On the Causes and Environmental Impact of Airborne Holdings at Major European Airports

Ramon Dalmau,* Philippe Very,1 and Gabriel Jarry,1
EUROCONTROL, Brétigny-Sur-Orge, France
*Corresponding author: ramon.dalmau-codina@eurocontrol.int

(Rceived 21 October 2023; revised ; accepted ; first published online )
(Editor: Tatiana Polishchuk; open reviewed by: )

Abstract
This paper introduces a data-driven technique for labelling airborne holdings based on their underlying causes, specifically distinguishing between adverse weather conditions and other causes, such as airport capacity. Utilising a dataset comprised of flight trajectories arriving at 45 European airports over a nine-month period, extracted from automatic dependent surveillance-broadcast data, this paper provides valuable insights into the causes behind airborne holdings and their relative environmental impact. The proposed approach involves employing an existing neural network to identify airborne holdings. Subsequently, these holdings are cross-referenced with actual weather observations obtained from meteorological aerodrome reports. Following this, a subset of the holdings is labelled as either weather-related or attributed to other causes, based on historical air traffic flow management regulations. Finally, the cause of the majority of unlabelled holdings is determined using semi-supervised learning. The findings indicate that at least one-quarter of the 30-minute time periods with airborne holdings identified by the neural network can be attributed to weather-related factors, with reduced visibility, strong winds, and convective weather, emerging as the primary contributing events. Intriguingly, weather-related causes account for approximately 40% of the total fuel consumption associated with these procedures.

Keywords: airborne holdings; environmental impact; adverse weather

1. Introduction

Arriving flights frequently encounter tactical control strategies to ensure safety. These control strategies involve level-offs, path stretching, and holding patterns, all of which can decrease flight efficiency [1]. A recent analysis by [2] examined two months of automatic dependent surveillance-broadcast (ADS-B) data for aircraft landing at five major European airports. The study revealed that holding patterns had the most significant adverse environmental impact, regardless of their cause. Based on these findings, this paper paves the way for a more in-depth investigation of airborne holdings at major European airports, facilitating an assessment of the relative environmental impact of various causes, specifically distinguishing between adverse weather conditions and other factors. By gaining insights into the primary drivers of these tactical control strategies, this study aims to equip the aviation community with the knowledge needed to facilitate the implementation of targeted measures aimed at mitigating both their environmental and economic consequences. This assessment, however, encounters several challenges, namely (1) the need to detect airborne holdings from flight trajectories and (2) the current lack of data providing information about their causes.
It is critical to emphasise that the primary goal of this paper does not revolve around the quantitative results. The primary focus of this paper is on the semi-supervised methodology, which has broad applicability to other tactical control strategies such as path stretching and level-offs.

After conducting a literature review in Section 2, Section 3 details the methodology employed to address the two aforementioned challenges. The setup of the experiment that showcases the effectiveness of the methodology is presented in Section 4. Section 5 presents the primary findings of the experiment, while Section 6 concludes the paper with key remarks and take-home messages.

2. Literature review

Recent research has underscored the substantial connection between aviation and environmental concerns. For instance, [3] highlighted that commercial aviation was accountable for nearly 818 megatons of CO2 emissions in 2018, suggesting a potential link with the economic prosperity of nations. This sentiment was further echoed by [4], who used ADS-B data and the open aircraft performance (OpenAP) framework to examine the environmental impact of aviation across Europe.

The terminal manoeuvring area (TMA) is where many of the environmental impact of aviation occurs. As an example, [5] investigated the environmental impact of air traffic congestion during peak hours at London Heathrow Airport, highlighting the contribution of holdings. More recently, [2] expanded upon this insight, examining environmental inefficiencies in arrival procedures and emphasised the detrimental environmental impact of holdings compared to other procedures. Simultaneously, [1] introduced a comprehensive set of valuable flight efficiency indicators for arrivals.

Weather remains a substantial factor affecting flight efficiency in the TMA, with adverse weather conditions often resulting in reduced airport capacity that triggers holdings or even diversions. The latter issue was addressed in [6], which introduced a tree-based model designed to predict diversions caused by adverse weather conditions. In a follow-up study [7], supervised clustering was used to categorise the causes behind these diversions, identifying events such as low visibility or snow.

Whether for evaluating the environmental impact or developing models to predict and mitigate these events, dedicated algorithms are essential for detecting them from surveillance data. [8] explored rule-based and statistical methods for detecting various events. Notably, some of these methodologies, such as a neural network for detecting holdings, have been integrated into traffic [9].

3. Method

The method starts by detecting holdings patterns from ADS-B trajectories with the neural network integrated into traffic [9]. These holdings are then grouped in 30-minute intervals, which is the typical frequency of weather updates at major airports, and enriched with weather observations from the closest meteorological aerodrome report (METAR). This enables the creation of an unlabelled dataset, with each observation corresponding to a 30-minute period in which at least one holding was identified, and where the various features represent the observed weather conditions.

Each observation is then assigned the label "weather" or "other" based on the cause of the air traffic flow management (ATFM) regulation in effect at the airport at that time (if any). Observations with no concurrent ATFM regulation at the airport remain unlabelled and are addressed in the next step.

Figure 1 provides examples of detected holding patterns at Zurich airport during three different days at the same hour. Figure 1a represents observations labelled as weather, Figure 1c depicts observations labelled as other, and Figure 1b represents scenarios where no holdings were present.
An ATFM regulation due to weather was present and the weather at 10:20 AM was VRB01KT 0200 FG VV001.

An ATFM regulation due to ATC capacity was present and the weather at 10:20 AM was 4000 -SHRA FEW004 SCT010.

No ATFM regulation was present and the weather at 10:20 AM was 18007KT 9999 FEW050.

Figure 1. Highlighted in blue are the arrival trajectories near Zurich Airport, covering a 50-nautical mile radius, between 10:00 AM and 11:00 AM. The segments of these trajectories marked in red represent holding patterns identified by the neural network implemented in the traffic library.

The cause of the unlabelled observations is estimated by using a simple yet effective self-training algorithm [10], which allows any base_classifier (e.g., a decision tree or a neural network) to learn from unlabelled data. The steps of the self-training algorithm are listed in Algorithm 1. The base_classifier is responsible for predicting labels for unlabelled observations during each iteration. Then, a portion of these observations may be transferred into the pseudo-labelled set. The selection of candidates can be accomplished through two methods: either by applying a threshold to the predicted probabilities or by selecting the k_best observations with the highest predicted probabilities. The reader is referred to the scikit-learn documentation for further details.

Algorithm 1 Self-training

Require: base_classifier, threshold (or k_best), max_iter, labelled set, unlabelled set

Initialise the empty set of pseudo-labelled observations

repeat

Fit the base_classifier using labelled and pseudo-labelled sets

Predict label probabilities for all observations in the unlabelled set

if threshold is used then

Unlabelled observations with predicted probabilities exceeding the threshold parameter are assigned that label and transferred to the pseudo-labelled set

else

(k_best is used) Transfer the k_best unlabelled observations with the highest predicted probabilities to the pseudo-labelled set, assigning them the corresponding most probable labels

end if

until No additional observations are added to the pseudo-labelled set, all unlabelled observations have been labelled, or after completing max_iter iterations

In practice, this procedure allows the majority of observations to be labelled as "weather" or "other," allowing the expected proportion of each cause to be determined. Some labels are given, while others are inferred by the model (the pseudo-labels). The fuel consumption associated with each holding can then be calculated using the OpenAP framework and attributed to the corresponding cause.

4. Experiment

The experimental setup for this study involved using ADS-B data for arrivals at the top 45 busiest airports in Europe during 2022, as determined by Wikipedia. This dataset covers the period from January 1st, 2022, to June 1st, 2023. metafora\(^2\) was used to extract the weather conditions from the raw METARs for the same airports and period. Table 1 provides an overview of the dataset.

<table>
<thead>
<tr>
<th>Name</th>
<th>Mean</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Name</th>
<th>Proportion of falses</th>
<th>Class</th>
<th>Occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>speed (m/s)</td>
<td>4.0</td>
<td>2.1</td>
<td>3.6</td>
<td>5.1</td>
<td>precipitation</td>
<td>0.87</td>
<td>Unlabelled</td>
<td>41620 (85%)</td>
</tr>
<tr>
<td>gust (m/s)</td>
<td>0.6</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>obscuration</td>
<td>0.94</td>
<td>Weather</td>
<td>3484 (7%)</td>
</tr>
<tr>
<td>visibility (m)</td>
<td>924</td>
<td>999</td>
<td>999</td>
<td>999</td>
<td>thunderstorms</td>
<td>0.98</td>
<td>Other</td>
<td>4051 (8%)</td>
</tr>
<tr>
<td>ceiling (m)</td>
<td>2252</td>
<td>1067</td>
<td>3048</td>
<td>3048</td>
<td>snow</td>
<td>0.99</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cover (oktas)</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>6</td>
<td>clouds</td>
<td>0.92</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The dataset comprises 41620 (85%) unlabelled observations, 3484 (7%) weather-related observations, and 4051 (8%) observations attributed to other causes. Remember that each observation represents a 30-minute period during which at least one holding was identified. To evaluate the model’s performance on unseen data, a subset of 10% randomly selected and labelled observations was set aside.

The experiment made use of two base classifiers to (1) cross-check the results and (2) demonstrate that even simple models can perform this task comparably to more sophisticated ones. The simple model is a DecisionTreeClassifier with max_depth set to 10 and min_samples_leaf set to 25. The complex model is a LGBMClassifier composed of 60 decision trees with the same hyper-parameters as the simple model but trained sequentially with gradient-boosting. The hyper-parameters listed above were chosen so as to prevent over-fitting while still ensuring adequate learning power. Furthermore, monotone constraints were applied to the LGBMClassifier to achieve consistent feature attribution. These constraints enforce that, all else being equal, higher values of wind speed, gust, sky cover, precipitation, obscuration, thunderstorms, snow, and presence of clouds increase the likelihood of an observation being classified as weather (the positive class). Lower visibility and ceiling values, on the other hand, must increase the likelihood of an observation being classified as weather.

In the context of the self-training algorithm (see Algorithm 1), the threshold selection criteria was adopted, with the threshold parameter set at 0.75, and the number of iterations was not limited.

Further investigation included computing Shapley values for the observations labelled as weather (both given and pseudo-labelled), which provided insights into the impact of the various weather events (e.g., obscuration, snow, thunderstorms). Then, the Birch clustering algorithm was applied to the Shapley values to group observations with comparable characteristics. Regarding the configuration of Birch, a threshold parameter of 0.25 was chosen and the development of 6 clusters was enforced, guided by a visual inspection of the data. Additionally, dimensionality reduction was carried out via principal component analysis (PCA) in order to facilitate the interpretation of results.

Lastly, environmental impact was quantified through fuel consumption estimation using the OpenAP framework, providing valuable insights into the environmental consequences associated with primary causes (i.e., weather and other) and weather events (e.g., obscuration, snow, thunderstorms).

\(^2\)https://github.com/ramondalmau/metafora
5. Results

This section presents the key findings of the experiment. Section 5.1 illustrates the distribution of observations by cause (weather or other) obtained through semi-supervised training. Section 5.2 delves into the specific weather events that likely prompted each observation labelled as weather.

5.1 Proportion of observations per cause

Figure 2 illustrates the evolution of the proportion of labels (weather, other, and unlabelled) as a function of the self-training iteration. Note that both given labels and pseudo-labels are considered.

As shown in Figure 2, DT and GBDT successfully labelled approximately 80% and 90% of the observations, respectively. The observations that remain unlabelled are observations for which the probability of being caused by adverse weather is higher than 25% but lower than 75%, indicating cases that cannot be attributed to a cause with sufficient confidence to satisfy the threshold criteria. It is worth noting that GBDT required twice the number of iterations compared to DT to complete the self-training process.

Table 2 shows the label occurrence on the entire dataset (including the 10% of labelled observations reserved for assessing the performance of the models on unseen data), after the self-training process.

<table>
<thead>
<tr>
<th>Model</th>
<th>Unlabelled</th>
<th>Weather</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>DT</td>
<td>9825 (20%)</td>
<td>10871 (22%)</td>
<td>28456 (58%)</td>
</tr>
<tr>
<td>GBDT</td>
<td>4708 (9%)</td>
<td>13618 (28%)</td>
<td>30829 (63%)</td>
</tr>
</tbody>
</table>

As indicated by the data in Table 2, both models allocate a comparable proportion of observations to different causes. Approximately two-thirds are ascribed to factors other than weather, one-quarter to weather-related factors, and the remaining observations are not classified due to insufficient confidence.

This categorisation is only valid, however, if the models have effectively learned the patterns that certainly lead to weather-related airborne holdings. In order to check that this condition is met, two steps will be taken: (1) measuring binary classification metrics on the 10% of reserved observations, and (2) computing the Shapley values of the model to investigate the attribution given to the features.

To start with, Table 3 presents the classification metrics on the 10% of labelled observations randomly sampled (with stratification) from the dataset before starting the self-training process.
Table 3. Classification metrics on the 10% of labelled observations randomly sampled with stratification from the dataset.

<table>
<thead>
<tr>
<th>Model</th>
<th>Weather (349 obs.)</th>
<th>Other (405 obs.)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>DT</td>
<td>0.82</td>
<td>0.77</td>
</tr>
<tr>
<td>GBDT</td>
<td>0.83</td>
<td>0.80</td>
</tr>
</tbody>
</table>

The metrics shown in Table 3 reveal comparable performance between the DT and GBDT models, with both excelling at predicting the cause for airborne holdings. Their accuracy is higher than 80%, and the precision and recall on the two categories are alike. Outstanding results are also observed for the average precision and the area under the receiver operator characteristic curve (ROC AUC).

Figure 3 shows the distribution of Shapley values for the two models. In this graph, the y-axis indicates the name of the features, in order of mean absolute Shapley value from the top to the bottom. Each dot in the x-axis shows the Shapley value of the associated feature on the prediction for one observation, and the colour indicates the magnitude of that feature: red indicates high, while blue indicates low. A positive Shapley value indicates that the feature contributes to the prediction for the observation by increasing the probability of weather-related airborne holding (i.e., the positive class) relative to the expected value in the train set, while a negative value indicates the opposite.

Figure 3 shows that the models have learned patterns that correspond to human intuition. Notably, visibility emerges as the most important feature, with higher values implying that the observed holdings are less likely to be driven by weather conditions, while lower visibility values indicate that the likelihood of weather-related holdings increases. Common sense also extends to the other features, where the presence of precipitation, obscuration, snow, and/or thunderstorms positively influences the model’s output. Figure 3 also showcases the importance of enforcing monotone constraints.

In contrast to the GBDT model, the DT model does not always adhere to these constraints. For instance, it exhibits occasional instances where strong wind gusts contribute to a decrease in the probability of weather-related holding. However, it’s worth noting that such undesirable behaviour, stemming from minor over-fitting to noise in the data, occurs infrequently.

It is important to remark that, in the context of self-training classification tasks employing a threshold criterion, the use of a well-calibrated classifier is imperative. The calibration curves of the DT and GBDT models after the last self-training iteration are shown in Figures 4a and 4b, respectively.
Figures 4a and 4b indicate that the models are relatively well calibrated, despite the fact that the separation from the perfectly calibrated line shows a pessimistic tendency to over-forecast low probabilities of weather-related holdings (i.e., the positive class).

Finally, Figure 5 shows the proportion of fuel consumption attributed to the various causes, demonstrating the alignment of outcomes between the two models, with the proportion of fuel consumption attributed to unknown causes closely resembling the frequency of these events. Interestingly, both models attributed approximately 40% of fuel consumption to weather-related causes, despite this class representing only a quarter of the observations. This intriguing finding suggests that while weather-related holdings occur less frequently, they have a higher impact on fuel consumption. In other words, periods with holdings caused by weather tend to be more severe.

Figure 5. Proportion of fuel consumption in airborne holdings attributed to the various causes.

5.2 Proportion of observations per weather cluster

Figure 6 shows the projection, into two components, of the Shapley values computed for the observations labelled as weather (either given labels or pseudo-labels after the self-training process) as a result of the PCA algorithm. Each point corresponds to one of these observations, and the colour indicates the cluster as detected by the Birch algorithm, which humanised identifiers, like "snow" instead of just "0", were based on the manual inspection of the feature distributions that will follow. Please keep in mind that the cluster names merely reflect the most significant weather event, but ultimately, a holding may be prompted by a combination of multiple weather events.
According to Figure 6, the PCA projection into two dimensions captures a large variance of the data. Generally speaking, the six clusters are well-separated, with similar or concurrent weather events appearing closely grouped in the lower-dimensional space. For instance, thunderstorms are in between clouds, obscuration and speed, while obscuration slightly overlaps with snow.

It should be noted that the cluster named "other" includes most of the observations labelled as "weather" but which predicted probability of belonging to that category is very low according to the model. To elaborate further, these are instances where holdings were observed during a weather-related ATFM regulation, but the weather conditions may not be exceptionally severe.

Figures 7 and 8 show the empirical cumulative distribution and the normalised histogram per cluster, respectively, for the DT model that were taken into account during the identification process. The equivalent graphs for the GBDT model are shown in Figures 9 and 10, respectively.
Figure 8. Normalised histogram for the DT model. Colours follow the same notation as in Fig. 6a.

Figure 9. Empirical cumulative distribution function for the GBDT model. Colours follow the same notation as in Fig. 6b.

Figure 10. Normalised histogram for the GBDT model. Colours follow the same notation as in Fig. 6b.

Figure 11 shows the Sankey diagram representing the relationships between observations within the 6 clusters generated from the Shapley values of the GBDT model (left) and their connections to either the same or different clusters based on the Shapley values of the DT model (right).
Figure 11 reveals that about two-thirds of the observations originally allocated to the obscuration cluster in the left categorisation are correlated with ceiling in the right categorisation. This flow of data is caused by the manner in which metafora encodes periods with vertical visibility information (VV), which often occurs during low-visibility times. In these circumstances, the vertical visibility is used to fill the ceiling feature, yet actually it reflects the vertical visibility. In contrast, around one-third of the observations originally classified as speed in the left clusters are now classified as ceiling in the right clusters. This flurry of observations is caused by the fact that severe winds frequently occur in the midst of storms, when the ceiling is low. Overall, it is worth noting that the two models have produced similar outcomes.

Lastly, Figure 12 depicts the distribution of fuel consumption in airborne holdings attributed to various weather events, each of which is associated with a specific cluster.

Figure 12 shows that, in both clusters generated by the DT and GBDT models, obscuration and ceiling together appear as the major contributors to fuel consumption. This fact is not surprising, as most of the observations belong to these clusters. Wind speed and the presence of convective weather also appear to have an important environmental impact, despite the presence of these occurrences is not as significant.
6. Conclusions

Leveraging the power of semi-supervised learning, this study revealed that approximately 25% of the 30-minute time periods with airborne holdings in Europe (independently of the severity) are weather-related, with reduced visibility, winds and convective weather as the main contributors, constituting around 40% of total fuel consumption during these procedures.

It is critical to emphasise that airborne holdings are just one of the many factors influencing flight efficiency within the TMA. This paper primarily focused on the methodology, which is why it specifically addressed one particular tactical control strategy for illustration purposes. However, it is important to note that other tactical control strategies, such as path stretching or level-offs, also have a significant impact and cannot be ignored [1]. Therefore, we strongly encourage the research community to explore the potential of extending the method proposed in this study to comprehensively unravel the causes of flight inefficiencies within the TMA in a more generalised manner.

Furthermore, it is worthwhile to investigate the various practises regarding the utilisation of airborne holdings at various airports. While some airports use holding patterns primarily in response to extreme weather conditions, as a safety measure to ensure the orderly and safe flow of air traffic during adverse conditions, others take a more strategic approach. Airborne holdings are not just a backup plan for bad weather at these airports; they are a deliberate strategy used during peak hours of air traffic congestion. Airborne holdings are useful in these situations for orchestrating the complex ballet of incoming and outgoing flights. This intriguing study, however, falls outside the scope of this paper, as our primary focus has consistently been on detailing the methodology for identifying the causes of airborne holdings in a semi-supervised fashion.

Finally, in order to improve reproducibility within the research community, this study used the OpenAP framework for fuel consumption computation. Nonetheless, we strongly advise using the base of aircraft data (BADA) performance model to achieve more accurate estimates.

Open data statement

The data that support the findings of this study are available at: https://zenodo.org/doi/10.5281/zenodo.10032729

Reproducibility statement

This work can be reproduced using the code available at: https://www.github.com/ramondalmau/holdings-opensky. Accessed on October 12th, 2023.

Reproducibility statement

- **R.D**: Conceptualization, Data Curation, Formal Analysis, Methodology, Software, Visualization, Writing – Original Draft;
- **P.V**: Conceptualization, Formal Analysis, Methodology, Supervision, Validation, Visualization, Writing – Review & Editing;
- **G.B**: Conceptualization, Formal Analysis, Methodology, Supervision, Validation, Visualization, Writing – Review & Editing.
References


