Journal of Open Aviation Science (2024), Vol.**2**
doi:10.59490/joas.2024.8078

JOAS

EDITORIAL

Reviews and Responses for Training a Machine Learning Model to Detect Holding Patterns in Aircraft Trajectories

Authors: Xavier Olive, Luis Basora, Junzi Sun, and Enrico Spinielli

Reviewers: Max Li and Christian Verdonk Gallego

Editor: Martin Strohmeier

1. Original paper

The DOI for the original paper is https://doi.org/10.59490/joas.2024.7943

2. Review - round 1

2.1 Reviewer 1

The authors describe the process of developing and training a machine learning model that is capable of distinguishing holding patterns in an aircraft trajectory. Such an open source tool is of great benefit to the overall aviation research community, and this paper walks readers through the research and development efforts required to produce such a tool. I have several comments / suggestions after reading this paper:

(1.) In the introduction section, it would be great if the authors could add a few more previous studies that examines airborne holding. In particular, airborne holding is used in different ways across different FIRs, e.g., in the US, airborne holding rarely occurs unless if there is severe and unexpected runway capacity degradations, whereas in other parts of the world, holding stacks are prominent parts of TRACON/TMA operations to help keep pressure on the runway. I think some more studies would also help to motivate why this is such an important data-driven library.

(2.) The wording of lines 58 through 68 is a bit confusing to me. Are these 6 conditions part of a "standard" definition of holding patterns, or are they a set of irregularly observed trajectory patterns that other holding pattern detection algorithms would miss? I think the authors are describing something important in lines 58-68, but the way it is currently written is vague. I think the problem is in the couple of sentences before the numbered list. I would suggest the authors rewrite this part to clarify their intentions.

(3.) This is sort of related to my first comment – the authors make no mention of the fact that EGLL observes a much higher proportion of holding compared to the other three airports, mostly because EGLL utilizes holding stacks as part of relatively nominal operations. Given that the percentage of flights observing holding patterns are magnitudes apart, perhaps even a small footnote explaining that holding is used differently for different airports would be helpful to readers.

(4.) In lines 107 through 113, the authors mention the use of latent space-based clustering. I would suggest adding a small discussion (i.e., a couple of sentences) that give some intuition for what a reasonable interpretation of a latent space that is "nice" for compressing aircraft trajectory data might be. I understand that there may not be a straightforward interpretation of the various components of the latent space for aircraft trajectories, but some discussion (or acknowledgment that additional work is needed to come up with such interpretations / intuitions would be great)

(5.) A more minor comment: In Table 2, I am not sure if the precision, recall, and F1 score require so many significant digits. Consistent rounding would be great, and would help de-clutter the table as well.

2.2 Reviewer 2

The paper presents the work undertaken by the authors in compiling and training various supervised algorithms capable of detecting trajectory segments within a holding pattern, tested under different experiments with destination airports. The authors' efforts are remarkable: their description of the methodology facilitates replicability and should help the broader community reproduce their results.

However, the final results shown might appear somewhat dissapointing at first glance. It would therefore be advisable for the authors to clarify whether the reported performance metrics (Recall, Precision, F1) refer to segments or entire flights. From the discussion section, it appears that no false positives or false negatives are reported for holding-pattern detection, which may be due to these metrics being applied segment-by-segment rather than to complete trajectories. So, it is not the same missing one segment, that missing the entire holding pattern. Therefore, it might also be worthwhile, for the sake of comparison, to indicate how many holding patterns were entirely missed.

In the introduction, the authors could include additional potential uses beyond those already mentioned. For instance, under the Single European Sky (SES) Performance Scheme, especially Commission Implementing Regulation (EU) 2019/317, Air Navigation Service Providers and Member States must monitor Key Performance Indicators that reflect delays in terminal airspace. Among these, airborne holding is specifically captured by the Additional Arrival Sequencing and Metering Area (ASMA) Time metric.

Essentially, any extra flight time within a given radius of the airport—often during racetrack holding or vectoring—contributes to this “additional ASMA” measure. The authors' algorithm could support systematic detection and measurement of such airborne delays using open data, which may supplement (and perhaps validate) the information provided by ANSPs, thereby facilitating a more transparent evaluation of operational efficiency, capacity constraints, and environmental impact. As part of this extended functionality, the authors might also consider identifying Continuous Descent Operations (CDOs), as described in Section 4.3 of the PRB Annual Monitoring Report 2022.

From a technical standpoint, the paper could reflect on whether the detection algorithm might benefit from methods such as those presented by T. Krauwth et al. (on Deep Generative Modelling of Aircraft Trajectories in Terminal Manoeuvring Areas), which incorporate temporal dependencies for generating latent spaces. Meanwhile, the discussion indicates that no misclassifications are observed. Since the test data appear relatively “regular” in terms of the type of flights that it evaluates, it would be valuable to see how the model behaves with other flight types—such as aerial surveys, test flights, or training holds—to assess generalisability. In these cases, flight context, altitude, and speed profiles might well prove distinctive. Finally, regarding explainability, it is not entirely clear whether the community needs a feature-by-feature explanation of why a segment is classified as a holding pattern. It might be more useful to derive reasons that determine the “state” of a given terminal manoeuvring area from an operational point of view, offering insights for FMPs or network-level decision-making. Such predictive or explanatory capabilities could indeed help stakeholders identify emerging traffic conditions or constraints before they become critical.

Minor Comments and Recommendations 1. Performance Metrics Table: Please clarify explicitly whether the metrics (Precision, Recall, F1) apply to entire flights or individual segments, and consider showing false positives/negatives at the full holding-pattern level.

2. Generalisability: Test the model on less typical or more varied data (e.g., aerial surveys, training holds) to strengthen claims of robustness.

3. Extended Use Cases: Highlight how the approach might assist in meeting SES Performance Scheme requirements (particularly around ASMA time and delay reporting), and consider referencing CDO monitoring where relevant.

4. Explainability: Clarify whether there is a genuine need to provide direct interpretability for each classification decision, or if a higher-level, operational viewpoint would be more beneficial.

3. Response - round 1

Response

The authors would like to thank the reviewers for their encouraging comments. We have taken them into account and improved the manuscript.

3.1 Response to reviewer 1

(1.) In the introduction section, it would be great if the authors could add a few more previous studies that examines airborne holding. In particular, airborne holding is used in different ways across different FIRs, e.g., in the US, airborne holding rarely occurs unless if there is severe and unexpected runway capacity degradations, whereas in other parts of the world, holding stacks are prominent parts of TRACON/TMA operations to help keep pressure on the runway. I think some more studies would also help to motivate why this is such an important data-driven library.

Response

A new paragraph has been added to the introduction section to highlight the importance of airborne holding and the motivation for the study.

(2.) The wording of lines 58 through 68 is a bit confusing to me. Are these 6 conditions part of a "standard" definition of holding patterns, or are they a set of irregularly observed trajectory patterns that other holding pattern detection algorithms would miss? I think the authors are describing something important in lines 58-68, but the way it is currently written is vague. I think the problem is in the couple of sentences before the numbered list. I would suggest the authors rewrite this part to clarify their intentions.

Response

We have rephrased the sentences in clarify the intentions of the authors: "An ML model allows to detect situations even when they do not perfectly match simple necessary conditions to define a holding pattern."

(3.) This is sort of related to my first comment – the authors make no mention of the fact that EGLL observes a much higher proportion of holding compared to the other three airports, mostly because EGLL utilizes holding stacks as part of relatively nominal operations. Given that the percentage of flights observing holding patterns are magnitudes apart, perhaps even a small footnote explaining that holding is used differently for different airports would be helpful to readers.

Response

Thanks for pointing this out, and we added an explanation in the revised manuscript.

(4.) In lines 107 through 113, the authors mention the use of latent space-based clustering. I would suggest adding a small discussion (i.e., a couple of sentences) that give some intuition for what a reasonable interpretation of a latent space that is "nice" for compressing aircraft trajectory data might be. I understand that there may not be a straightforward interpretation of the various components of the latent space for aircraft trajectories, but some discussion (or acknowledgment that additional work is needed to come up with such interpretations / intuitions would be great)

Response

We have added information on Autoencoders, which are used to compress the data from high-dimensional space to a lower-dimensional space, in the revised manuscript. We hope this will help the readers to understand the latent space-based clustering.

(5.) A more minor comment: In Table 2, I am not sure if the precision, recall, and F1 score require so many significant digits. Consistent rounding would be great, and would help de-clutter the table as well.

Response

We have rounded the precision, recall, and F1 scores in Table 2 to less decimal places.

3.2 Response to reviewer 2

However, the final results shown might appear somewhat dissapointing at first glance. It would therefore be advisable for the authors to clarify whether the reported performance metrics (Recall, Precision, F1) refer to segments or entire flights. From the discussion section, it appears that no false positives or false negatives are reported for holding-pattern detection, which may be due to these metrics being applied segment-by-segment rather than to complete trajectories. So, it is not the same missing one segment, that missing the entire holding pattern. Therefore, it might also be worthwhile, for the sake of comparison, to indicate how many holding patterns were entirely missed.

1. Performance Metrics Table: Please clarify explicitly whether the metrics (Precision, Recall, F1) apply to entire flights or individual segments, and consider showing false positives/negatives at the full holding-pattern level.

Response

We have clarified that the reported performance metrics (Recall, Precision, F1) refer to segments in the revised manuscript.

Based on the recall, around 0.7, we can infer that 30% of the holding patterns were missed based on our labeled dataset. However, due to the lack of ground truth data, we cannot provide the exact number of holding patterns that were entirely missed.

In the introduction, the authors could include additional potential uses beyond those already mentioned. For instance, under the Single European Sky (SES) Performance Scheme, especially Commission Implementing Regulation (EU) 2019/317, Air Navigation Service Providers and Member States must monitor Key Performance Indicators that reflect delays in terminal airspace. Among these, airborne holding is specifically captured by the Additional Arrival Sequencing and Metering Area (ASMA) Time metric.

> Response
>
> We have added a paragraph in the introduction section to highlight the potential uses of the approach in the SES Performance Scheme.

Essentially, any extra flight time within a given radius of the airport—often during racetrack holding or vectoring—contributes to this "additional ASMA" measure. The authors' algorithm could support systematic detection and measurement of such airborne delays using open data, which may supplement (and perhaps validate) the information provided by ANSPs, thereby facilitating a more transparent evaluation of operational efficiency, capacity constraints, and environmental impact. As part of this extended functionality, the authors might also consider identifying Continuous Descent Operations (CDOs), as described in Section 4.3 of the PRB Annual Monitoring Report 2022.

3. Extended Use Cases: Highlight how the approach might assist in meeting SES Performance Scheme requirements (particularly around ASMA time and delay reporting), and consider referencing CDO monitoring where relevant.

> Response
>
> The continuous descent operations (CDOs) are a good point. However, we think simpler algorithms than neural networks could be used to detect CDOs. Perhaps, this could be a topic for future papers.

From a technical standpoint, the paper could reflect on whether the detection algorithm might benefit from methods such as those presented by T. Krauwth et al. (on Deep Generative Modelling of Aircraft Trajectories in Terminal Manoeuvring Areas), which incorporate temporal dependencies for generating latent spaces. Meanwhile, the discussion indicates that no misclassifications are observed. Since the test data appear relatively "regular" in terms of the type of flights that it evaluates, it would be valuable to see how the model behaves with other flight types—such as aerial surveys, test flights, or training holds—to assess generalisability. In these cases, flight context, altitude, and speed profiles might well prove distinctive.

2. Generalisability: Test the model on less typical or more varied data (e.g., aerial surveys, training holds) to strengthen claims of robustness.

> Response
>
> The generalization of neural network models is a more complex topic, and we think this would be a good topic for future research. The clustering approach we employed in this paper indeed has similarities with the work of Krauwth et al. This helped us to preselect the holding pattern. However, a lot of manual efforts is still needed to examine them one by one. The actual model performed best for holding pattern detection is a convocational neural network model (Figure 7), which is not directly comparable to the work of Krauwth et al.

Finally, regarding explainability, it is not entirely clear whether the community needs a feature-by-feature explanation of why a segment is classified as a holding pattern. It might be more useful to derive reasons that determine the "state" of a given terminal manoeuvring area from an operational point of view, offering insights for FMPs or network-level decision-making. Such predictive or explanatory capabilities could indeed help stakeholders identify emerging traffic conditions or constraints before they become critical.

4. Explainability: Clarify whether there is a genuine need to provide direct interpretability for each classification decision, or if a higher-level, operational viewpoint would be more beneficial.

Response

We agree with the reviewer that the operational viewpoint is more beneficial than a feature-by-feature explanation. However, this would require more features to be included beyong just the open trajectory data. This is a good point for future research.

4. Editor Note

The paper was accepted for publication as submitted, with both reviewers recommending acceptance. The authors also addressed the comments and suggestions made by the reviewers in their response in a revised version of the manuscript.