International Civil Aviation Organization

INFORMATION PAPER

A41-WP/378
TE/142
22/8/22
(Information paper)
English only

ASSEMBLY — 41ST SESSION

TECHNICAL COMMISSION

Agenda Item 30: Aviation Safety and Air Navigation Policy
30.3 Relevant Outcomes of the High-level Conference on COVID-19, Safety Stream (HLCC 2021)

AUTOMATION CAPABILITIES TO MONITOR SAFETY METRICS IN THE UNITED STATES NATIONAL AIRSPACE SYSTEM

(Presented by the United States)

EXECUTIVE SUMMARY

This paper describes the computational methods the Federal Aviation Administration (FAA) Air Traffic Organization (ATO) has employed to monitor the safety performance in the National Airspace System (NAS). Several machine learning models utilizing Natural Language Processing (NLP) approaches have been developed to categorize unstructured text reports from a variety of sources to filter for indicators of accidents and safety-impacting incidents. Combined numerical metrics were derived for surface and airborne events, and trends tracking these safety metrics are presented to stakeholders on a monthly basis. The paper also provides an overview on a prototype model the ATO has developed to detect encounters between an Unmanned Aircraft System (UAS) and manned aircraft by leveraging voice communication between pilots and controllers without relying on reported incident data.

Strategic Objectives:	This information paper relates to the Safety and Air Navigation Capacity and Efficiency Strategic objectives.
Financial implications:	This paper contains no significant financial implications.
References:	Bati et al., Risk Metrics to Measure Safety Performance of the National Airspace System: *Implementation Using Machine Learning. The 40th DASC Conference, San Antonio, Texas, USA – October 2021*

1. INTRODUCTION

1.1 Metrics are the key tools to monitor the safety performance of complex systems like air traffic control. Risk metrics may involve prediction of risk and identification of different potential outcomes of undesired safety events. For instance, a metric that helps monitor the risk on the surface in an airport environment needs to detect and identify events such as runway excursions, runway incursions, and taxiway incidents and assign appropriate numerical indices proportional to the outcome of each event. An accident with a fatality, injury and aircraft damage will have a corresponding weight for each outcome and an incident that involves no outcome can be assigned a severity weight based on its probability of becoming an accident. Models that support such metrics need to be able to process data from diverse sources. In this paper, we discuss key aspects of comprehensive metrics that the Office of Safety in the ATO (ATO Safety) has deployed recently: 1) How to employ automation to detect all relevant events from different data sources, and 2) how to assign severity weights for different types of events (accidents and incidents).

1.2 The ATO employs various performance metrics to monitor different aspects of the NAS. For safety performance monitoring, some of these metrics are based on risk prediction and others are based on actual outcomes. The metrics explained in this paper are outcome-based, and the key logic is that the ultimate worst outcome in the NAS is an accident involving a fatality and a hull loss. Hence, it assigns the highest severity weight to a catastrophic accident; all other event types are measured according to their relative "proximity" to a catastrophic accident. As such, the metrics incorporate all relevant types of events that can occur in the NAS. For surface safety, the metric scores events such as runway collisions, runway excursions, runway incursions and taxiway incidents. For airborne safety, the metric scores events related to flight into terrain, mid-air collisions and turbulence. By incorporating all relevant safety events, the metrics measure the overall safety performance of the system more accurately and depicts what is occurring in the airport environment and during flight.

1.3 An important element of a comprehensive metric is an automated detection of relevant events by categories. Unlike in aviation, the use of Artificial Intelligence (AI) and Machine Learning (ML) have permeated most industries. Aviation is a safety-critical domain in which there is very little tolerance for failure. The stringent requirements of aviation have contributed to the slow adoption of AI and ML in aviation, in general. However, many aviation sectors are increasingly adopting AI systems, ranging from automating simple yet tedious and repetitive tasks, to a more sophisticated application of a complex autonomous air traffic control system to de-conflict traffic. In this paper, we outline how machine learning models were employed to identify relevant accidents and incidents to support two metrics; a surface safety metric and an airborne safety metric.

1.4 For a metric to be a comprehensive measure and indicate the overall performance of the system, it needs to account for various types of accidents and incidents (precursors) that occur in the system. This paper shows a weighting scheme we developed to measure the outcome of accidents, such as injuries to people and damage to property, as well as probabilistically determine the severity of incidents with no outcome. The aggregation of all weights along with the frequency of occurrence in a given period represent the overall safety performance of the system.

1.5 The metrics account for changes in the NAS indirectly using penalty and credit terms. The NAS has gone through various improvements over the years; e.g. technological, policy/regulation, procedural, training, etc.Therefore, it is very difficult to isolate the effect of each change on the overall safety improvements of the system. The penalty term in the metric measures the undesired outcomes of events, such as injury to people and damage to property, while the credit term accounts for preventing those outcomes. Analysis shows that, particularly in the surface environment, what has improved significantly is the reduction in the negative consequence of events, not the frequency of incidents. For instance, a runway

excursion (overrun) that occurs in the current system is less likely to be catastrophic than a similar incident several decades ago due to technological improvements like an Engineered Materials Arrestor System (EMAS). Similarly, in the airborne environment, a Near Mid-Air Collision (NMAC) is less likely to be catastrophic due to the use of the Traffic Avoidance Collision System.

1.6 The approach in the two metrics involves a heavy application of ML for event identification. Due to a large number of minor events in the system, it is impractical to manually identify all relevant incidents and accidents without automation capabilities. Several ML techniques were employed to identify events that are relevant for scoring. In addition, odds ratio was employed to compute the relative distance of incidents to a catastrophic outcome using 20 years of historical accident database. To categorize relevant events further into appropriate airborne and surface subcategories, an ensemble classification model was employed. The scoring process runs every month capturing the preceding 12 months' worth of data and the running average of the overall safety performance.

1.7 A valuable new source of information has recently emerged in the form of voice communication transcriptions. Radio communications between pilots and air traffic controllers are automatically transcribed by using a speech-to-text technology. A trial project has been conducted to detect UAS sighting reports within the voice transcriptions by natural language (NLP) processing with high levels of accuracy achieved.

2. DATA PROCESSING

2.1 There are several data sources enabling these safety metrics, some of which are maintained by the FAA while others are owned by industry partners. To identify relevant events, some structured fields were leveraged as part of merging the disparate data sources. Various approaches were applied to remove duplicates such as airport ID, date, and time of the event. The pre-processing also helped normalize common spelling errors and abbreviations used by pilots and air traffic controllers in the reports. The ATO assembled a dictionary of most frequent spelling variants to map to canonical terms. The ATO also normalized different ways by which an empty narrative was indicated ("no data", "no narrative available", "no narrative given", etc.) to avoid spurious correlations with target outcomes.

3. EVENT IDENTIFICATION

3.1 Automation was of great importance to identify the relevant events from different sources and sustainably monitor the systems using data-driven metrics. Actual event outcomes and classification of events into subcategories are necessary to set severity weights for the incidents. Several supervised ML and deep learning models were trained and the best performing models were used for event classification.

3.2 **Training Sample**: The ATO started creating the training samples by using the structured fields populated by investigators and reporters and mostly by reviewing the event narratives. The ATO then cleaned the text narratives employing basic NLP functionalities such as, converting words into lower-case, removing irrelevant items such as standalone numbers, punctuations, white spaces, stop words and applying word stemming. An iterative pattern identification of the event narratives was performed using a text extracting tool to determine subcategories by enhancing the pattern identification and locating key phrases. Other steps of text mining include feature extraction from the narratives. Customized terms unique to the aviation safety domain were captured by employing regular expressions, which helped to improve the feature extraction process and overall performance of the models. Finally, we produced the training sample with a number of features and corresponding target subcategories.

3.3 **Model Development & Evaluation**: The performance of models vary depending on the contexts or use cases. We experimented with various traditional ML and recent NLP models. We also experimented with a majority voting approach with a set of thresholds for outputs from different models that differ largely in their underlying algorithms. The generalizing ability of the supervised models was evaluated by using independent test sets. Performance of each model was evaluated by calculating a set of metrics such as, accuracy, precision, and recall.

4. OUTCOME-BASED WEIGHTING

4.1 The metrics assign different levels of severity weight to each event proportional to their proximity to a catastrophic accident. For the purpose of safety risk management, FAA defines a catastrophic accident as an accident with three or more fatalities or a hull loss with any fatality. The weight is the result of a combination of quantitative measures derived from data and assumptions appropriate for airborne and surface operations. This technique employs the following three-step approach to calculate the relative severity index for each event:

a) accident outcomes are ordered such that injury to people is treated as more severe than damage to aircraft;

b) accident weights are assigned according to the event's proximity to a catastrophic outcome; and

c) incident weights are assigned according to the weights of their corresponding accident types.

4.2 **Accident Outcome Ordering**: In addition to informing safety decisions, the purpose of the metrics is to communicate with the public about the overall safety performance of the NAS based on actual outcomes, not predicted risk. This objective assumes that life is more valuable than property, and the metrics are designed to reflect this assumption. Therefore, to align the metrics with this objective, the weighting scheme was constructed such that a higher weight is assigned to injury than to aircraft damage. In addition, different accident outcomes are categorized cumulatively. That is, a high severity outcome is also considered a low severity outcome. The following is the outcome order maintained in the probability estimation.

$$\textit{Fatal Injury} \rightarrow \textit{Serious Injury} \rightarrow \textit{Minor Injury} \rightarrow \textit{Destroyed Damage} \rightarrow \textit{Substantial Damage} \rightarrow \textit{Minor Damage}$$

4.3 The counterintuitive nature of this ordering is that there may be cases where an accident involving just a minor injury is weighted more heavily than an aircraft damage that resulted in a hull loss, although such cases are rare due to the fact that with hull loss there is likely injury to passengers.

4.4 **Accident Weights**: Severity weights are assigned to accidents (i.e., events with outcomes such as injury and/or aircraft damage) according to their proximity to a catastrophic accident by using odds ratio. Odds ratio is a measure to compare the odds of one event type to another. As such, it helps to quantify the strength of the association between two events. If the two events are independent, the odds ratio becomes 1. Odds ratio of greater or less than 1 is an indication of a correlation between the events. Odds of an event is the ratio of probability of one event with another, typically with its complement event. It is another representation of probability of an event.

4.5 **Incident Weights**: In this metrics application, when an event doesn't result in any outcome (no injury and no aircraft damage), it is categorized as an incident. As described earlier, relevant incident types both in the airborne and surface environments are included. In general, accidents and incidents fall under similar categories by type. Based on domain knowledge, most event types have some probability of becoming accidents. For instance, in the surface environment, runway incursion incidents have some probability of becoming runway collisions. Similarly, in the airborne environment, NMAC incidents have some probability of becoming a MAC. Therefore, by mapping the incident types to the appropriate accident types, weights can be indirectly assigned to incidents proportional to their relative counts. Because incidents are less severe than accidents based on the actual outcome, an adjustment term must be introduced to complete the mapping. For each incident type, a reasonable adjustment term is the inverse frequency of that incident type, which is consistent with the Heinrich Triangle.

5. APPLICATION OF SEVERITY WEIGHTS

5.1 Severity weights are calculated for each outcome of an accident; that is, a corresponding weight exists for each type of injury and extent of aircraft damage. However, each incident type is assigned a single weight because there is no outcome in incidents. Therefore, the weight application differs slightly depending on the type of event; that is, for accidents we would consider the number of injuries and aircraft damage to compute the overall weight.

5.2 **Penalty Term**: For this application, penalty is a term that indicates the overall severity of an event. It is a direct measure of all the consequences of the event. In the aggregation of weights, the weights are summed over all the undesired outcomes. For instance, for an event involving both injuries and aircraft damage, the product of the weight for each injury type and the number of injuries is added to the product of weight for each damage type and number of damaged aircraft.

5.3 **Credit Term**: A safety performance metric should account for the deficiency of the system as well as the safety improvements and technological advances that contribute to the reduction in safety event consequences. Over the years, numerous procedural and technological solutions have been deployed in the NAS to reduce the frequency and impact of safety event consequences. Such solutions for the runway environment include Airport Surface Detection Equipment (ASDE) and the EMAS. ASDE is a surveillance system that uses radar, multilateration, and satellite technologies to enable air traffic controllers to track the surface movement of aircraft and vehicles. Through early detection of conflicts on the runway, ASDE minimizes the severity of runway events and reduces the likelihood of more severe events and potential collisions. Similarly, EMAS is placed at the end of runways and breaks down under the weight of an aircraft to minimize human injury and aircraft damage. The ATO designed the metric to account for such improvements indirectly by incorporating a credit term proportional to non-injured people and non-damaged aircraft involved in an event. Events with fewer people injured in proportion to the total number of people onboard and minimally damaged aircraft are assigned relatively smaller weights using a credit term. The credit term, therefore, reduces the penalty term assigned to an event proportional to the extent of non-injured people and undamaged aircraft.

6. RESULT

6.1 The surface metric demonstrates a critical element of domain knowledge: although events still happen (and in some cases at relatively the same rate as before), their severity has decreased significantly. Even though the number of reported accidents in the surface environment is relatively unchanged, and the number of reported incidents has increased, the NAS has become safer overall. There

are two potential explanations for this. First, the solutions that have been deployed in the NAS over the last few decades (e.g., training, procedures, technological advances, etc.) are collectively resulting in significant safety improvements. Second, the adoption of a Just Culture that promotes increased reporting of unsafe actions and incidents has resulted in a highly efficient, continuously improving system.

6.2 Therefore, even though the number of reported incidents is steadily increasing, the overall severity of events is trending downward. A fatal commercial accident in the runway environment has not occurred in nearly a decade. Thus, an increasing number of reported incidents may be required to increase general knowledge and improve the overall safety performance of the NAS. As the metrics demonstrate, the increasing number of reported incidents is not necessarily a negative outcome and, therefore, optimal safety decisions cannot be based on the number of incidents alone.

7. ADDITIONAL CAPABILITIES USING VOICE DATA

7.1 Voice data in the form of machine transcribed communications between pilots and air traffic controllers can become a very important new source of information. ATO Safety has recently explored voice data to provide additional context and automation of risk detection using voice data. Several natural language processing models were successfully trained to detect drone activities (encounters between a UAS and manned aircraft, Class-B excursions, and Pilot Reports, with accuracy levels exceeding 90%. An ensemble approach combining the output of several different models to create a high accuracy agglomerative model was used to achieve 91% accuracy for the UAS encounter detection.

7.2 A comparison with the FAA's Mandatory Occurrence Reports (MOR) database was performed to verify the overlap between the events detected via voice communication analysis and the events officially reported to the MOR. The overlap was found to be only at 6% meaning that 94% of UAS sightings which can be detected from radio communications were never officially reported. This does not necessarily mean a safety violation has occurred but it demonstrates the value of the voice transcription analysis.

8. CONCLUSION

8.1 With the increasing air traffic volume the amount of information with potential safety implications is growing beyond the capacity of human subject matter experts to review. This necessitates the usage of automated systems which can transform unstructured text data into a numerical performance metric. High accuracy models and safety metrics have been developed to filter and classify safety-related events and track the nationwide trends for the US airspace.

8.2 Transcribed voice communications between pilots and controllers is a large, and so far under-utilized source of data which can facilitate detection of safety-impacting occurrences. The events detected through voice transcriptions are complimentary to the other sources of reports.

8.3 The Assembly is invited to note the information provided in this paper.

— END —