how close the fuel burn estimates are to actual fuel burn, and how consistent the model is across different airlines and markets. An application of the validation to analyze a model change is presented as an example, considering the distance correction proposed to be implemented. This document also summarizes the recommendations that were shared at the 3rd Advisory Committee (AC) meeting in January 2024 and considers the discussions and suggestions from two task group meetings held in the fourth quarter of 2023. At the 3rd AC meeting, the group agreed to incorporate the proposed validation methodology into the TIM and since then, this methodology has been applied to analyze any model changes that affect the fuel burn estimates.

2. VALIDATION BACKGROUND: CURRENT VALIDATION AND MAIN CHALLENGES

One of the main challenges of the TIM model validation is having access to reliable data on past flights and their fuel burn, from different airlines and different markets, at the needed granularity. Ideally, model validation would be carried out using fuel burn data at the flight level, the data specificity of the TIM estimates. The Google team has access to real-world flight data reported by some partner airlines. However, this data is private.

To be transparent, the TIM validation would ideally adopt public datasets containing fuel burn at the flight or the route level to freely test the model and disclose findings. This kind of data is, unfortunately, very scarce. There are very few markets with public flight and fuel burn data. U.S. Department of Transportation Form 41,[1] for example, provides fuel burn data, but at the aircraft type level only, aggregated by airline and month. Brazil's Microdata[2] was the only public dataset we were able to identify that provides fuel burn data at the flight level. The lack of appropriate public fuel burn

1 This is available at https://www.transtats.bts.gov/Tables.asp?QO_VQ=EGI&QO_anzr=Nv4%FD-Pn44vr4%FDSv0n0pvny%FDer21465%FD%FLS14z%FDHE%FDSv0n0pvny%FDQn6n%FM&QO_fu146_anzr=Nv4%FDPn44vr4%FDSv0n0pvny

2 Brazil's Microdata available at https://www.gov.br/anac/pt-br/assuntos/regulados/empresas-aereas/Instrucoes-para-a-elaboracao-e-apresentacao-das-demonstracoes-contabeis/envio-de-informacoes

data from different markets to test the model also increases the risk of not having geographic representativeness.

Public data, whenever available, will be prioritized, as it contributes to the validation transparency. We expand the validation sample with the private data to increase aircraft model coverage and market representativeness, although this data combination limits transparency. Currently, the analyzed sample combines ANAC data with an anonymized version of the private data shared by airlines.

2.1. ANAC Microdata

Brazil's flight data ("Microdata") is published monthly by the National Civil Aviation Agency (ANAC) and contains information about all domestic and international flights operated to and from Brazil since the year 2000. Besides the flight information, including flight number, airline, aircraft model, and airport-pair, it provides operational data such as the number of passengers transported and cargo mass carried for a given flight. For Brazilian airlines, some additional information is provided, such as actual fuel burn and aircraft tail number. Table 1 summarizes the information available in the ANAC Microdata.

Table 1. Data available in ANAC Microdata

Category	Data
Flight Information	• Flight number • Airline • Service type (pax or dedicated cargo) (only for Brazilian airlines) • Departure and arrival date and time
Airport-pair	• Origin and destination airports/cities/countries • Great circle distance
Aircraft	• Model • Seats • Maximum payload • Tail number (only for Brazilian airlines)
Operations	• Number of passenger transported and/or mass of cargo • Baggage mass (only for Brazilian airlines) • Fuel consumed (only for Brazilian airlines)

For the validation analysis, we selected ANAC data from 2019 to avoid including operations impacted by the COVID-19 pandemic in our sample. For the validation analysis, we initially selected ANAC data from 2019 to avoid including operations impacted by the COVID-19 pandemic in our sample. We do not use older data at this point, because airline practices and aircraft in use change over time. We recommend including more years of operation as a refinement.

Table 2 presents a summary of ANAC Microdata, considering flights from 2019. We see that flights performed by Brazilian airlines—the flights with fuel burn information—represented 90% of the total flights performed that year, and 87% of passengers transported. The sample considered in our analysis is highlighted in red in the table

and includes only passenger flights performed by Brazilian airlines. We remove flights operated by foreign airlines as the corresponding fuel burn information is not available, and we disregard dedicated cargo flights as the TIM focuses on passenger flights. More information about Brazil's commercial aviation market and the sample selection is detailed in Appendix 1.

Table 2. Summary of Brazilian flights in 2019

Airline nationality	Route type	Service type	Number of flights	Number of airlines	Total passengers (million)	Total fuel (million L)	Number of aircraft models	Number of aircraft
Brazilian	Domestic	Passenger	789,072	10	98.14	3,412.6	26	525
Brazilian	International	Passenger	50,457	7	9.03	1,483.8	19	362
Brazilian	Domestic	Dedicated cargo	13,279	7	0	90.4	6	36
Brazilian	International	Dedicated cargo	1,581	2	0	38.9	2	5
Foreign	Domestic	Unreported	1,213	35	0.27	Unreported	26	Unreported
Foreign	International	Unreported	84,417	89	15.42	Unreported	40	Unreported

Considering the scope of the validation work, the main strength of the ANAC Microdata is that it contains fuel burn and other operational information at the flight level, which means it can be directly compared with the TIM estimates. Also, data is available since the year 2000, providing a long time series for testing. However, fuel burn data is only available for Brazilian airlines, reducing the airline and aircraft coverage, and limiting the analysis, especially for international flights (and long-haul flights, as a consequence). In addition, Brazilian domestic flights are highly concentrated on a few airlines, and the traditional low-cost-versus-legacy-carriers market dynamics seen in other countries may not apply to Brazil. Lastly, data is self-declared, and there is no third-party audit, which means reporting errors may not be identified.

2.2. Private airline data

The validation sample includes passenger flight data, privately shared by airlines. Given data confidentiality restrictions, the airlines are anonymized and always treated as a block, never as individuals. Currently, the sample includes data from several airlines, considering operations from the years 2019, 2021, 2022, and 2023. However, data is not always provided at the flight level. For some cases, fuel burn is provided at the aircraft and route level, being represented by the mean fuel burn of all flights performed in a given route and year. Table 3 presents a summary of the private airlines' data.

Table 3. Summary of private airline data

Service type	Number of flights	Total fuel (million L)	Number of aircraft models
Passenger	2,306,758	24,450	41

The Google Engineering Team is continuously working with airlines to expand the flight data sample and market representativeness. But, like the ANAC data, the private airline data is self-declared and there is no third-party audit.

2.3. Market representativeness of the data

To analyze if the sample's fuel burn data represents the global market, we calculate its distance and aircraft coverage. Figure 1 shows the flight distance distribution of the sample, considering the number of flights (a) and the total fuel burn (b), compared to global flight distributions. Global flight distance data were sourced from OAG Aviation Worldwide Limited (2019)[3], and global fuel burn come from ICCT's Global Aviation Carbon Assessment (2019)[4]. In Figure 1 (a), we see that most of the flights in our sample have distances shorter than 1000 nautical miles and that there are very few flights longer than 5500 nautical miles. The distribution of flight distances is roughly similar to the global distribution of flight distances. Figure 1 (b) shows that longer flights contribute (proportionally) more to emissions than shorter flights, given that flights from groups of long distances burn more fuel, even with relatively lower flight frequency.

3 Historical flight schedules data provided by OAG Aviation Worldwide Limited is available at https://www.oag.com/airline-schedules-data

4 Brandon Graver, Dan Rutherford, and Sola Zheng, *CO_2 Emissions from Commercial Aviation: 2013, 2018, and 2019* (International Council on Clean Transportation, 2020), https://theicct.org/publication/co2-emissions-from-commercial-aviation-2013-2018-and-2019/.

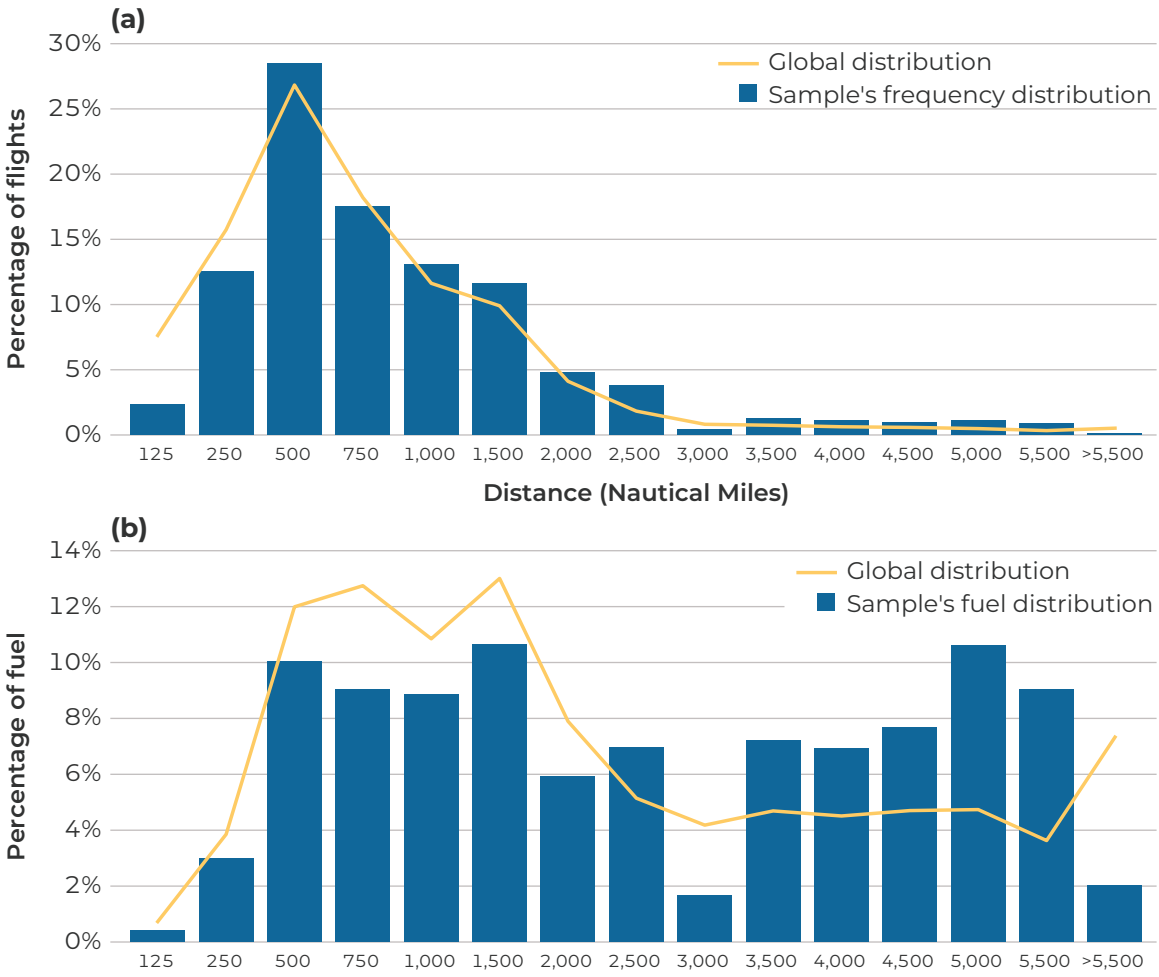

Figure 1. Comparing the distance distribution of the validation sample (ANAC + private airline data) and global operations in 2019 based on flight frequency (a) and fuel (b)

In relation to aircraft model coverage, we calculated that 60.4% of 2019 global flights[5] are performed using aircraft models that are available in ANAC Microdata from 2019. Considering private data shared by the airlines that have a partnership with Google, we observe that the aircraft models they operate account for 71.0% of global flights. When combining both data sources, the data coverage increases to 75.3%. Table 4 summarizes these results.

Table 4. Global flight coverage according to the aircraft models available in different flight data sources

ANAC	Private airlines data	All sources (ANAC + private)
60.4%	71.0%	75.3%

5 See footnote 3.

3. COMPARING REAL FUEL BURN WITH THE TIM

The fuel burn of a flight tends to increase with the route distance, but it also depends on the aircraft technology, weather conditions, operational factors such as payload mass and speed, and many other factors, including aircraft maintenance and airport congestion. As a consequence, fuel burn variations are expected even for a single aircraft of a single airline, operating on a specific route. We can observe this behavior in Figure 2. This figure presents a scatterplot of fuel burn and distance for individual flights on the four most used aircraft models by Brazilian airlines in 2019, compared to the TIM estimates (dashed line).

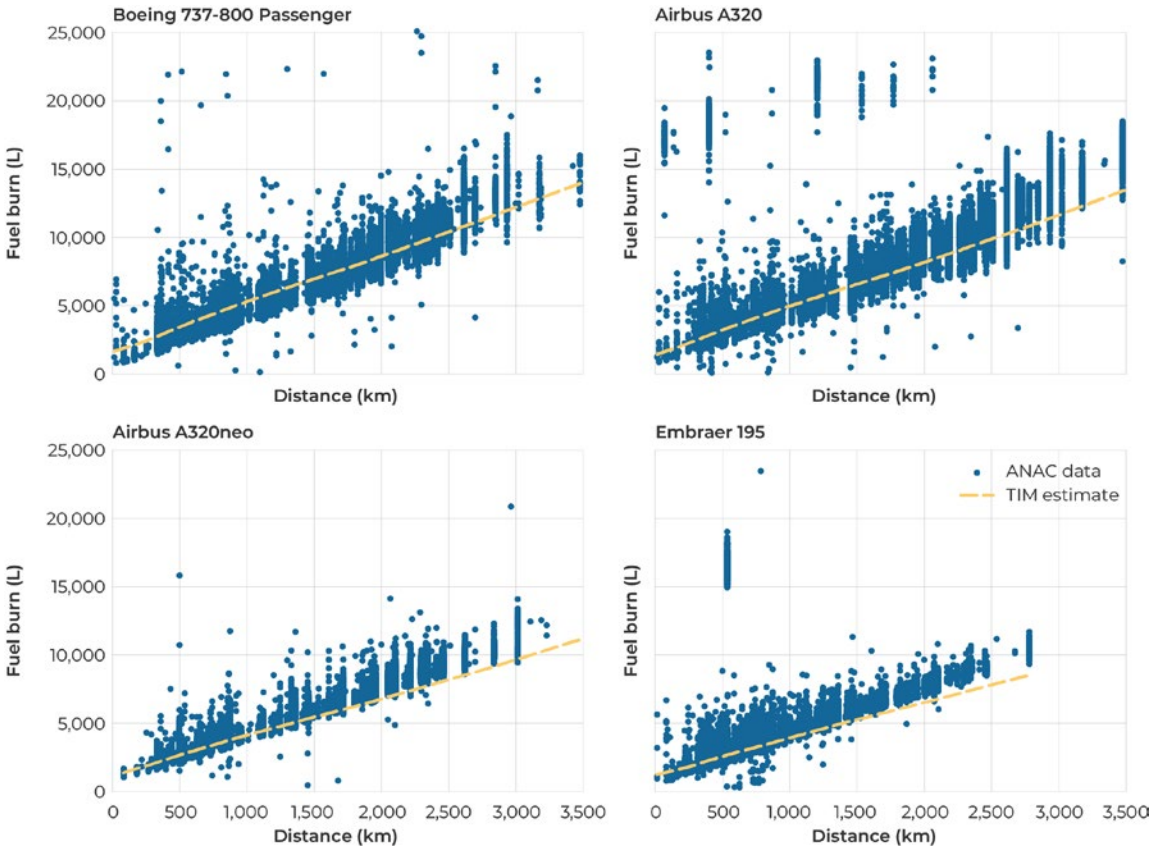

Figure 2. Fuel burn versus distance for each individual flight by the four most common aircraft types in the ANAC data in 2019

4. VALIDATION METHODOLOGY

Validation occurs by the following three steps:

1. Clean sample to remove flights that are not relevant.
    a. Remove flights that have no reported fuel burn or are reported as 0 L.
    b. Remove dedicated freighter flights, which are not supported by the TIM.
2. Aggregate fuel burn data to make it comparable with estimated emissions.

a. Considering that there is typically a large variance of fuel burn for even the same aircraft operating the same route on different occurrences, define a representative fuel burn value for a given group, for example mean or median fuel burn for an aircraft/carrier/route. This is needed to make data from different sources compatible, as part of the private airline data is not reported at the flight level, but at the route level.

b. Process:
    i. Start with raw data at the flight level (flight/airport-pair/airline/date/time).
    ii. Aggregate by carrier, aircraft, origin, and destination.
    iii. Remove aggregations that have fewer than 50 individual samples.[6]
    iv. Define a representative fuel burn value for each group. We suggest that the median of the distribution be adopted as a representative value. However, we are currently adopting the mean to make data from different sources compatible, as some airlines that have privately shared data reported the mean fuel burn of their routes.
    v. Some additional data cleaning may be needed if there are outliers biasing the average calculation. A possible mitigation process would be removing the extreme fuel burn values of each group (keeping data within the 90th percentile, for example). We have not applied this process yet, but it could be a future refinement if the need is identified.

The aggregation should allow a direct comparison between the actual fuel burn values and the model estimates. We are currently using fuel burn aggregated by aircraft/carrier/route as the base model adopted by the TIM provides a single function of fuel burn for each aircraft that depends only on distance. In addition, with this aggregation, we avoid having the metrics skewed toward more frequent routes.

3. Apply a set of metrics to evaluate the model, for example:
    a. General error metrics: median absolute error, under/overestimation percentages
    b. Distribution of errors
    c. Error as a threshold (what % of results are within X% of actuals)
    d. Distance-based error metrics
    e. Distance- and aircraft-based error metrics
    f. Other metrics as defined

These metrics are detailed in the following section.

6 The limit of 50 flights to include a group of airline-aircraft-route in the sample is a working assumption, and we are investigating some statistical tests for the sample size definition to avoid unnecessary data elimination.

4.1. Error metrics

For each unique carrier, aircraft, and route combination, calculate the relative difference in the TIM estimate to the mean real-world fuel burn reported in the sample, as defined in Equation (1). The analyses reported in this document consider the estimates provided by the TIM 1.8.0.

$$\frac{[Fuel\ burn]_{TIM} - [Fuel\ burn]_{Mean\ value}}{[Fuel\ burn]_{Mean\ value}} \quad (1)$$

A positive error value indicates that the TIM is overestimating the fuel burn while a negative value indicates an underestimation of the fuel burn. Considering the sample composed of ANAC data and the private data from partner airlines, we find that the TIM underestimates fuel burn 76% of the time while it overestimates 24% of the time.

For the same sample, taking the absolute value of the errors, the median error over all aggregate groups is 8.0%. We adopt the absolute values to avoid negative errors (underestimation) being canceled out by positive errors (overestimation). While this single value is useful to characterize the accuracy of the model, evaluating the precision of the model requires looking at the distribution of errors.

4.2. Distribution of errors

There is variation in the error calculated for each aircraft-carrier-origin-destination combination. The distribution of the errors is shown in Figure 3. The frequency distribution on the left shows the peak of the distribution being lower than 0, which indicates a tendency to underestimate fuel consumption. The distribution is also not mirrored on either side of the peak. It has a longer tail on the right side (overestimation) than on the left (underestimation). On further investigation, it was found to be due to larger errors found at shorter route lengths. These trends are confirmed by the cumulative distribution of errors shown on the right.

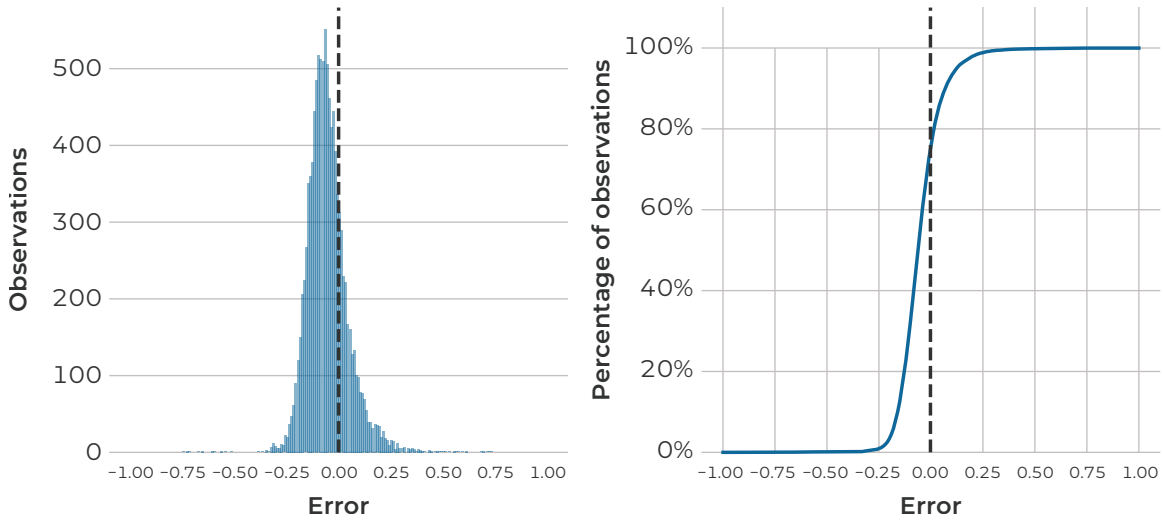

Figure 3. Frequency (left) and cumulative (right) distributions of the error in the TIM's fuel burn estimates, considering the real-world fuel burn from the combination of ANAC 2019 and private airlines data

4.3. Error thresholds

One aim is to use this validation methodology to quantify how suggested changes to the model improve its performance. For this purpose, it is useful to define the percentage of estimates that are below specific error thresholds. We choose threshold levels of 2%, 5%, 10%, and 20% and show the percentage of estimates that are below these levels (Table 5), considering absolute errors. To illustrate how to read the table, 13% of the estimates from the TIM are within 2% of the median fuel burn reported in the actual flight sample and 62% of the estimates are within 10% of the median fuel burn reported in the sample.

Table 5. Percentage of observations that are below certain absolute error thresholds for the baseline TIM estimates

Weights	Error threshold	Percentage of observations
0.1	<2%	13%
0.2	<5%	31%
0.3	<10%	62%
0.4	<20%	95%
	Weighted average	**63.8%**

To aggregate these metrics further, we use weightings to describe the relative importance of each error threshold. The priority is to reduce the largest errors, while getting the error to less than 2% is less important. Consequently, we ascribe increasing weights: 0.1, 0.2, 0.3, and 0.4 to the <2%, <5%, <10%, and <20% thresholds, respectively. The weighted average number for the baseline TIM estimates is 63.8%. While the exact value of the number cannot be ascribed to a physical meaning, a higher value is desirable. The best possible score is 100% and would only happen if all errors are <2%. The worst possible score is 0% and would only happen if all errors are >20%.

The table of values and the weighted average number provide numerical quantifications of the threshold. The error thresholds can also be visually represented by a cumulative distribution of the absolute errors (Figure 4).

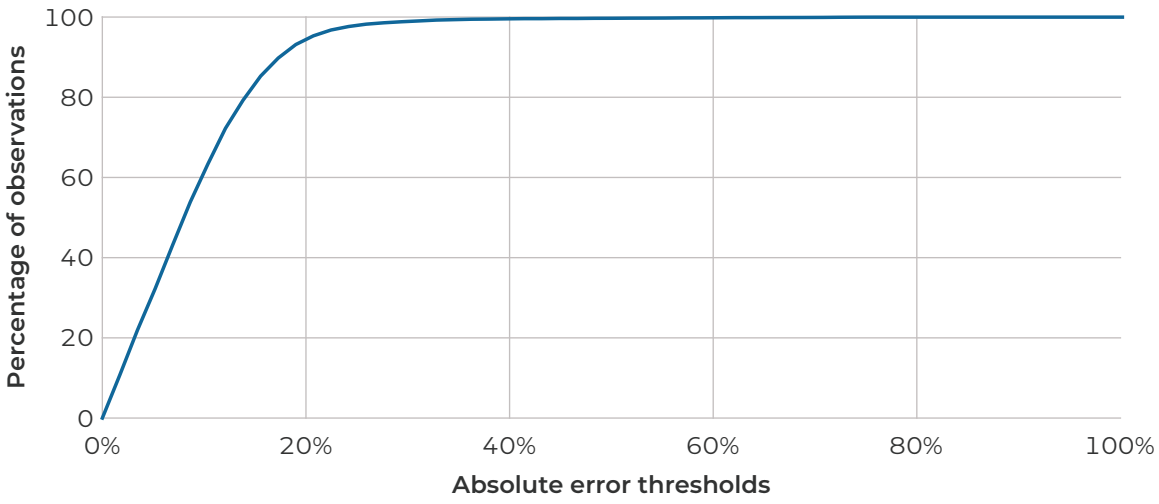

Figure 4. Cumulative distribution of the absolute error thresholds

Given that the fuel burnt on a flight is highly dependent on the aircraft and the distance it flies, we also characterize the errors along these dimensions.

4.4. Distance-based error analysis

To show trends of the error with the distance flown, the carrier-aircraft-origin-destination combinations are binned based on the distance. Then the average of the errors for all the combinations in that bin is taken, weighted by the number of flights. Figure 5 plots a bar graph of the trend for ANAC 2019 (a) and the private data shared by partner airlines (b). Here the binning done is "right inclusive." The 2,000 nautical mile bin, for example, includes all combinations that were greater than 1,500 nautical miles (NM) and less than or equal to 2,000 NM.

The figure shows that the TIM 1.8.0 tends to underestimate fuel burn, except for the short-haul flights reported by ANAC, for which we observe an overestimation pattern. This overestimation trend is seen for the flights shorter than 500 NM, and the highest error is seen for the shortest flights, in the 125 NM bin. There are a few potential reasons for this overestimation behavior. 125 NM is the shortest distance for which the underlying EEA model provides fuel burn data. As a consequence, any flights shorter than 125 NM require extrapolation. Additionally, the default LTO cycle is ~30 minutes long, which may be a poor assumption for short flights, which are likely to operate from smaller airports with more flexible landing and takeoff procedures that could shorten the LTO cycle time and reduce emissions. Lastly, the EEA model calculates LTO fuel burn assuming operating at European airports, which may be more congested than Brazilian airports and less representative of other markets. Further analysis of short-haul flights and the LTO cycle fuel burn will be conducted in Workstream 7 (improving model granularity).

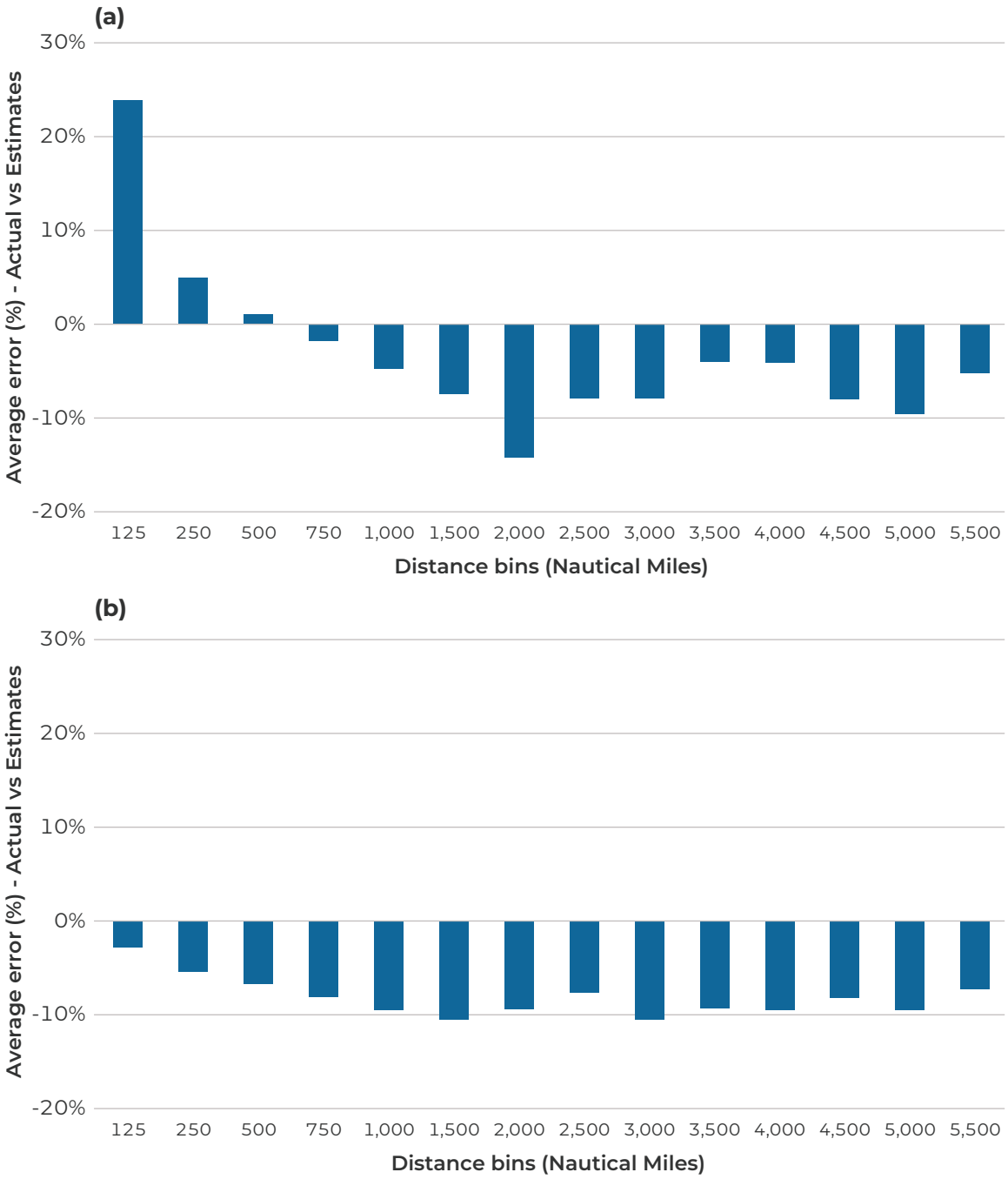

Figure 5. Weighted average error by distance bins based on the distance traveled for ANAC 2019 (a) and private airline data (b).

In general, we observe a consistent underestimation trend for longer-distance flights, while the trend is inconsistent for shorter-distance flights, depending on the data source. Figure 6 shows the weighted average error for the combined sample, including both ANAC and private airline data. With the sample combination, the underestimation trend prevails for the fuel burn estimation of longer flights, and we observe a small overestimation trend for flights shorter than 250 NM.

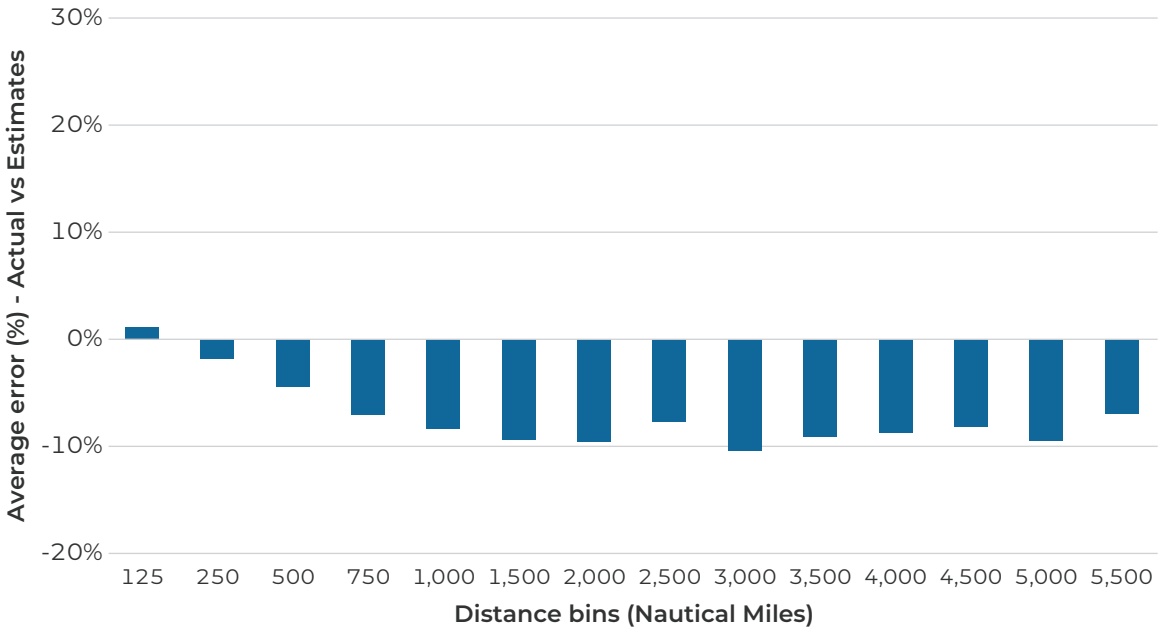

Figure 6. Weighted average error by distance bins based on the distance traveled for the complete sample, combining ANAC 2019 and private airline data

The analysis shown above considers the real value of the error, including negative (underestimation) and positive (overestimation) errors. Although this is useful to analyze the general error trend in a given distance bin, negative values cancel out positive values, and it does not give a reliable average error magnitude. To capture that, we generate a version of the chart presented in Figure 6 but considering the absolute errors (Figure 7).

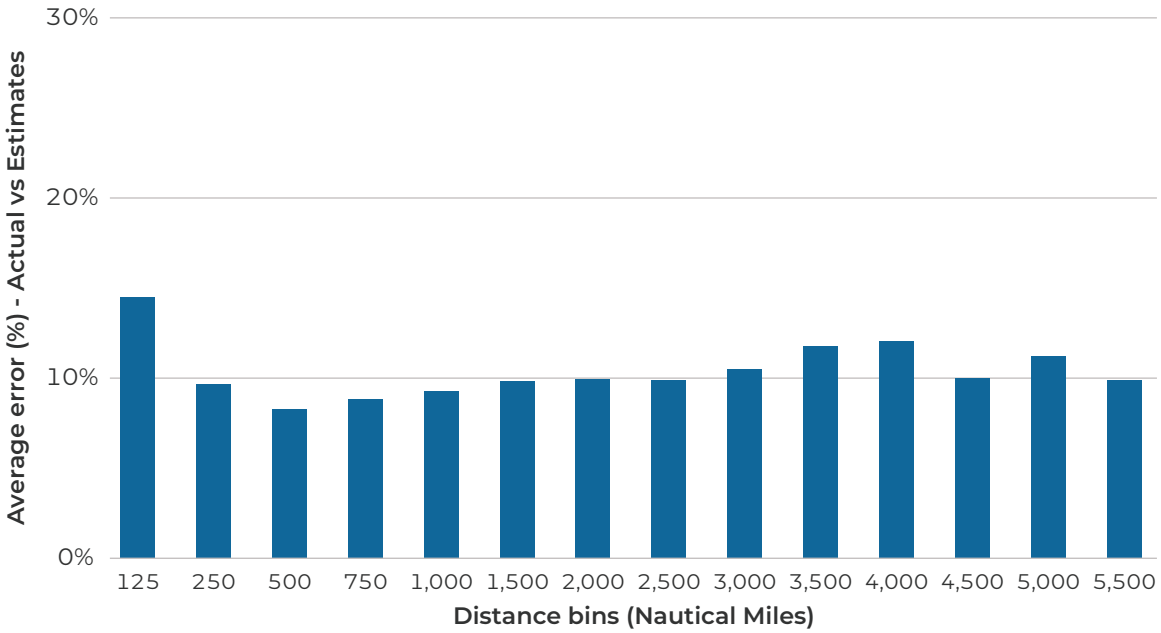

Figure 7. Weighted average absolute error by distance bins based on the distance traveled for the complete sample, combining ANAC 2019 and private airline data

To reflect the distance distribution from the global market and not from the sample (see Figure 1), we calculate an average error metric weighting results from the combined sample by the global distribution. For the weighting, two metrics are used:

1. The percentage of global flights in each distance bin
2. The percentage of global emissions in each distance bin

These two weightings help scale the results of the sample to the global context, but they highlight two differing objectives. Weighting by the number of flights puts more importance on getting the emissions correct for each individual flight and emphasizes the performance on shorter distance flights which occur more frequently. Weighting by emissions puts more importance on getting the emissions correct for the more polluting flights and emphasizes the performance on longer distance flights which emit more CO_2.

When weighted by the distribution of flights, the weighted average absolute error metric is 9.5% while when weighted by the distribution of emissions, the weighted average absolute error metric is 9.8%. While there is no physical interpretation of this metric, a value closer to zero is desirable.

4.5. Distance- and aircraft-based error analysis

To investigate the trend in fuel burn in greater detail, we separate out the fuel burn trends of each aircraft type. Still working with the aggregated data, we now filter for each aircraft type and plot the mean fuel burn versus route distance. Figure 8 presents the mean fuel burn from the ANAC sample as blue dots and the fuel burn estimate from the TIM as a dashed line. Each subplot represents one of the four most common aircraft types in the analyzed sample. There is also a R^2 value for each subplot. The R^2 value is a metric used in linear regression and is a quantification of how well the fuel burn estimates from the TIM fit the real-world fuel burn data. It is calculated as described by Equation (2).

$$R^2 = 1 - \frac{\Sigma_i (y_i - \hat{y}_i)^2}{\Sigma_i (y_i - \bar{y}_i)^2} \tag{2}$$

For a given aircraft, i is the route-airline combination, y_i is the median fuel burn of real operation data, $\hat{y}_i$ is the TIM estimate and $\bar{y}$ is the mean fuel burn for that aircraft. Higher values of R^2 indicate a better fit.

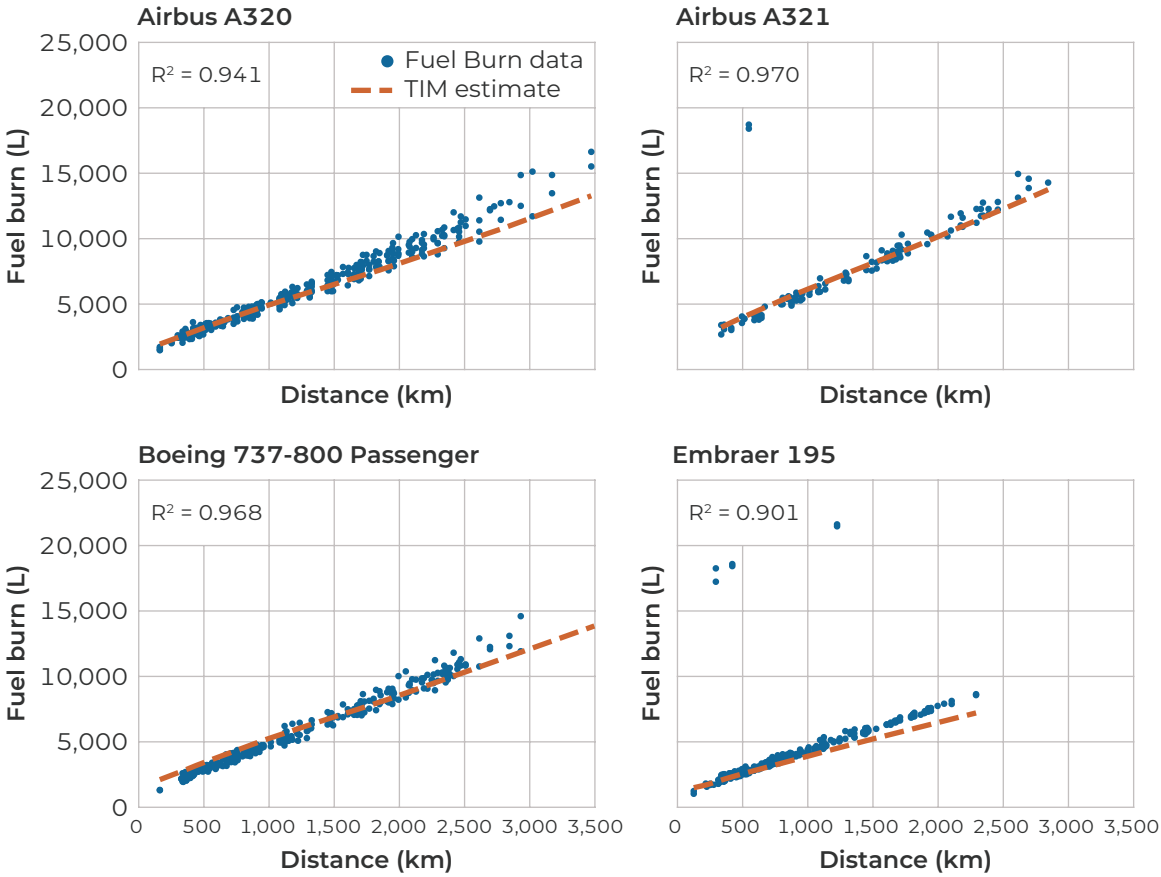

Figure 8. Fuel burn versus distance trends for the four most common aircraft in the ANAC data

Similar to the distance-based error analysis, there is a need to aggregate the performance over each aircraft into a single error metric. In the case of this distance and aircraft based error analysis, we calculate a weighted average of the R^2 values for all the aircraft that are represented in the validation data. The weighting is determined by the percentage of global flights that are flown by the specific aircraft. The aircraft represented in the validation sample cover about 76.2% of global flights. The weighted average R^2 value of all the aircraft, after normalizing by the total aircraft coverage, is 0.877. Higher values are desirable, and the maximum score would be 1.0.

5. VALIDATING THE APPLICATION OF A DISTANCE CORRECTION FACTOR

In this section, we present an application of the validation methodology to analyze if the adoption of a distance correction factor is improving the model. This model change was being discussed when the validation framework was being developed. The distance correction factors derive from work developed by the Imperial College London.[7] They are market-based averages that correct for actual distance flown, which

7 Roger Teoh, Zebediah Engberg, Marc Shapiro, Lynette Dray, and Marc Stettler, "A High-Resolution Global Aviation Emissions Inventory Based on ADS-B (GAIA) for 2019–2021," *Atmospheric Chemistry and Physics, 24*, (2024), 725-744. https://doi.org/10.5194/acp-24-725-2024.

is usually higher than the direct distances (great circle distance, or GCD) between the origin and destination due to airspace restrictions, avoidance of bad weather conditions, and airport congestion.

All the error metrics were calculated considering the validation sample, which combines the ANAC data from 2019 and the data privately shared by the partner airlines, aggregated by route-aircraft-carrier groups.

5.1. Average model error and trends

Table 6 compares the average model error of the baseline TIM model (TIM 1.8.0) with the version of the TIM adopting the distance correction. We choose as the "average model error" the median absolute error, considering the absolute fuel burn error for each group of route-aircraft-carrier (error defined in Equation (1)). Table 6 also shows the general behavior and trends of the error distribution, showing the percentage of observations in the validation sample that is being overestimated or underestimated by each model. The color code indicates which model had the better (green) and the worse (yellow) for a specific error metric.

Table 6. Error metrics for baseline and alternative model, considering the application of the distance correction factors

	Baseline (TIM 1.8.0)	TIM 1.8.0 + distance correction
Median absolute error	8.0%	6.6%
Overestimation	24%	31%
Underestimation	76%	69%

We observe that the median absolute error has reduced from 8% to 6.6% with the distance correction application, showing that the general accuracy of the model has increased with this correction factor. The baseline model overestimates fuel burn for 24% of the flights and underestimates for 76% of the flights. The application of the distance correction factors decreases the underestimation trend. With the distance correction, we find that it overestimates fuel burn in 31% of the occurrences. We do not use the overestimation and underestimation values as error metrics as it is not possible to analyze the error magnitude, only the error direction, being unable to define a preferred value for them.

These aggregated metrics are helpful to analyze the models' accuracy (how close the estimates are to real fuel burn) and general trend (over or underestimation), but they provide no information about the models' precision (how close the estimates are to each other). The precision can be evaluated with the distribution of errors, discussed in the following section.

5.2. Distribution of errors

A model change does not necessarily impact the estimates' distribution evenly. Figure 9 compares the frequency distribution of errors calculated for each aircraft-airline-

route group for the baseline model and the model with distance correction. The figure shows that the underestimation trend of the model has reduced with the application of the distance correction, given that the graphic has shifted to the right.

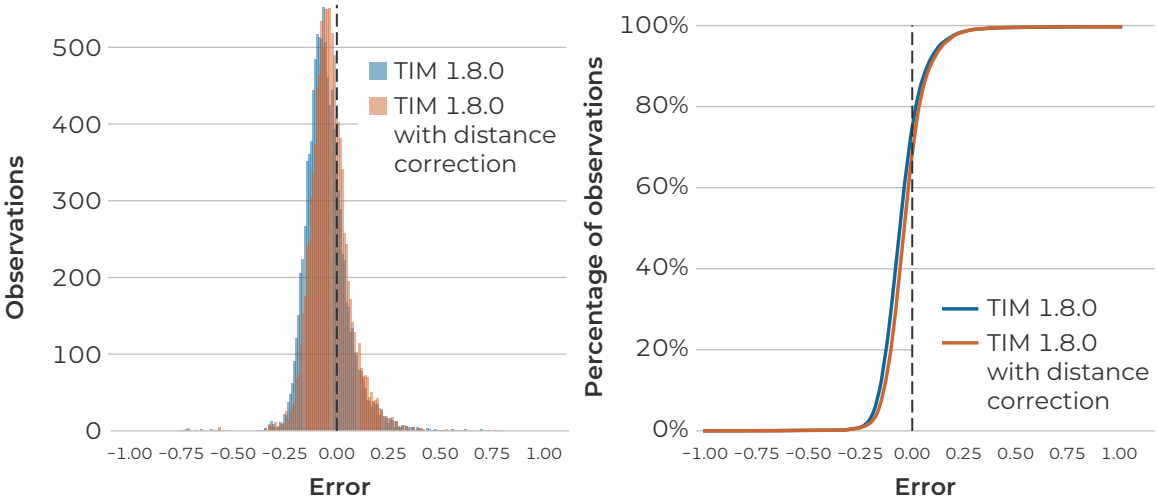

Figure 9. Frequency distributions of errors of the TIM's fuel burn estimates, comparing the baseline model to the model with distance correction

The curve of the model has kept a roughly similar shape, indicating that the precision of the estimates has not significantly changed with the application of this correction.

5.3. Error thresholds

We also analyze how the correction factors application has impacted the error distribution by calculating the percentage of estimates that are below specific error thresholds for each alternative model, as shown in Table 7. We consider 2%, 5%, 10%, and 20% as threshold levels and adopt the same color code used in the previous table: green indicates the model that had the better performance in a given error threshold, and yellow indicates worse performance.

Table 7. Percentage of observations that are below certain error thresholds for the baseline and the alternative model (TIM with the application of distance correction)

		Percentage of observations	
Weights	**Error thresholds**	**Baseline (TIM 1.8.0)**	**TIM 1.8.0 + distance correction**
0.1	<2%	13%	16%
0.2	<5%	31%	39%
0.3	<10%	62%	70%
0.4	<20%	95%	96%
	Weighted average	**63.8%**	**68.8%**

Results show that with the distance correction application, the frequency of estimates within the limit error has increased for all error categories analyzed. The number of estimates that are within 10% of the median fuel burn, for example, has increased from 62% to 70%.

In aggregating the results of the error thresholds, a weighted average is used, as described in Section 4.3 and considering the weights presented in the table. Using the weighted average, we see that the distance correction factor improves the TIM on the error threshold metric.

The visualization of the cumulative distribution of the absolute errors helps to analyze the error thresholds differences between the models. Figure 10 plots the cumulative distribution curves for the model with the distance correction and the baseline. The figure shows that the distance correction provides absolute errors lower than ~7% for ~50% of the observations. The baseline model performs slightly worse for the same error threshold, having ~45% of its observations with errors within ~7%.

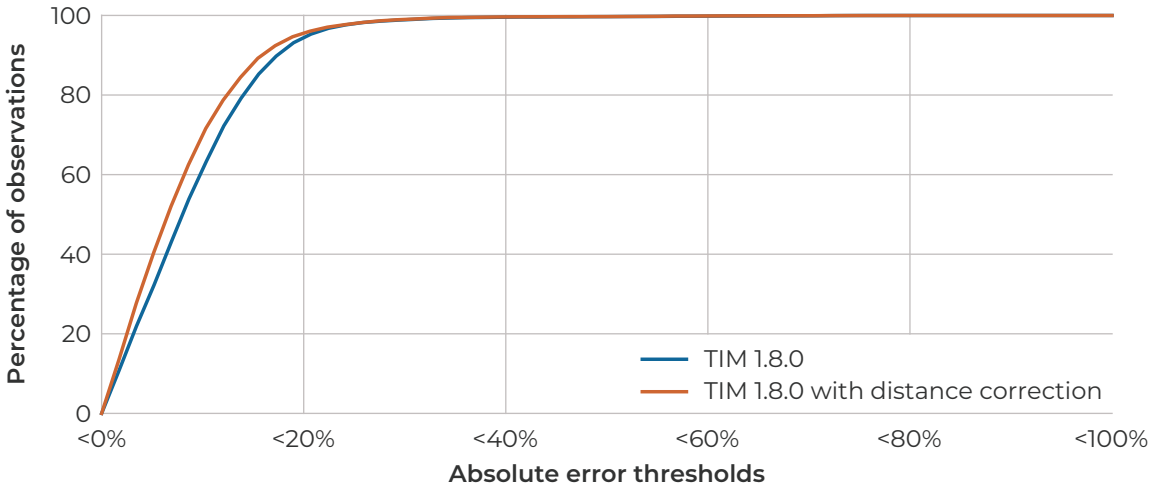

Figure 10. Comparison of the cumulative error distribution of the baseline and alternative models (models with the application of distance correction and fuel burn correction)

5.4. Distance-based metrics

Another important feature to be analyzed is how the errors of the estimates vary with the distance flown. Figure 11 shows the weighted average of the errors for all carrier-aircraft-route groups of the sample considering ANAC 2019 data and private airline data, binned by distance, and weighted by the number of flights. Figure 12 shows the same analysis, but with the absolute errors.

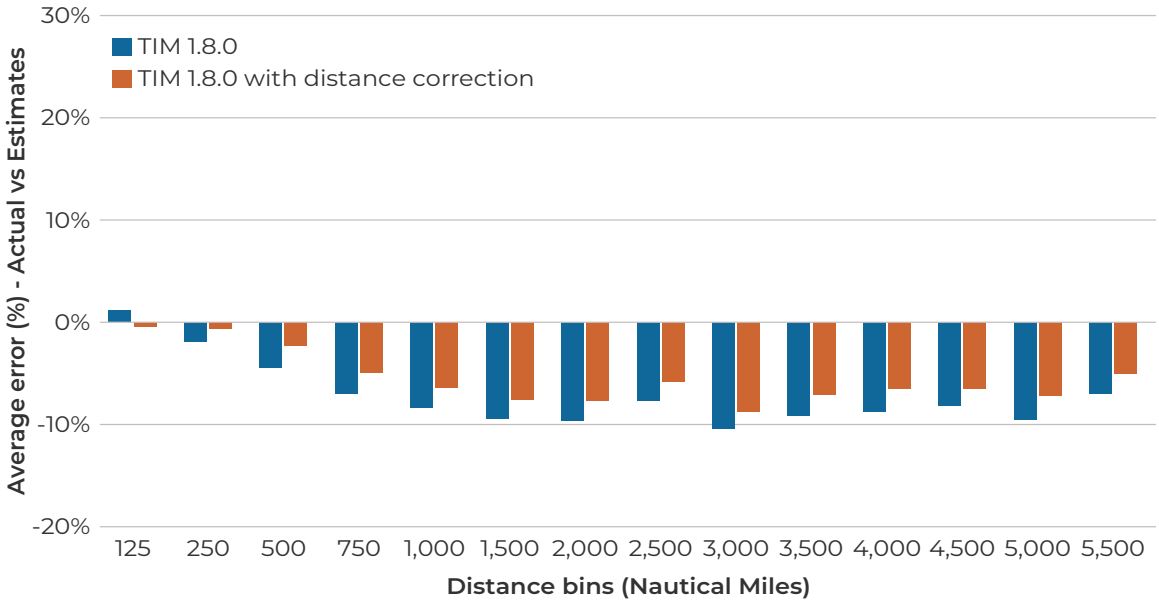

Figure 11. Weighted average error observed for each distance bin, considering the baseline and alternative models (with the application of distance correction and fuel burn correction)

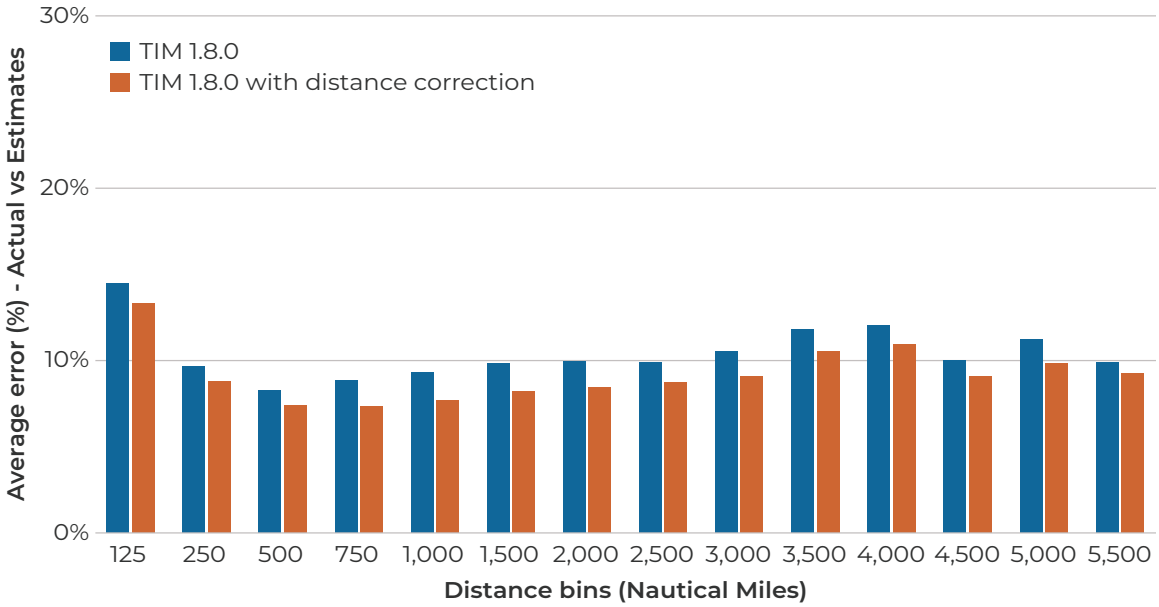

Figure 12. Weighted average absolute error observed for each distance bin, considering the baseline and alternative models (with the application of distance correction and fuel burn correction)

For both models, we observe in Figure 11 a general underestimation trend for almost all distance bins, except for the shortest flights, with distances shorter than 125 NM. The baseline model presents a slight overestimation trend for this distance bin only. As discussed in section 4.4, considering the baseline model, we observed some important discrepancies between fuel burn estimates of ANAC and private airline data for short-haul flights. For ANAC data, the TIM tends to overestimate fuel burn for flights shorter than 500 NM, while we see an underestimation trend for the private airline data in the same distance bins. Further analyses are required to understand these differences and to check if the default LTO cycle adopted by EEA is appropriate for shorter flights, especially in non-European markets. When combining both datasets, the weighted average error for the 125 NM bin is positive but lower than 2%. However, this is a consequence of having negative and positive errors in the sample, as they compensate for each other in the average calculation.

The error magnitude can be better investigated by analyzing the average absolute error by distance, presented in Figure 12. We observe that the median absolute errors have reduced across all distance bins with the distance correction application. The average absolute errors vary between 8% and 14% by distance bin for the baseline model and between 7% and 13% by distance bin for the model with the application of the distance correction factor.

Using the weighted average metrics explained in Section 4.4 to scale for global context, we can compare the performance across the models. Table 8 presents each of the metrics for the two models being considered. Metric values closer to zero are desirable. We observe that the model with the distance correction application performs better for both global frequency and emissions weighted metrics.

Table 8. Weighted average metrics for the distance-based error analysis comparing the baseline model performance with the application of distance correction and fuel burn corrections.

Weighted average metrics		
Weighting	**Baseline (TIM 1.8.0)**	**TIM 1.8.0 + distance correction**
Global flight distribution	9.54%	8.33%
Global emissions distribution	9.79%	8.50%

5.5. Distance- and aircraft-based error analysis

An additional characteristic to be analyzed is how fuel burn varies with distance flown for each aircraft, and how this variation changes with the correction factors adoption. Figure 13 plots the median fuel burn of the ANAC sample as blue dots and the fuel burn estimate for each variation of the TIM model as dashed lines. Each panel represents one of the four most common aircraft models in the ANAC data. For each model, we calculate the R2 value, considering the median fuel burn versus the estimated fuel burn. For a given aircraft model, higher R2 values indicate a better fit of the model to the median fuel burn.

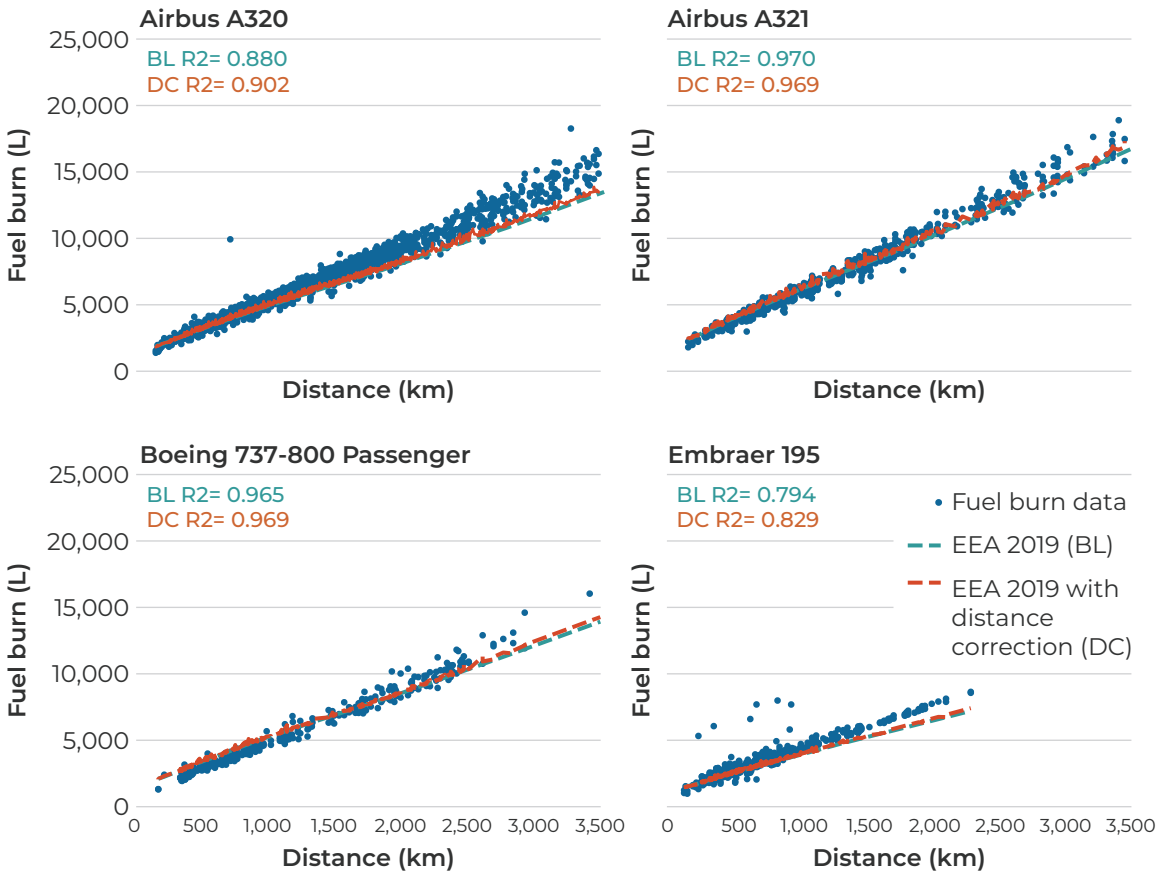

Figure 13. Comparison of the fuel burn versus distance trends for the baseline (BL) and distance correction (DC) models, considering the four most common aircraft in the ANAC database. The actual fuel burn presented is the median of all flights performed for a given distance.

We should remark that the R2-values can be compared across the TIM estimates (baseline vs. alternative models) for a specific aircraft type, given that they are calculated using the same fuel burn distribution. For this same reason, the R2-values of different aircraft cannot be compared across each other. The R2-value will depend highly on the density of observations, which is unique for each aircraft, and on the number of routes operated by each aircraft and the diversity of distances that are represented.

The performance across all the aircraft is aggregated using a weighted average where the weighting represents the percentage of global flights that are flown on the specific aircraft (explained in Section 4.5).

Table 9. Weighted average metric for the aircraft- and distance-based error analysis comparing the baseline model performance with the model with the application of distance correction

Weighting	Baseline (TIM 1.8.0)	TIM 1.8.0 + distance correction
Global flight coverage	0.877	0.886

6. HOW TO DECIDE WHETHER A CHANGE GETS IMPLEMENTED?

Across the different types of error analysis, certain metrics have been calculated to quantify the performance of each model. The individual error analyses may have conflicting trends that need to be aggregated to make an informed decision on whether a change gets implemented or not. We present only the aggregate metrics in Table 10. Each error metric has a different ideal value that is included in the first column. The colors represent the relative performance of the models for each error metric. Green represents the best performing model, and yellow the worst. The color code can be adapted, depending on the number of model versions being analyzed. If there were three models, the baseline and two new features, then green, yellow, and red would be used, yellow representing the midpoint, and red the worst performance.

Table 10. All error metrics for the validation methodology comparing the baseline model performance with the application of distance correction and fuel burn corrections

Error metric	Baseline	Distance correction
Median absolute error **Ideal value: 0%**	8.0%	6.6%
Error threshold analysis **Ideal value: 100%**	63.8%	68.8%
Frequency-weighted distance metric **Ideal value: 0%**	9.54%	8.33%
Emissions-weighted distance metric **Ideal value: 0%**	9.79%	8.50%
Distance and aircraft error metric **Ideal value: 1.0**	0.877	0.886

It is important to note that other than the median absolute error, the other metrics do not have a specific physical interpretation and so the magnitude of the values does not convey any specific information. They should only be used to compare across models to see which model provides a more desirable value. The results show that the distance correction factor does better than the baseline model on all the five metrics.

As has been shown with this validation methodology, there might be a lot of nuance in the way a change to the model will affect the results of the validation process. Although the example presented here shows a model change that has performed better in all the error metrics, this might not be true for a different feature. The validation necessitates a careful consideration of the different error metrics and what each metric represents. It is tempting to calculate a simplistic sum of the relative performance across the aggregated metrics, ascribing a value of +1 to the green cells, and use that sum to decide whether to approve a change.

However, we suggest a more nuanced approach. The process of approving changes should be:

1. The analysts (Secretariat or Google) run the baseline model and the model change through the validation process to get the performance of the models across all the error metrics. They summarize the results of the validation process which includes details of the analysis as shown in Section 5 and the error metrics as shown in Section 4.

2. The group investigating the change (TG or AC) and the analysts meet to discuss the results. Where discussed at the TG level, reach a consensus on a recommendation to the AC.

3. The AC decides by taking into account the results of the validation process and the recommendation of the TG, as appropriate.

7. FINAL CONSIDERATIONS

The proposed validation framework can be applied to analyze any TIM model changes that impact the fuel burn estimation. It allows investigating if a new feature is effectively improving the model and providing estimates that are closer to real operation. In addition, a model change may unevenly impact the estimates. It may improve the model for a given portion of the flights, such as specific routes, or aircraft, but worsen for others. The validation framework helps to characterize this behavior and to identify possible model flaws, which can motivate new improvements.

The framework will keep evolving as new features and alternative models are tested, as well as incorporating more actual fuel burn data to the validation sample to increase its market representativeness. Some opportunities for future improvements:

- **Aggregation of the actual fuel burn data and sample size:** We aggregate all flights of a given route/aircraft/airline/year and define a representative fuel burn value per group to be compared with the TIM estimates. Currently we eliminate all groups with less than 50 flights registered. We are investigating some statistical tests to define an appropriate sample size and avoid eliminating unnecessary groups.

- **Definition of the representative fuel burn value:** To define a representative fuel burn for every group of route/aircraft/airline/year, we suggest adopting the median if possible as it represents the central value of the fuel burn distribution. However, given that some of the airlines that have privately shared the data reported the mean fuel burn of their routes, we adopted the mean for all the different data sources. A future refinement could be requesting airlines to report median instead of mean, if sharing data at the flight level is not possible. Also, if any outlier is biasing the representative value definition for a given group, a mitigation strategy could be to remove flights with extreme fuel burn value (and keep data within a specific interval, such as the 90th percentile, for example).

- **Considering different years of operation:** Currently, the analyses were concentrated in a few years of operation, mainly 2019, 2021, and 2022, avoiding 2020 due to the impact of the COVID-19 pandemic on the aviation sector. As the partnership with airlines increases and more data is available, we suggest testing other years of operations separately and rely on data from recent information as much as possible to reflect the latest airlines' practices. The inclusion of older years would augment the validation sample, which would increase the model coverage and statistical significance. It would also allow testing if year and aircraft age have a strong influence on fuel burn and should be considered as a relevant input to model fuel burn and emissions.
- **Communicating validation results:** The TIM aspires to be the most trusted and transparent emissions model in the industry. Ideally validation results should be shared publicly to demonstrate the accuracy of the model. Every time a model change is implemented, a document summarizing the validation results should be published.

APPENDIX A: OVERVIEW OF BRAZIL'S COMMERCIAL AVIATION MARKET

The analyses presented in this appendix were developed using ANAC Microdata. Our objective is to provide an overview of Brazil's commercial aviation market and to present more details about the sample analyzed in this study. Figure A.1. shows the number of flights performed in Brazil from 2015 to 2022, considering domestic and international flights. For the international segment, the data comprises flights that have departed from or have arrived in Brazil. Depending on the year, 85%–92% of the Brazilian flights were domestic. Figure A.2. presents more details about these segments, now showing the share of Brazilian and foreign airlines by route type (domestic or international).

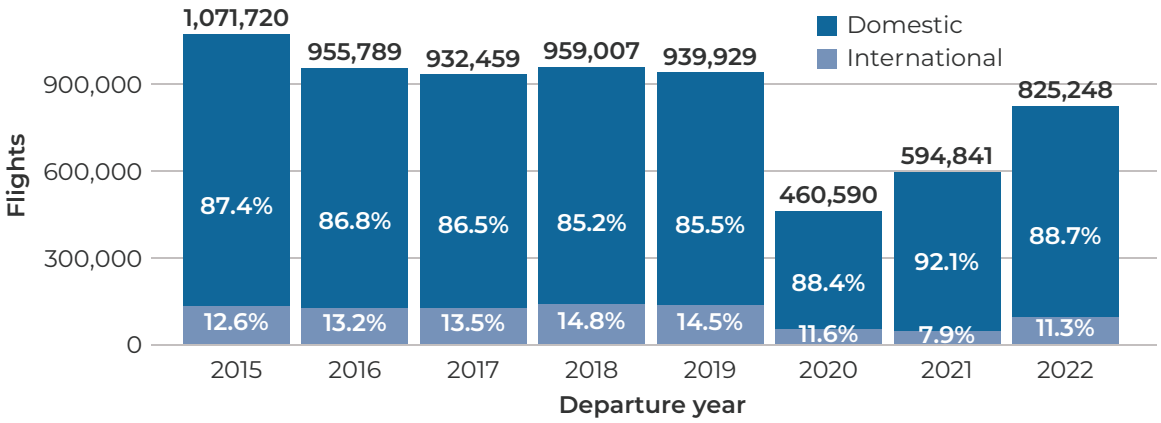

Figure A.1. Number of flights performed in Brazil from 2015 to 2022 by route type (domestic or international)

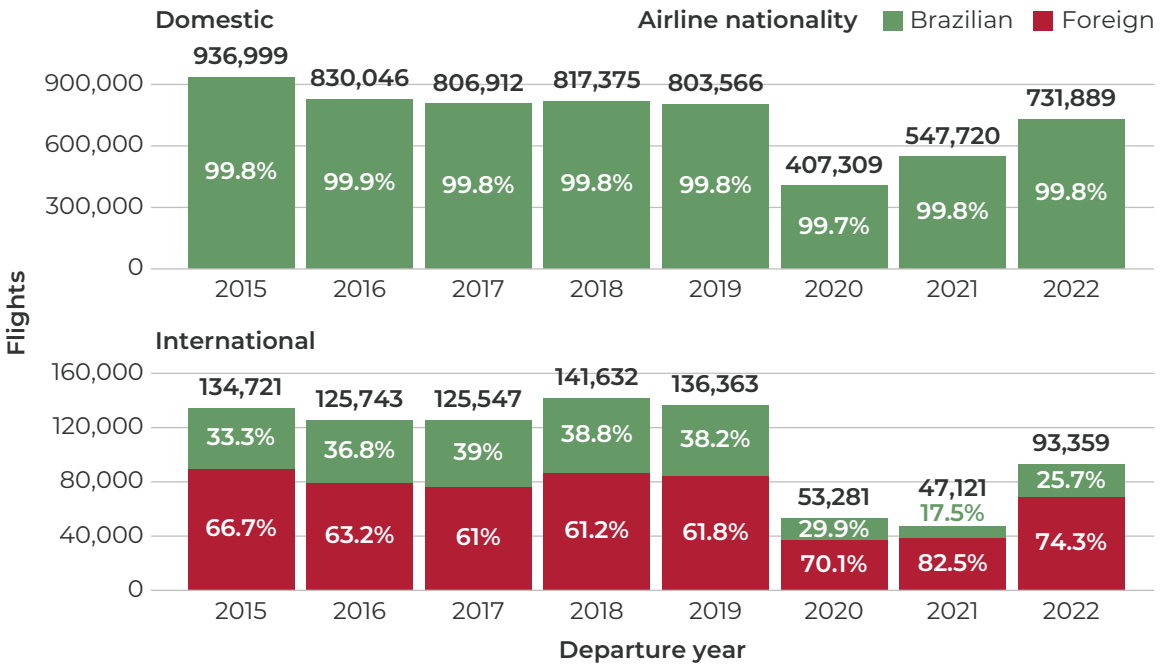

Figure A.2. Number of flights in Brazil from 2015 to 2022, segmented by route type (domestic or international) and airline nationality (Brazilian or Foreign)

Domestic flights are almost all performed by Brazilian airlines, while the larger share of international flights is operated by foreign carriers. The fuel burn data is only available for flights performed by Brazilian airlines, which means there is fuel information for basically all domestic flights, but for a smaller share of the international segment. Flights performed by foreign airlines are eliminated from our sample.

Considering only flights performed by Brazilian airlines, Figure A.3 presents the share of service type (dedicated cargo or passenger) by route type (domestic or international). Usually, more than 96% of domestic and international flights are passenger flights, except international flights from 2020 until 2022, probably impacted by the COVID-19 pandemic. Given that TIM does not include cargo in its modeling process, we eliminate cargo flights from our analyzed sample.

Figure A.3. Number of flights in Brazil from 2015 to 2022, considering only Brazilian airlines, by service type (dedicated cargo or passenger)

Our final sample consists of domestic and international passenger flights, performed by Brazilian airlines only. For testing the validation process, we selected the 2019 data to consider a scenario not impacted by the pandemic.

Another important information is the market share of airlines for domestic and international flights in Brazil. Table A.1 presents the 5 airlines with the highest number of flights and of passengers transported for domestic passenger flights in 2019. We see that the Brazilian domestic flights market is very concentrated in a few airlines, with the largest 3 (Azul, Gol, and Latam) being responsible for 93.7% of flights and 95.9% of passengers transported that year. The fourth largest airline in 2019, Avianca, declared bankruptcy in 2020 but had its operations suspended in May 2019.

Table A.1. Airline market share for domestic passenger flights in Brazil in 2019

		Frequency		Passengers	
Airline (ICAO)	**Airline**	**Number of flights**	**Ranking (market share)**	**Number of passengers**	**Ranking (market share)**
AZU	Azul	278,847	1 (35.3%)	26,912,180	3 (27.4%)
GLO	Gol	242,316	2 (30.7%)	35,268,116	1 (35.9%)
TAM	Latam	218,826	3 (27.7%)	31,955,581	2 (32.6%)
ONE	Avianca	25,105	4 (3.2%)	3,204,950	4 (3.3%)
PTB	Passaredo	12,457	5 (1.6%)	546,810	5 (0.6%)
Others	Others	11,521	Others (1.5%)	250,915	Others (0.3%)
		789,072		**98,138,552**	

For international flights, the market is less concentrated. Table A.2 shows the 7 airlines with the highest number of flights and of passengers transported for the international segment in Brazil in 2019. Fuel burn information is only available for Brazilian airlines, highlighted in the table. For this reason, we eliminate foreign airlines from our sample. Service type information (passenger or dedicated cargo flight) is also only available for Brazilian airlines.

Table A.2. Airline market share for international flights in Brazil in 2019

			Frequency		Passengers	
Airline (ICAO)	**Airline**	**Service type**	**Number of flights**	**Ranking (market share)**	**Number of passengers**	**Ranking (market share)**
TAM	Latam	passenger	26,332	1 (19.3%)	5,401,919	1 (22.1%)
GLO	Gol	passenger	15,641	2 (11.5%)	2,110,612	2 (8.6%)
ARG	Aerolineas Argentinas	not identified	8,078	3 (5.9%)	1,027,933	7 (4.2%)
CMP	Copa Airlines	not identified	7,880	4 (5.8%)	1,041,709	6 (4.3%)
TAP	TAP	not identified	7,830	5 (5.7%)	1,810,826	3 (7.4%)
AZU	Azul	passenger	7,598	6 (5.6%)	1,382,789	4 (5.7%)
AAL	American Airlines	not identified	5,517	7 (4.0%)	1,301,926	5 (5.3%)
Others	Others	-	57,487	Others (42.2%)	10,355,653	-
			136,363		**24,433,367**	

We also present the aircraft share of our sample. Table A.3 shows the 8 most common aircraft models for domestic flights in Brazil in 2019 and which airlines operate each one of them. Table A.4 shows equivalent information for international flights, considering only Brazilian airlines.

Table A.3. Aircraft share for domestic passenger flights in Brazil in 2019 (8 most common aircraft models used to operate this segment)

Aircraft (ICAO)	Number of flights	Share (%)	Airlines
B738	192,042	24.3%	GLO
E195	136,574	17.3%	AZU
A320	129,696	16.4%	TAM, ONE, AZU
A319	57,949	7.3%	ONE, TAM
A20N	54,215	6.9%	TAM, AZU
A321	52,410	6.6%	TAM
AT72	45,287	5.7%	PAM, AZU, PTB
B737	31,529	4.0%	GLO
Others	89,370	11.3%	
	789,072		

Table A.4. Aircraft share for international passenger flights in Brazil in 2019, considering only Brazilian airlines (8 most common aircraft models used to operate this segment)

Aircraft (ICAO)	Number of flights	Share (%)	Airlines
B738	13,899	27.5%	GLO
A320	9,684	19.2%	TAM, ONE, AZU
B763	7,069	14.0%	TAM
B77W	3,597	7.1%	TAM
A321	3,350	6.6%	TAM
A359	2,863	5.7%	TAM
A20N	2,395	4.7%	AZU
A332	2,250	4.5%	AZU
Others	5,347	10.6%	
	50,454		