Journal of Open Aviation Science (2023), Vol.**1**
doi:10.59490/joas.2023.7358

JOAS

EDITORIAL

Reviews and responses for

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

Authors: Ramon Dalmau, Philippe Very, and Gabriel Jarry

Reviewers: Antonio Franco, Max Li, and Enrico Spinielli

Editor: Editor Name

1. Original paper

DOI for the original paper: https://doi.org/10.59490/joas.2023.7198

2. Review - round 1

2.1 Reviewer 1

In this paper, authors present a data-driven technique to assign an underlying cause to airborne holdings, with a specific emphasis on distinguishing between adverse weather conditions and other causes. To do so, they first identify holdings by applying a neural network-based functionality of an existing open-source tool. Then, holdings are clustered into thirty-minute intervals and each cluster is assigned a weather observation from the METAR of the closest aerodrome. This way, each observation of the dataset is a half-an-hour interval when a holding took place and contains the weather conditions. The next step authors perform is the cause labelling of observations. Specifically, if the closest airport to an observation had an ATFM regulation in effect, the observation is labeled as "weather" or "other" based on the said ATFM regulation. The cause of the unlabelled holdings (those without an ATFM regulation in place) is determined by applying a semi-supervised learning approach.

The references cited are pertinent and up-to-date, but quite scarce for a journal paper. The methodology is sound, although a bit shallow for a journal paper. The description of the experiment is concise, since deeper details are provided in the following section addressing results. As for the results, the authors do not settle for just presenting the proportion of holdings due to weather and their impact on fuel consumption, but they provide an interesting and thorough analysis including a validation of the outcoming model and an explanation of the effects of the different features considered. Finally, conclusions and final remarks are sensible.

Notwithstanding the above, there are a couple of comments the authors must address:

1) On the one hand, the reader would appreciate for authors to elaborate more on the performance evaluation based on the 10% labelled observations set aside. Specifically, the possibility for the trained model to provide unlabelled results raises some questions unanswered. If those validation observations are classified with the trained model, then one would expect a non-negligible percentage of unlabelled results, which is not provided, though. Furthermore, when classification may lead

to unlabelled results, the definition of classification metrics (accuracy, precision, recall ...) is less obvious, so they should be explicated.

2) On the other hand, and less importantly, on page 10, lines 196-197, the authors state "around one-third of the observations originally classified as speed in the left clusters are now classified as ceiling in the right clusters.". This assertion is not correct; in fact, I would say this is the other way around: around one-third of the observations originally classified as speed in the RIGHT clusters are now classified as ceiling in the LEFT clusters.

All in all, my recommendation is to accept the paper subject to minor revisions (those needed to address the previous comments).

2.2 Reviewer 2

The authors propose an application of deep learning for classifying the weather-related causes of airborne holding at European airports. Additionally, the authors conduct a Shapley value-based analysis of feature importance, plus an examination of the per-holding emissions impact, differentiating between the emissions impact due to weather-related airborne holding, versus other causes. Although the application of deep learning and the deep learning architecture is straightforward, I enjoyed reading the authors' discussion of the importance of dissecting airborne holdings in a more systematic manner. I have several comments/suggestions that I would like to see the authors address, which I detail below:

1. In Figure 1(b), there seems to be some irregular holding patterns to the north of Zurich. These could have just been extra vectors given for perhaps sequencing reasons, but they still would induce some non-negligible amount of airborne delay. I would suggest that the authors clarify if their approach is only to identify standardized holding patterns, e.g., such as those depicted in standard arrival routes, or if it's to detect any circuitous patterns that result in airborne delays.

2. In terms of the features that the authors used, they include mostly features related to weather/convective conditions. It would be great if the authors could comment/discuss the possibility of integrating additional data sets, such as those describing active ATFM measures and/or data sets that contain airline schedule information. The reason being that those two data sets should be relatively observable, and may contribute to airborne holding occurrences as well. Even better might be a way to somehow incorporate features that describe on-airport emergency situations, e.g., an emergency aircraft on the runway – those can be a common reason for go arounds and holding. However, I would assume that these happen much more rarely, and may be difficult to get comprehensive data on.

3. A minor note, but I would suggest that the authors normalize the x-axis in Figure 2 such that both go up to 12 iterations. This will make it very apparent that GBDT requires 2x the iterations as DT to achieve similar levels of label proportions.

4. The Shapley value-based analysis of the contributions per feature is certainly helpful in interpreting the results of the classifier. I also would be interested in seeing the set of (arrival) airports at which these classifications were made, specifically the ILS categories of the dominant arrival runways/arrival runway configurations active at each airport. I would assume that there might be a relationship between the approach categories (e.g., Cat III vs. Cat II) and ILS equipment, and the chances of an aircraft having to hold in order to wait for weather minimums to improve.

5. The authors emphasize in several places that this methodology can be applied more broadly than just for airborne holding, which I agree with. I would suggest that the authors add to their discussions/conclusion, a bit more about how they would guide future researchers to apply their method to the case of path stretching or non-continuous descent approach procedures. Of course, not all of the details need to be fleshed out, but a cursory discussion on what might be the important

features involved, how might one set up the neural network architecture/perform the parameter tuning procedures, etc. – these would be great extended discussions to have.

2.3 Reviewer 3

The paper is clear and interesting. A few observations worth considering by the authors:

1. Lines 88-93: what led to the decision to use the specific parameters used? Why max_depth 10, min_samples_leaf 25, 60 decision trees? The text mentions (Lines 92-93) that the hyper-parameters were chosen to prevent over-fitting: the question is what was the method chosen to choose them?

2. Lines 99-100 and the rest of section 4: similar to the item above what were the criteria, methods used to define threshold and no limit on iterations?

3. On the data: In my humble opinion, I think that a small sample of the ATFM regulations and a code snippet to work with it (the 'final step' of merging with ATFM regulations in README.md), would make the full cycle of reproducibility complete.

4. On the Software: I cloned the repo and followed the instructions but failed quite soon...

On Apple M1 Max with MacOS Sonoma version 14.1, using mamba 1.4.2 / conda 23.3.1, I created an environment based on Python 10 as suggested in the README.md and I activated it. The installation of the requirements, unfortunately, failed as per attached file[1].

Maybe the required installation needs a specific platform/OS to be successful and if so it should be specified in the README.md

3. Response - round 1

3.1 Response to review 1

On the one hand, the reader would appreciate for authors to elaborate more on the performance evaluation based on the 10% labelled observations set aside. Specifically, the possibility for the trained model to provide unlabelled results raises some questions unanswered. If those validation observations are classified with the trained model, then one would expect a non-negligible percentage of unlabelled results, which is not provided, though. Furthermore, when classification may lead to unlabelled results, the definition of classification metrics (accuracy, precision, recall ...) is less obvious, so they should be explicated.

We appreciate the reviewer's valuable feedback and acknowledge the necessity for enhanced clarity in our paper's methodology. Our approach leverages a self-training technique to unravel as many labels as possible within a partially unlabelled dataset. This technique is adaptable to any classifier capable of providing a probability distribution across potential classes, from rudimentary decision trees to sophisticated neural networks. In our research, we utilised two types of models: Decision Tree (DT) and Gradient Boosting Decision Tree (GBDT). The final iteration of the self-learning algorithm results in a model trained exclusively on labelled and pseudo-labelled observations. It's important to note that these models predict the likelihood of an observation belonging to either the 'weather' or 'other reasons' classes, but not the 'unlabelled' class. To clarify, the self-training algorithm operates independently of the performance evaluation conducted on 10% of the labelled observations. For instance, an observation that yields a prediction of 55% for 'weather' and 45% for 'other reasons' would be considered unlabelled within the self-learning algorithm, as it doesn't meet the threshold criteria. However, during the separate evaluation phase, a conventional binary classi-

[1] Editor note: not included in review report

fication evaluation is carried out. In this scenario, the observation in question would be classified as a 'weather' observation, given the default threshold of 50%.

We hope this detailed explanation enhances the comprehension of our methodology. We have incorporated the provided text into the manuscript to specifically address your comment. We look forward to further discussions and feedback.

On the other hand, and less importantly, on page 10, lines 196-197, authors state "around one-third of the observations originally classified as speed in the left clusters are now classified as ceilings in the right clusters.". This assertion is not correct; in fact, I would say this is the other way around: around one-third of the observations originally classified as speed in the RIGHT clusters are now classified as ceilings in the LEFT clusters.

Your observation is correct. Thank you for pointing out the error.

3.2 Response to review 2

Reviewer Point P 3.1 — In Figure 1(b), there seems to be some irregular holding patterns to the north of Zurich. These could have just been extra vectors given for perhaps sequencing reasons, but they still would induce some non-negligible amount of airborne delay. I would suggest that the authors clarify if their approach is only to identify standardized holding patterns, e.g., such as those depicted in standard arrival routes, or if it's to detect any circuitous patterns that result in airborne delays.

In Figure 1, you can see some irregular holding patterns to the north of Zurich. These patterns are called 360s or orbits, and they are circular patterns in which the aircraft maintains a constant rate of turn. They are used mainly for sequencing and spacing reasons. In this study, we would like to consider them as holdings, as they still induce some non-negligible amount of airborne delay. However, the neural network implemented in the traffic library sometimes detects these patterns as holdings, and sometimes not. All in all, as with any machine learning model, the neural network is not perfect, and some false positives or negatives could be found. The exact performance of the neural network is still unknown, as a publication presenting the details is not yet available. Nevertheless, we have observed that it performs very well, and that missed predictions are very rare.

The provided text has been incorporated into the manuscript to specifically address your comment. We appreciate your valuable input.

In terms of the features that the authors used, they include mostly features related to weather/convective conditions. It would be great if the authors could comment/discuss the possibility of integrating additional data sets, such as those describing active ATFM measures and/or data sets that contain airline schedule information. The reason being that those two data sets should be relatively observable, and may contribute to airborne holding occurrences as well. Even better might be a way to somehow incorporate features that describe on-airport emergency situations, e.g., an emergency aircraft on the runway – those can be a common reason for go arounds and holding. However, I would assume that these happen much more rarely, and may be difficult to get comprehensive data on.

The dataset mostly includes features related to weather conditions. However, other features like airport congestion (which could be expressed as the ratio between the scheduled demand and the declared capacity) as well as more detailed information related to on-airport emergency situations, e.g., an emergency aircraft on the runway, or even information from NOTAMs (notice to airmen) could be included to further help the model identify holdings caused by reasons different from weather during the self-supervision process. However, we kept the study as simple as possible for the sake of reproducibility. Obtaining airline schedules, airport capacities, and NOTAMs is not straightforward, and including these features would have made the reproducibility of the study more complex. The

provided text has been incorporated into the manuscript to specifically address your comment. We appreciate your valuable input.

A minor note, but I would suggest that the authors normalize the x-axis in Figure 2 such that both go up to 12 iterations. This will make it very apparent that GBDT requires 2x the iterations as DT to achieve similar levels of label proportions.

Done. Thank you.

The Shapley value-based analysis of the contributions per feature is certainly helpful in interpreting the results of the classifier. I also would be interested in seeing the set of (arrival) airports at which these classifications were made, specifically the ILS categories of the dominant arrival runways/arrival runway configurations active at each airport. I would assume that there might be a relationship between the approach categories (e.g., Cat III vs. Cat II) and ILS equipment, and the chances of an aircraft having to hold in order to wait for weather minimums to improve.

Grateful for your insightful feedback and constructive suggestions — thank you. This analysis could be enhanced by further categorising airports based on the precision approach categories associated with the active runway configuration at the time of observed holdings (e.g., Cat III, Cat II). Investigating such a relationship between the precision approach category and the likelihood of aircraft holding for weather improvement could yield valuable insights. Unfortunately, to the best of the authors' knowledge, publicly available datasets detailing historical runway configurations at each airport and precision approach categories per runway and airport, are not available. The former dataset could potentially be acquired through the traffic library, utilising some of the provided methods, whereas obtaining the latter dataset necessitates manual extraction from the aeronautical information publications (AIPs). It is crucial to acknowledge that the extraction of such data falls outside the defined scope of this paper. However, we strongly encourage future research endeavours to explore this avenue.

The provided text has been incorporated into the manuscript to specifically address your comment. We appreciate your valuable input.

The authors emphasize in several places that this methodology can be applied more broadly than just for airborne holding, which I agree with. I would suggest that the authors add to their discussions/conclusion, a bit more about how they would guide future researchers to apply their method to the case of path stretching or non-continuous descent approach procedures. Of course, not all of the details need to be fleshed out, but a cursory discussion on what might be the important features involved, how might one set up the neural network architecture/perform the parameter tuning procedures, etc. – these would be great extended discussions to have.

Reply: 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. 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.

For example, the current study's methodology could be reproduced by integrating the additional ASMA time[2] rather than relying solely on the binary indicator of the presence or absence of a holding pattern. Following the approach outlined in this paper, each observation could correspond to the

[2]ASMA stands for arrival and sequencing metering area, representing a 40NM cylinder around the airport. The additional ASMA time provides an approximate measure of the average inbound queuing time on the inbound traffic flow during congested airport periods.

weather conditions during a specified time period (e.g., 30 minutes), enriched with the associated additional ASMA time. These observations could then be categorised as weather-related or others based on the presence of ATFM regulations. Subsequently, employing the semi-supervised approach introduced in this study would facilitate the extrapolation of labels for unclassified observations. This method promises to provide a more exhaustive understanding of the contributing factors to overall airborne delay within the TMA, attributable to airborne holdings and other tactical control techniques.

The provided text has been incorporated into the manuscript to specifically address your comment. We appreciate your valuable input.

3.3 Response to review 3

Lines 88-93: what led to the decision to use the specific parameters used? Why max depth 10, min samples leaf 25, 60 decision trees? The text mentions (Lines 92-93) that the hyper-parameters were chosen to prevent over-fitting: the question is what was the method chosen to choose them?

We manually selected the hyper-parameters of the models based on our understanding of the problem and the characteristics of the data. The hyper-parameters were chosen to create models capable of capturing the complexity in the data while also being robust to over-fitting. The decision to not use a validation set was primarily driven by the limited amount of labelled observations (less than 10K). Although we did not use a validation set, we believe that our choices of hyper-parameters are justified given our understanding of the problem and the data. More specifically:

- We set the maximum depth of each decision tree (max_depth) to 10 considering the complexity of the problem and the number of features in the dataset. A depth of 10 allows the models to learn complex patterns in the data, but not so intricate that they fit to the noise.
- We also set the minimum number of samples required to be at a leaf node (min_samples_leaf) to 25 with the size of the dataset in mind. This parameter ensures that the models make decisions based on a substantial amount of data, which helps to prevent over-fitting by avoiding rules that are too specific and based on a small number of observations.
- Finally, we set the number of trees in the gradient-boosting ensemble (n_estimators) to 60 to strike a balance between model performance and the risk of over-fitting. While a higher number of trees can potentially lead to better performance, it can also increase the risk of fitting to the noise.

The provided text has been incorporated into the manuscript to specifically address your comment. We appreciate your valuable input.

Lines 99-100 and the rest of section 4: similar to the item above what were the criteria, and methods used to define threshold and no limit on iterations?

In the context of the self-training algorithm, we adopted a threshold selection criterion with the threshold parameter set at 0.75. This high threshold ensures that only predictions made with high confidence are added to the training set, which helps to maintain the quality of the labels and prevent the degradation of the model. We chose not to limit the number of iterations, allowing the model to learn as much as possible from the unlabelled data. These parameters also happen to be the default settings in the scikit-learn implementation. While we did use the default settings, our decision was not solely based on convenience or lack of consideration. Instead, it was a deliberate choice informed by our understanding of the problem, the data, and the model's behaviour. The alignment of our choices with the scikit-learn defaults further substantiates our decisions.

The provided text has been incorporated into the manuscript to specifically address your comment. We appreciate your valuable input.

On the data. In my humble opinion, I think that a small sample of the ATFM regulations and a code snippet to work with it (the 'final step' of merging with ATFM regulations in README.md), would make the full cycle of reproducibility complete.

Grateful for your insightful feedback and constructive suggestions — thank you. Your comment has been implemented, and we've now included a dummy example of the ATFM regulations data format along with a code snippet demonstrating the final step of merging with weather and holdings data in the README.md for a more comprehensive and reproducible cycle.

On the Software I cloned the repo and followed the instructions but failed quite soon... On Apple M1 Max with MacOS Sonoma version 14.1, using mamba 1.4.2 conda 23.3.1, I created an environment based on Python 10 as suggested in the README.md and I activated it. The installation of the requirements unfortunately failed as per attached file 'holdings-dalmau_installation-errors.txt'. Maybe the required installation needs a specific platform OS to be successful and if so it should be specified in the README.md

We appreciate your feedback and apologise for the inconvenience you faced. While we couldn't reproduce the error, we suggest following the instructions meticulously: (1) create a clean conda environment and (2) use 'pip install -r requirements.txt'. It's worth noting that Mamba was not specified in the README.md, and our testing encompassed Linux, Mac, and Windows machines. We also improved the README.md. We recommend revisiting the installation steps, and if issues persist, feel free to share any additional details for further assistance.

4. Review - round 2

4.1 Reviewer 1

The revised version provided by the authors successfully addresses all questions raised. Hence, my recommendation is to accept the paper.

4.2 Reviewer 2

The authors have addressed all of my comments – I am happy to recommend acceptance.

4.3 Reviewer 3

I am fully happy with the review work on both my comments and those of the other reviewers. I am really happy because the result is really shining in terms of clarity and reproducibility.

The GitHub repo has greatly improved and I was able to set up everything as described and execute the provided notebook.