Development of a Metric Concept that Differentiates Between Normal and Abnormal Operational Aviation Data

Matthew Stogsdill,1,* Daniele Baranzini2 and Pernilla Ulfvengren1

There is a strong and growing interest in using the large amount of high-quality operational data available within an airline. One reason for this is the push by regulators to use data to demonstrate safety performance by monitoring the outputs of Safety Performance Indicators relative to targeted goals. However, the current exceedance-based approaches alone do not provide sufficient operational risk information to support managers and operators making proximate real-time data-driven decisions. The purpose of this study was to develop and test a set of metrics which can complement the current exceedance-based methods. The approach was to develop two construct variables that were designed with the aim to: (1) create an aggregate construct variable that can differentiate between normal and abnormal landings (row_mean); and (2) determine if temporal sequence patterns can be detected within the data set that can differentiate between the two landing groups (row_sequence). To assess the differentiation ability of the aggregate constructs, a set of both statistical and visual tests were run in order to detect quantitative and qualitative differences between the data series representing two landing groups prior to touchdown. The result, verified with a time series k-means cluster analysis, show that the composite constructs seem to differentiate normal and abnormal landings by capturing time-varying importance of individual variables in the final 300 seconds before touchdown. Together the approaches discussed in this article present an interesting and complementary way forward that should be further pursued.

KEY WORDS: Flight data; safety metrics; safety performance

1. INTRODUCTION

The advent of the digital era has revolutionized the collection and storage of data within the Air Transport System (ATS). This data can be used to assess and improve safety in the ATS generally, and airlines more specifically. Consequently, there has been increasing interest from both practitioners (and ATR, 2016; Civil Aviation Authority [CAA], 2013; International Civil Aviation Organization [ICAO], 2013, 2018) and academics (Fernández et al., 2019; Oehling & Barry, 2019; Stolzer, Halford, & Goglia, 2015) to leverage this data to further aviation safety initiatives. It has been shown that current approaches used in operations do not sufficiently address the industry’s need for generating the required risk information needed to improve safety performance (Ulfvengren et al., 2013), which has resulted in a safety plateau using the current methods. For further improvements to be realized, methods will need to be developed that can integrate and analyze data, not only in retrospect for managerial/operational post hoc decision making (as is the current practice), but also to share actionable information to crews in real time. Such advancements, which are still in their infancy in the commercial aviation industry, would be able to differentiate between “normal” and
“abnormal” situations based not upon event identification, but rather on detecting differences within the data risk patterns prior to, or co-occurring, with target events (Baranzini & Zanin, 2015; Distefano & Leonardi, 2018; Wagner & Barker, 2014).

The rarity of airline accidents has created a Catch-22, where the information needed to improve the system using the traditional “fly-fix-fly” approach is no longer being generated in sufficient quantities (or in generalizable conditions) (Leveson, 2011; Walker, 2017). While the standard practice since 2014 has been to monitor certain outcomes that are viewed as precursors to more serious events (Stolzer et al., 2015), this logic is predicated on the idea that accidents will occur in repeatable ways. This assumption is based upon the belief that safety is controllable using the Plan, Do, Check, Act (PDCA) cycle and is based on the same underlying logics and assumptions that govern quality management systems (ICAO, 2018; Stolzer et al., 2015). However, accidents like Air France 447, U.S. Airways 1549, Colgan Air 3407, and Malaysia Airlines 370 all represent cases where the accident was “novel” in some way; that is, the event concatenations and order sequences had never occurred before (i.e., no repeating history paths), and therefore would not be preventable/detectable using these types of analyses/logics. While this is an inherent weakness in most popular current approaches, it does not mean that these approaches should be discarded. In fact, they are good at accounting for those events expressing some stationarity and historical repetition. However, such methods would fall short when dealing with nonstationary conditions and true “novel” events.

This study uses flight data collected on board the aircraft to investigate ways of differentiating between “normal” and “abnormal” operations prior to an event occurring. Previous studies have been able to identify the difference between “normal” and “abnormal” approaches/landings (Fernández et al., 2019; Oehling & Barry, 2019; Wang, Ren, & Wu, 2018), but such event driven insights will prove to be preventative only if: (1) the identified differences find (risk) patterns that are highly correlated with unwanted events, or (2) abnormal sequences can be detected in real time and forwarded to the pilots so they are warned about a potential impending event. The former can be used by management and training personnel in order to benchmark and improve standard operating procedures, while the latter can be used by operational personnel as a support tool which helps them more quickly detect undesirable operational contexts.

Historically, there has been very little distinction between these two goals, as they both have been lumped together under the “safety improvement” banner. Though the current methods create information that is interesting for managers, it is of little direct use to operational personnel. The purpose of this study was to develop and test a set of metrics which can complement the current exceedance-based methods:

1. How well does a simple aggregated construct variable perform relative to the individual constituent variables when differentiating between normal and firm landings prior to touchdown?
2. How does adding an explicit time component effect the simple construct’s performance?
3. How do the two created constructs compare to a more advanced dynamic clustering method?

2. STATE OF ART

Currently the most popular quantitative approaches used to evaluate airline safety performance are based upon exceedance detection metrics (CAA, 2013). These metrics are in turn used to create rate-based Safety Performance indicators (SPI), which are most often calculated as the number of events per 1,000 flights (ATR, 2016; CAA, 2013; ICAO, 2018; Stolzer et al., 2015). These rate-based SPIs primarily describe the current state of the system relative to previous system states. An example of this type of metric is the Hard Landing SPI, which describes the number of landings with a normal acceleration greater than a specified threshold per 1,000 flights, which is then compared to the same metric for previous months (and in some cases years). While this data can ideally be used to drive safety improvements, most often the complexity and variability within operations makes it difficult to study and assess risk if a “trend” is due to an underlying factor or random chance (Schilling, 1990).

Without understanding the function and mechanisms that drive a process, the knowledge of the resulting outcome variation is of limited value when trying to improve the performance of the process, if not completely useless. To address this issue, academics have become increasingly interested in utilizing Artificial Intelligence (AI) or Machine Learning (ML) methods as a way to discover new underlying patterns within the data. While such methods have
been widely adopted by some industries with great success, their application and adoption into safety critical domains has been relatively slow (Hegde & Rokseth, 2020). This slower adoption rate is likely due to a number of factors. First, many organizations lack the maturity to apply the most recent advances in data analytics (Big Data and AI/ML predictive models) to their Safety Management System (SMS). Second, the cost of failure when a predictive model generates a false negative (e.g., the model does not predict a risk, but an accident occurs) is typically unacceptable. Lastly, they can lack interpretability, a concept that Molnar (2020) describes as the ability to understand an entire model at once (i.e., the model, the underlying algorithm, and the data). While not all AI/ML algorithms have low degrees of interpretability, it is of utmost importance to carefully select the most “readable” models if they are going to be used within a safety critical domain such as commercial aviation.

In reviewing the literature, we found that studies can be roughly split into two categories: outcome focused, and process focused. The outcome focused approaches are primarily interested in attempting to find some interesting/unknown patterns within the flight data (see Das, Li, Srivastava, & Hansman, 2012; Fernández et al., 2019; Oehling & Barry, 2019). The process focused approaches have been more interested in determining differences between “normal” and “abnormal” outcomes (Wang et al., 2018). Though both approaches produce thought-provoking results, the general lack of interpretability of the outcome focused metrics, along with their increased computational demands, led us to start exploring alternative process focused methods. Though the process focused approaches are not as sophisticated as many of the outcome focused methods mentioned above, they are nonetheless still able to find patterns within the data that are not easily identified by the standard methods the aviation industry is currently employing.

3. METHOD

Accidents within the ATS are fortunately rare, which makes preventing them difficult. While accidents are unwanted events, the methods currently used to improve safety draw heavily from such events to learn what within the system needs to be improved going forward (Walker, 2017). This difficulty is further compounded by the lack of control over exogenous factors, which results in a limited ability to determine cause and effect relationships. In an attempt to minimize the influence of exogenous factors and stochastic noise, the following criteria were used when selecting the data set to be analyzed:

1. All aircraft were of the same type (B737–800).
2. The flight route had the same departure/destination pair.
3. Two runways were used to compare and validate results.
4. The collected data contained at least three self-identified hard landings (maximum normal acceleration ≥ 1.7 Gs).

4. DATA COLLECTION

The obtained data set consists of 1,000 Boeing 737–800 flights occurring between July 2016 and March 2018. The destination airport in the data has a single runway 01-19 that is 45 m wide and over 2,000 m long. The quick access recorder data for each flight consisted of 75 variables. However, several of these variables had duplicate information (e.g., the flight_date_time variable contained all the data in the flight_date variable) allowing for 17 variables to be removed, creating a data set consisting of 58 multimodal variables. Drawing from the work of Wang et al. (2018), seven key variables were then selected to be analyzed in this study, as described in Table I. The selected variables represent inputs over which pilots have direct control and can be used to determine the aircrafts kinematic state.

The data was further refined to include only data from the 301 seconds prior to and including the touchdown moment, which is roughly equivalent to the final approach to landing flight phase. The data was collected at a rate of 1 Hz, except the normal acceleration at landing which, while extracted at 1 Hz, is based upon the maximum 8 Hz value recorded by the accelerometers. While it would have been ideal to conduct the entire analysis using 8 Hz data, it was determined that 1 Hz was a more reasonable data rate given that it is closer to the processing time of the pilots who are making the moment-by-moment control input decisions.

4.1. Data Preparation

Upon being imported to Python 3.7, the following variables were created: touchdown_point (TDP), n1_sum, and landing_group, the latter of which categorized each flight into either the “normal” or “firm”
Table I. Variables Selected for Analysis Along with a Description, Abbreviation, and the Unit Being Measured

| Variable                  | Description                  | Abbreviation | Units |
|---------------------------|------------------------------|--------------|-------|
| Aircraft pitch            | Degree of rotation around aircrafts horizontal axis (wingtip to wingtip) | pitch        | Degrees |
| Bank (Roll)               | Degree of rotation around aircrafts longitudinal axis (nose to tail) | bank         | Degrees |
| Flaps position handle     | Selected degree of flaps      | flaps        | Degrees |
| Sum of N1 engine speeds   | Summed N1 fan speeds for all engines | n1_sum      | Rotations/Minute |
| Instantaneous vertical velocity | The rate at which the aircraft is changing altitude | ivv          | Feet/Second |
| Computed airspeed         | The corrected and calibrated speed the aircraft is moving through the air | airspeed     | Knots   |
| Calculated gross weight   | The calculated total weight of the aircraft | gross_weight | Pounds (US) |

Table II. Description of the Collected Data Set Categorized by Landing Runway

| Variable   | Runway 01 | Runway 19 |
|------------|-----------|-----------|
| ivv         | 130,032   | 170,968   |
| bank        | 130,032   | 170,968   |
| pitch       | 130,032   | 170,968   |
| n1_sum      | 130,032   | 170,968   |
| computed_airspeed | 130,032   | 170,968   |
| flap_actual | 130,032   | 170,968   |
| gross_weight| 130,032   | 170,968   |

landing groups. The categorization was based upon a 1.5 G normal acceleration at touchdown threshold, in which flights with normal accelerations below the threshold were considered to be “normal” \( n_{normal} = 809 \), while flights above this threshold were considered to be “firm” landings \( n_{firm} = 191 \). This resulted in a 0.236 firm to normal landing ratio. The 1.5 G threshold was selected as a realistic boundary beyond which passengers may start feeling uncomfortable with the landing, and perhaps even concerned that an event occurred. While aircraft are designed to take stresses in excess of 2 G’s, this “low” threshold is also sufficiently conservative that enough events occur to conduct analyses without severely violating several statistical assumptions. The data was further separated by the runway on which the aircraft landed. This also increases control over possible confounding effects in the subsequent analyses. Of the 1,000 flights analyzed, 432 flew an approach to Runway 01 \( n_{normal, rwy01} = 348 \), \( n_{firm, rwy01} = 84 \) with the remaining 568 landing on Runway 19 \( n_{normal, rwy19} = 461 \), \( n_{firm, rwy19} = 107 \). Results are shown in Table II.

4.2. Construct Development

The following two construct variables represent a basic approach which will complement existing analysis types to create proactive or predictive airline insights at both managerial and operational levels. While these approaches provide alternative perspectives to the other approaches previously discussed, there are still major limitations that will need to be addressed in future research.
4.2.1. Instant Differentiation Construct (row_mean)

The first of the two constructs is the mean unweighted linear combination of the seven variables described in Table I, with values normalized between −1 and 1 and a mean of zero. This approach was selected in order to account for the interdependent nature of the variables based upon changes in the operational context. Each variable is treated as a partial component of the latent row_mean construct and used to create an information signature for each observation during the approach. By making the variables have a mean of zero, small patterns in variable combinations encountered in the approach are magnified (and have directionality). The result of this is that relatively high or low values in one or more of the selected variables result in a row_mean value that deviates from the “ideal” norm value of zero. If similar deviations occur repeatedly and can be differentiated through the use of a cluster analysis (and/or other statistical methods), this could be a good indicator for management and training staff to use in detecting odd/risky patterns that need to be further investigated or mitigated.

As there are an almost infinite number of ways that these seven variables can interact to create the row_mean total, it is impossible to say if a particular value is “good” or “bad,” but the total does contain the raw values of the information collected in that moment as well as how those variables linearly relate to one another. To better explain this, consider the following example: An aircraft in straight and level flight has balanced all four flight forces (i.e., lift, weight, thrust, drag), and would have a certain row_mean value depending upon the context. However, in a different context (e.g., a descending turn during an approach), it is entirely possible that the same row_mean value could be observed, which makes the row_mean value extremely sensitive to the operational context and is only comparable when the operational contexts are very similar. Though, when a comparison is possible, as it is in this study, the underlying differences in the data’s aggregated structure should appear as aggregated differences between the normal and firm landing groups.

4.2.2. Temporal Differentiation Construct (row_sequence)

The row_sequence construct was developed to explicitly address the temporal aspects encountered in an approach, specifically those that can be called path dependent. While the row_mean construct represents momentary snapshots that can be strung together, such a representation does not fully capture the temporal interconnections that exist in operations; more specifically, how the inputs at $t_0$ effect the system not only at $t_1$ and $t_2$ but also at $t_n$. By treating each landing sequence as though it is occurring in real time and accounting for the number of observations which have thus far been made, the row_sequence construct is able to determine if a deviation is increasing or starting to return to a more normal range. Whereas, the previous construct is more geared toward aggregated managerial level analyses, the row_sequence construct is more operationally driven as it focuses on the approach path of the aircraft as it is evolving.

The row_sequence construct is dependent upon the logic that a pilot on an approach must make decisions about how to achieve the desired goal (i.e., a safe landing) at touchdown based only on the data available at or before the present. Though basic, this construct does create the potential for individual flights to be compared to historically commensurate approaches that ended successfully in the past. A deviation from those past successes does not necessarily mean that an event will occur, but it does indicate that the aircraft is entering a context that is less understood, which in turn suggests a rising chance that an undesired outcome will occur. It is important that the row_sequence calculation occurs frequently so that when a deviation does start to emerge a correction can be implemented as soon as possible.

As an example, consider an aircraft on an approach to an airport. During the approach the pilot is not only making control inputs based on what is going on in the present moment, but also what has happened in the past. This makes the inputs path dependent, meaning that as an approach evolves, an increasing amount of information needs to be considered by the pilot when deciding how to achieve a safe landing. While the calculations occurring in the pilot’s head are far more complex than is being done to compute the row_sequence construct, the underlying logic remains the same; that is, the temporal evaluation pattern is as important, if not more so, than any single value or set of values.

4.3. Time Series Clustering

To create a baseline against which the developed metrics can be tested, a time-series k-means clustering algorithm using a dynamic time warping (DTW)
metric was carried out (Ives, 2016). Given the aim of this work is to differentiate between normal and firm landing groups, we chose to search for two clusters within the data with the explicit intent of separating out normal from firm landings into two different clusters generated by the algorithm. The DTW method was selected to find the best alignment and similarity between a collection of time series exposed to clustering where the similarity match is invariant to certain non-linear variations in the time dimension. This type of analysis was originally developed in the field of AI and Deep Learning for time movement and speech recognition (Lundtrop Olsen, Markussen, & Lau Rakêt, 2018; Permanasari, Harahap, & Ali, 2019) and has become a commonly used method by which different time series can be compared. This allows for the impact of differing speeds and accelerations within the data to be minimized.

The use of the DTW method was chosen as a way to cope with the contextual differences inherent in the approach/landing phase of any flight. While we have limited the length of each data set, how the pilots chose to use that time is entirely dependent upon the externalities which the pilots encountered during their particular approach (i.e., head/tail wind, gusty conditions, etc.). Such contextual differences make differentiating between two outcomes even more difficult, especially given the highly similar distributions of the two groups shown in Table II. DTW works by comparing and then aligning different time sequences so that the linearity of time is not necessarily maintained, thereby minimizing the distance between two (or more) time series (Ratanamahatana & Keogh, 2005).

4.4. Data Analysis

In order to determine which statistical tests are most appropriate for the collected data, an abbreviation of the preanalysis protocol outlined by Pallant (2011) was followed. The revised protocol checked the data to be analyzed for the following attributes: random sampling, independence of observations, and normal distribution of data point values (Pallant, 2011, pp. 205–206). The rarity of event data requires a design that precluded a completely random sampling of data (since such a sample would be very unlikely to include any true hard landings\(^1\)). However, even though the individual flight sequences are assumed to be independent, the observations within each sequence are interdependent, since an observation at time \(t\) is at least partially dependent upon what occurred in observations \(t_1, t_2, t_3, \ldots t_n\).

When the analysis variables were examined for normality, several of the variables were found to be only roughly normally distributed, which led to the decision to use the more robust non-parametric Mann–Whitney \(U\) test instead of the more common parametric \(T\)-test (which was used by Wang et al., 2018, 2014). The Mann–Whitney \(U\) test is based on rank sums of two different distributions (Johnson, Miller, & Freund, 2018). The observations of the two distributions are ranked with the result being compared to a critical value at less than equal to 0.05. However, while statistical differences are interesting, they do not necessarily indicate practical differences (Rodgers, 2010; Kanji, 2006).

To assess if a practical difference exists between the two groups, Cohen-\(D\) effect size and overlap coefficient tests were also conducted to determine the magnitude of their difference and to what degree the two distributions overlap. The Cohen-\(D\) is a well-known measure of effect size that “...presents difference between groups in terms of standard deviation units” (Pallant, 2011, p. 210). The overlap coefficient is the degree of overlap that two distributions have (Inman & Bradley Jr, 1989). Instead of comparing point estimates, as the Cohen-\(D\) does, the overlap coefficient assesses the degree two distributions share value ranges (Goldstein, 1994).

Due to the exploratory nature of these constructs, and to more clearly understand any temporal patterns within the data, a more qualitative visualization-based analysis approach was also used. While such qualitative differences do not necessarily indicate either statistical or practical significance, given the number of flights being analyzed, any visual differences could indicate potential patterns in the data. The visual analysis utilized line plots, histograms, and 1-D/2-D Kernel density estimation (KDE) plots. The line plots were used to plot individual flight sequences over time and the histograms and 1-D KDE plots were used to compare the differences between the observed values and generalized distributions between the two landing groups. The lesser known 2-D KDE plots were used to show the density of points for each \(X/Y\) coordinate pair, creating a generalized “contour map” (or “landscape”) which shows how the value distributions change over time.

The final analysis conducted in this study compared the diagnostic odds ratios of the cluster

\(^1\)True hard landings being defined as landings with a maximum normal acceleration \(\geq 1.7\ \text{Gs}^\)
analysis to determine how well the constituent and construct variables could predict a firm landing. The diagnostic odds ratio is calculated as follows (Šimundić, 2009, p. 209):

\[
\text{Diagnostic Odds Ratio} = \frac{\text{True Positive}/\text{False Negative}}{\text{False Positive}/\text{True Negative}}
\]

The diagnostic odds ratio was used to assess the accuracy of the differentiation prediction and was calculated by the clustering algorithm or based on a selected threshold value. For the constituent and construct variables, this was the mean of the two group’s means. This ratio gives us the opportunity to directly compare the ability of each variable, construct, or cluster to correctly predict the outcome at any given moment during the approach, the higher the ratio the higher the differentiation ability. Another advantage of this approach is that it does not depend upon equal prevalence between true positive/negatives and false positives/negatives, meaning that the approach can be used when events are extremely rare.

5. RESULTS AND ANALYSIS

5.1. Differentiation Between Landing Groups by Observation

Two sets (one for each runway) of independent-sample Mann–Whitney U, Cohen-D effect size, and overlap coefficient tests were conducted for the following seven time windows: 300, 200, 100, 50, 25, 15, and 5 seconds before touchdown, the results of which can be seen in Table III and Figs. 1 and 2. The results showed that the majority of the tests throughout the time windows for ivv, bank, pitch, and computed_airspeed were not significant when a Bonferroni correction for inflated alpha levels was applied. Moreover, the distributions had relatively low effect sizes and high amounts of overlaps. The variable n1_sum was particularly interesting since despite having several statistically significant instances, the effect sizes remained small and had high degrees of overlap between the normal and firm distributions. The remaining two variables and
| Time Until Touchdown | Statistic | p_value | cohen_d | Overlap |
|---------------------|-----------|---------|---------|---------|
| Runway 01           |           |         |         |         |
| ivy                 | 14,166    | 0.331   | −0.052  | 0.930   |
|                     | 14,326    | 0.389   | 0.039   | 0.928   |
|                     | 14,380    | 0.409   | −0.003  | 0.993   |
|                     | 14,227.5  | 0.353   | 0.001   | 0.997   |
|                     | 12,253    | 0.011   | 0.230   | 0.544   |
|                     | 13,914    | 0.247   | 0.142   | 0.848   |
|                     | 13,152.5  | 0.077   | 0.198   | 0.541   |
| bank                | 14,376.5  | 0.408   | −0.029  | 0.939   |
|                     | 12,152.5  | 0.008   | −0.256  | 0.268   |
|                     | 13,552.5  | 0.150   | 0.086   | 0.241   |
|                     | 14,163    | 0.330   | −0.055  | 0.687   |
|                     | 14,162.5  | 0.320   | −0.091  | 0.610   |
|                     | 12,817.5  | 0.040   | −0.319  | 0.460   |
|                     | 13,368.5  | 0.112   | 0.182   | 0.746   |
|                     | 13,432.5  | 0.124   | 0.155   | 0.794   |
| pitch               | 12,823    | 0.040   | −0.013  | 0.980   |
|                     | 12,993    | 0.057   | −0.332  | 0.572   |
|                     | 13,887.5  | 0.239   | −0.208  | 0.780   |
|                     | 11,908.5  | 0.004   | −0.296  | 0.660   |
|                     | 10,238    | 0.000   | −0.447  | 0.544   |
|                     | 12,142.5  | 0.008   | −0.286  | 0.640   |
|                     | 10,452    | 0.000   | −0.464  | 0.540   |
|                     | 13,353    | 0.150   | 0.125   | 0.849   |
|                     | 13,222.5  | 0.087   | 0.130   | 0.813   |
|                     | 13,432.5  | 0.124   | 0.155   | 0.794   |
| n1_sum              | 12,823    | 0.040   | −0.013  | 0.980   |
|                     | 12,993    | 0.057   | −0.332  | 0.572   |
|                     | 13,887.5  | 0.239   | −0.208  | 0.780   |
|                     | 11,908.5  | 0.004   | −0.296  | 0.660   |
|                     | 10,238    | 0.000   | −0.447  | 0.544   |
|                     | 12,142.5  | 0.008   | −0.286  | 0.640   |
|                     | 10,452    | 0.000   | −0.464  | 0.540   |
|                     | 13,353    | 0.150   | 0.125   | 0.849   |
|                     | 13,222.5  | 0.087   | 0.130   | 0.813   |
|                     | 13,432.5  | 0.124   | 0.155   | 0.794   |
| computed_airspeed   | 12,553    | 0.022   | 0.262   | 0.655   |
|                     | 12,435    | 0.017   | 0.289   | 0.454   |
|                     | 13,982.5  | 0.269   | 0.095   | 0.859   |
|                     | 12,584.5  | 0.024   | 0.220   | 0.695   |
|                     | 13,218    | 0.087   | 0.144   | 0.806   |
|                     | 13,984.5  | 0.269   | 0.032   | 0.957   |
|                     | 13,105.5  | 0.071   | 0.119   | 0.812   |
| flaps_actual        | 13,354    | 0.096   | −0.206  | 0.833   |
|                     | 12,869    | 0.031   | −0.281  | 0.669   |
|                     | 12,500.5  | 0.014   | −0.312  | 0.725   |
|                     | 10,404    | 0.000   | −0.738  | 0.635   |
|                     | 10,404    | 0.000   | −0.738  | 0.635   |
|                     | 10,404    | 0.000   | −0.738  | 0.635   |
| gross_weight        | 10,294.5  | 0.000   | −0.477  | 0.421   |
|                     | 10,306.5  | 0.000   | −0.476  | 0.422   |
|                     | 10,336    | 0.000   | −0.475  | 0.422   |
|                     | 10,338    | 0.000   | −0.475  | 0.422   |
|                     | 10,344    | 0.000   | −0.475  | 0.423   |
|                     | 10,330.5  | 0.000   | −0.475  | 0.422   |
|                     | 10,327    | 0.000   | −0.475  | 0.423   |
| row_mean            | 12,151    | 0.008   | −0.279  | 0.210   |
|                     | 10,901    | 0.000   | −0.458  | 0.023   |
|                     | 10,761    | 0.000   | −0.426  | 0.057   |

(Continued)
the two created constructs all had statistically significant differences as well as practical differences indicated by medium to large effect sizes and low amounts of distribution overlap. These results were not unexpected, since the two constructs are based upon aggregating the seven selected variables, several of which had significant differences. However, the constructs still generally outperform any of the selected variables, meaning that the additional information contained in the construct is able to help more clearly separate the landing groups prior to touchdown.

To get a higher resolution of the differences between the normal and firm landing groups throughout the approach to landing phase, 2-D KDE plots for each of the seven selected variables and two constructs were created (Fig. 3). In viewing the data as a 2D-KDE plot, it is possible to see the peaks and valleys of the two groups relative to one another. It is also a convenient way to quickly determine if the variables show any differences that are large enough to be considered practically significant. While large qualitative differences do not appear within this data set, of all the tested variables, the row_mean and row_sequence constructs do appear to have the least amount of overlap, especially as the time to touchdown approaches zero.

5.2. Temporal Differentiation

In order to test the criterion validity of the row_mean and row_sequence constructs, the aggregated values (with a 95% confidence interval) were plotted for the entire approach (Fig. 4). While the values for the row_mean construct do overlap and remain relatively close together for the majority of the approach until about 75 seconds prior to touchdown, the row_sequence construct remains distinct for almost the entire approach. Such descriptive evidence is promising as it shows a relevant separation of the landing paths over 300 seconds with minimal mean confidence interval overlap across the point estimates of the time series. Findings which are quite promising but represent an aggregate and not the ability to create a specific prediction.

Furthermore, when a diagnostic odds ratio was calculated for each of the constitute and construct variables 300, 200, 100, 50, 25, 15, and 5 seconds to touchdown both row_mean and row_sequence generally outperformed the constitute variables, as shown in Table IV.

In order to determine if the two constructs can create the insights needed by industry, a more granular analysis approach is needed (Fig. 5). While qualitatively it can be seen that both the row_mean and row_sequence constructs gravitate toward higher values, there is not a clear distinction between the two groups at the operational level. In order to see if this lack of distinction is due to the unbalanced nature of the sample or is representative of the actual variations within the operations, two edge cases were examined (Figs. 6 and 7).

The results of the two edge cases raises questions about the ability of the row_seq construct to create the insights needed by industry. The first case examines the 20 hardest and softest landings within the data set (Fig. 6). While there is a substantial amount of overlap between the two groups five minutes prior to touchdown, as the time until touchdown decreases, the two groups cleanly separate. Since that pattern

| Time Until Touchdown | Statistic | p_value | cohen_d | Overlap | Statistic | p_value | cohen_d | Overlap |
|----------------------|-----------|---------|---------|---------|-----------|---------|---------|---------|
| Runway 01            |           |         |         |         |           |         |         |         |
| −50                  | 9,184     | 0.000   | −0.748  | 0.010   | 15,106    | 0.000   | −0.722  | 0.021   |
| −25                  | 9,467     | 0.000   | −0.682  | 0.032   | 15,943    | 0.000   | −0.664  | 0.034   |
| −15                  | 9,280     | 0.000   | −0.728  | 0.019   | 14,471    | 0.000   | −0.774  | 0.014   |
| −5                   | 9,586     | 0.000   | −0.666  | 0.031   | 15,253    | 0.000   | −0.754  | 0.018   |
| row_sequence         |           |         |         |         |           |         |         |         |
| −300                 | 12,151    | 0.008   | −0.279  | 0.210   | 19,118    | 0.000   | −0.425  | 0.123   |
| −200                 | 10,975    | 0.000   | −0.421  | 0.009   | 17,497    | 0.000   | −0.529  | 0.003   |
| −100                 | 10,096    | 0.000   | −0.517  | 0.001   | 17,976    | 0.000   | −0.491  | 0.002   |
| −50                  | 9,643     | 0.000   | −0.588  | 0.000   | 17,069    | 0.000   | −0.547  | 0.001   |
| −25                  | 9,129     | 0.000   | −0.659  | 0.000   | 16,135    | 0.000   | −0.612  | 0.000   |
| −15                  | 8,982     | 0.000   | −0.679  | 0.000   | 15,778    | 0.000   | −0.634  | 0.000   |
| −5                   | 8,868     | 0.000   | −0.697  | 0.000   | 15,487    | 0.000   | −0.655  | 0.000   |

Table III. (Continued)
was not as clear in Fig. 4, a final edge case test was conducted.

The final test case examined the 20 hardest landings and the 20 most representative landings from the normal landing group (Fig. 7). The result showed similar mixing five minutes prior to touchdown, but the qualitative separation seen in Fig. 6 was not repeated as clearly. Instead, while the same general pattern is seen (firm landings gravitating toward higher \textit{row\_mean} and \textit{row\_sequence} values), the ability to differentiate the two groups based upon the value of either the \textit{row\_mean} or \textit{row\_sequence} has essentially disappeared.

5.3. Clustering Algorithm

In order to assess the efficacy of the created construct variables, the results in Fig. 4 were tested against those obtained using a \textit{k}\-means time series cluster analysis implementing the DTW metric. This clustering technique was applied to flights on approach to Runway 1 time series only. The application here is to benchmark across this AI derived method and search for convergence (confirmation) of the presented visual results differentiating normal and firm landings.

The \textit{k}\-Means clustering algorithm was indeed capable of identifying and reliably differentiating two clusters for both the \textit{row\_mean} and \textit{row\_sequence} constructs. Fig. 8 shows the individual time sequences in black; and the time series barycenters in red, which are the average of the cluster time series. Most importantly, Fig. 8 shows that it was possible to reliably separate the normal and firm landings into two clusters for both constructs. \textit{Row\_mean}_Cluster 1 could correctly predict normal landings \( \sim 63.5\% \) of the time, while \textit{row\_mean}_Cluster 2 predicted firm landings \( \sim 60.3\% \) of the time. Whereas the clusters created from \textit{row\_sequence} performed slightly worse, with \textit{row\_sequence}_Cluster 1 predicting firm landings \( \sim 57\% \) of the time, and \textit{row\_sequence}_Cluster 2 predicting
Fig 3. 2-D Kernel density estimation (KDE) plots for both landing groups (normal = blue, firm = red) for each variable and construct variable.
Fig 3. Continued
normal landings 62.2% of the time. This provides preliminary evidence showing that such clusters do not only differentiate but also contain opposite ratios of normal versus firm landings, a finding that aligns with the other results of this study.

6. DISCUSSION

The results of this study showed that both the row_mean and row_sequence constructs had some of the most significant and consistent Mann–Whitney U results, along with the largest effect sizes and smallest overlap coefficients of the variables analyzed, thereby answering the first two research questions posited in this article. The third question was answered when the results of the construct were compared with those of the time series k-means clustering algorithm. Those results indicate that these two basic aggregate constructs do generally outperform the constituent’s individual variables and is an important step forward in evaluating the operational risk of landing in a particular context. The data also shows that approaches like those proposed in this study are worthy of further exploration, as they could create important alternative perspectives from which safety and risk insights may arise since both the simple and more complex methods appear to converge on similar findings.

6.1. Construct Evolution

As an exploratory study, the two constructs developed in this article were the final iterations of several other attempts that need to be discussed briefly. The formulation of the row_mean construct was initially based on the absolute value of each selected variable that were then summed together. It was assumed that this would prevent the different
variables from “canceling” each other out by nullifying variables with opposing values leading to a reduction in the relational information within the data. For example, there is generally a negative correlation between \( ivv \) and \( pitch \) during an approach. This is because when the aircraft pitches up, the angle of attack of the wing increases. This results in an increase of the amount of lift created, thus increasing the vertical velocity. However, this assumption was proven wrong. The \( row_mean \) construct was underperforming relative to each individual variables' differentiation ability. The initial misstep proved to be quite serendipitous, since it stressed the importance of exploring and attempting to understand the underlying patterns within the data. Indeed, Rasmussen, Pejtersen, and Goodstein (1994) argue that such findings are just as, if not more, important than understanding individual variable relationships. This realization prompted us to modify the algorithm, creating the \( row_mean \) construct presented in this article.

The \( row_sequence \) construct was originally envisioned as a rolling average function. However, it quickly became evident that such a function would have a very arbitrary limit as to what observations were included and could not be theoretically justified. Instead, we decided to allow all the observations up to the moment of analysis to be included so that any temporal interdependencies and/or interconnections could be accounted for. This choice did have the unfortunate effect of uniformly weighting the observations so that the first and last observations are considered equally relevant, which is rarely true, and will need to be addressed and refined in later iterations. In its original iteration, the \( row_sequence \) construct was calculated using the data collected for the entire approach all at once, but we decided that this was counter to the second research question posed by this study and reformulated the algorithm to do the calculations step by step to represent the incremental information increase which occurs during an actual approach. Though the \( row_sequence \) construct is not sufficiently refined to address industry needs, this research does highlight a path forward. Further studies need to be done to critically validate and evaluate currently used metrics, as well as working to develop new ones with the ability to assess individual flight sequences in relation to operational risk.

### 6.2. Scrutiny

While the construct metrics cannot yet differentiate between landing groups with a sufficient degree of accuracy or precision to be implemented operationally in their current form, the results are nonetheless promising. The emphasis on simple and easily interpreted construct metrics is simultaneously this study’s greatest weakness and greatest strength. However, the proposed simplified metrics context pattern matching is similar to the processing that is continually occurring in a pilot’s mental model throughout the approach, which makes them a valuable tool in trying to assess real time operational risk.

---

**Table IV.** Diagnostic Odds Ratios of a Firm Landing Given the Mean of the Means of Each Selected Variables at 300, 200, 100, 50, 25, 15, and 5 Seconds Before Touchdown

| Time Until Touchdown | \( ivv \) | \( bank \) | \( pitch \) | \( n1_{sum} \) | \( computed_{airspeed} \) | \( flaps \) | \( gross_weight \) | \( row_mean \) | \( row_sequence \) |
|----------------------|---------|---------|---------|-------------|-----------------|--------|----------------|-------------|--------------|
| Runway 01            |         |         |         |             |                 |        |                |             |              |
| −300                 | 1.047   | 0.787   | 1.559   | 1.291       | 0.739           | 1.053  | 1.988          | 1.655       | 1.655        |
| −200                 | 1.109   | 1.666   | 1.490   | 1.897       | 0.637           | 1.319  | 1.988          | 2.182       | 2.436        |
| −100                 | 0.971   | 0.710   | 1.181   | 1.486       | 0.834           | 1.346  | 1.988          | 2.182       | 3.194        |
| −50                  | 1.036   | 1.104   | 0.868   | 2.118       | 0.590           | 4.481  | 1.988          | 3.752       | 2.706        |
| −25                  | 0.508   | 1.098   | 0.921   | 2.446       | 0.681           | 4.481  | 1.988          | 2.730       | 2.602        |
| −15                  | 0.822   | 1.354   | 0.633   | 1.963       | 0.927           | 4.382  | 1.988          | 3.056       | 2.706        |
| −5                   | 0.780   | 0.804   | 0.834   | 3.378       | 0.597           | 4.382  | 1.988          | 2.739       | 2.643        |
| Runway 19            |         |         |         |             |                 |        |                |             |              |
| −300                 | 0.855   | 1.458   | 1.436   | 1.634       | 0.542           | 2.420  | 3.803          | 1.950       | 1.950        |
| −200                 | 1.041   | 1.069   | 1.298   | 1.601       | 0.707           | 1.638  | 3.768          | 1.728       | 3.029        |
| −100                 | 0.676   | 1.409   | 1.155   | 0.931       | 1.411           | 0.776  | 3.699          | 1.572       | 2.332        |
| −50                  | 0.862   | 1.951   | 0.701   | 0.830       | 0.958           | 4.572  | 3.699          | 2.864       | 2.332        |
| −25                  | 0.956   | 1.106   | 0.547   | 1.030       | 0.769           | 4.674  | 3.699          | 1.897       | 2.787        |
| −15                  | 0.659   | 1.782   | 0.755   | 1.029       | 0.967           | 4.674  | 3.699          | 2.822       | 2.662        |
| −5                   | 0.848   | 1.079   | 0.476   | 2.613       | 0.885           | 4.674  | 3.699          | 2.674       | 2.959        |
Further scrutiny of our data and approach identified three additional limitations. First, there is a lack of normality of our data. This does technically violate an underlying assumption of many parametric statistics (which is why we opted instead for the more robust Mann–Whitney \( U \) test) but does not pose a major issue given the tenants of the central limit theorem. However, since we are examining a system that is changing with time (both in terms of aircraft configuration and the surrounding environment) we see little benefit in assuming that parametric methods will be appropriate for the wide range of conditions encountered in airline operations. We feel that this is an especially important point if the data under analysis is highly contextualized, as the random stochastic noise that would be expected of larger more varied data set will not exist. For this reason, it was decided to maintain a conservative approach throughout the data analysis.

The second limitation was the firm/normal event ratio in our data set. When comparing this work to a somewhat similar study done by Wang et al. (2018), the \(~0.83\) hard to normal ratio Wang et al. had in their data set was far higher than the \(~0.24\) firm to normal ratio in this study. While both studies used conservative thresholds, this made the event to non-event ratio’s much higher when using more standard threshold cutoffs; for our data set the reported hard landings to normal landings was \(~0.004\). It is for this reason that we have attempted to develop an individual flight focused metric, rather than rely upon the aggregated differences that others have found (see Oehling & Barry, 2019; Wang et al., 2018; Mugtsusids, 2000 for examples of some aggregate focused analyses).

The final major limitation of this work was our inability to cross-validate our sample set with that of other contexts. We split our data by runway, and,
6.3. Future Work

As part of a larger project series, the results from this article will be further explored. First, we plan to refine the variable aggregation algorithm so that certain variables are treated as more important (i.e., pitch, bank, computed_airspeed), as these variables represent the current immediate state of the aircraft. The next step will be to introduce a temporal minimization function so that the oldest values in each sequence have less relative weight than those observed more recently. Such an approach will draw heavily upon applications of generalized linear mixed models and will be used to test and model the variance-covariance between the variables over time. The final step will be to introduce a dynamic weighting function that will emphasize larger deviations, as such instances are unusual and therefore of greater interest. By emphasizing the larger deviations and taking into account the temporal minimization weighting we anticipate that differences between the two groups on an individual level will start to emerge more clearly. This anticipated increase in differentiation ability is based upon the reasoning that recent extreme inputs from one or more variables are likely indicative of some externality that must be overcome. In cases where pilots do not recognize a potentially abnormal event, an effect size or overlap coefficient-based method could prove to be a relevant classification approach.
criterion upon which to base warnings/alerts. This call for attention would help the operators address any externalities faster and more reliably by evaluating the context and then using their expertise.

Additional work already being undertaken will pursue the use of ML ensemble models to better differentiate between different outcome groups. Though these approaches will require additional competencies to create and interpret, they will very likely be instrumental in creating new safety insights in the future, since they have the ability to detect patterns invisible to humans (assuming such methods remain interpretable). Furthermore, while some algorithms will not be able to produce real-time data, some preliminary work has already showed that the ensemble ML algorithms if properly set up can produce results in less than five seconds using a relatively small amount computational power that could easily be installed on aircraft today. Given the preliminary results from that work we believe that there is a great deal of potential use of such algorithms by managers/training staff as well as those conducting the actual real-time operations, again with the caveat that the algorithms being used are interpretable.
7. CONCLUSIONS

While the exceedance approaches examined at the beginning of this article will continue to play an invaluable role in ensuring safe operations, they are generally limited to being used to assess the likelihood that an event will reoccur given a set of parameters, rather than whether a deviation is occurring that could lead to a novel event. The purpose of this study was to develop and test a set of metrics to complement the current exceedance-based methods. To do this, two construct variables were constructed (row_mean and row_sequence) and tested for their ability to differentiate between normal and firm landings prior to touchdown. While the constructs did generally outperform the individual variables in the aggregate, their ability to proactively/predictively identify differences at the individual operational level was not as clear. To further analyze the effectiveness of the selected approach, we also analyzed the data using a Time Series k-means clustering algorithm, which was also able to effectively differentiate between normal and firm landings at the aggregate level. Furthermore, the clustering algorithm was shown to have potential predictive value in differentiating between normal and firm landings.

This study shows that the types of metrics developed in this article have potential to improve risk information and add value to both operators and managers. While operators need direct feedback and support to ensure that learning experiences are maximized, managers need support in knowing how to allocate resources to create the most value. For operators, the individualized sequence metrics can help identify abnormal patterns which can then be used to provide pilots with feedback on how normal their approaches are. Such information could also be used to help personalize recurrent training and ensure that each individual pilot is able to improve upon areas of weakness. Safety managers could use the aggregated sequence plots (e.g., Fig. 4) to ensure that the differentiation between the two groups is not changing and if it is, to try and determine why such a change is occurring. Furthermore, an airline’s safety office could use the overlay of individual flights on top of a 2-D KDE plot of normal flights for that context to determine if a major deviation occurred when investigating an event. Together the approaches discussed in this article are complementary to current risk management approaches and present exciting avenues for further study.

ACKNOWLEDGMENTS

The authors would like to thank the reviewers for helping us clarify our message, our partner airline for providing us with the data used within the paper, and the members of the FP7 PROSPERO project for their insights which helped formulate the ideas presented above.

REFERENCES

ATR. (2016). Flight data monitoring: On ATR aircraft. Blagnac, France: ATR.
Baranzini, D., & Zanin, M. (2015). Risk prediction and risk intelligence in aviation—the next generation of aviation risk concepts from PROSPERO FP7 Project. Proceeding of 25th European Safety and Reliability Conference, September 7—10, Zurich, Switzerland.
Civil Aviation Authority [CAA]. (2013). CAP 739: Flight data monitoring. London, UK: The Stationery Office.
Das, S., Li, L., Srivastava, A., & Hansman, R. J. (2012). Comparison of algorithms for anomaly detection in flight recorder data of airline operations. Proceedings from 12th AIAA Aviation Technology, Integration, and Operations (ATIO) Conference and 14th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference, Reston, Virigina.
Distefano, N., & Leonardi, S. (2018). Aircraft runway excursion features: A multiple correspondence analysis. Aircraft Engineering and Aerospace Technology, 91. https://doi.org/10.1108/AEAT-11-2017-0244/full.html
Fernández, A., Martínez, D., Hernández, P., Cristóbal, S., Schweiger, F., Nunez, J. M., & Ruiz, J. M. (2019). Flight data monitoring (FDM) unknown hazards detection during approach phase using clustering techniques and AutoEncoders. Paper presented at 9th SESAR Innovation Days December 2—5, Athens, Greece.
Goldstein, R. (1994). The overlapping coefficient and an “improved” rank-sum statistic. Stata Technical Bulletin, 22, 12–15.
Hagge, J., & Rokseth, B. (2020). Applications of machine learning methods for engineering risk assessment – A review. Safety Science, 122, 104492. https://doi.org/10.1016/j.ssci.2019.09.015
International Civil Aviation Organization (ICAO). (2013). Flight data analysis programme manual. Quebec, Canada: International Civil Aviation Organization.
International Civil Aviation Organization [ICAO]. (2018). Safety management manual. Quebec, Canada: International Civil Aviation Organization.
Inman, H. F., & Bradley, Jr, E. L. (1989). The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities. Communications in Statistics-Theory and Methods, 18(10), 3851–3874.Retrieved from http://scholar.google.com/scholar?output=instlink&nors=1&q=inforRV_5fIO9k92UU&scholar.google.com/chi=en&as_sd=1.5&scillep=3161915846228853349&oi=lle
Ives, Z. G. (2016). Technical perspective: K-shape: Efficient and accurate clustering of time series. ACM SIGMOD Record, 45(1), 68–68.Retrieved from https://dl.acm.org/doi/pdf/10.1145/2949741.2949757
Johnson, R. A., Miller, I., & Freund, J. (2018). Miller & Freund’s probability and statistics for engineers (Ninth Global Edition ed.). London, UK: Pearson Higher Education.
Kanji, G. K. (2006). 100 Statistical Tests (3rd Edition ed.). Sage Publications.
Leveson, N. G. (2011). Engineering a safer world: Systems thinking applied to safety (engineering systems). Cambridge, MA: The MIT Press.
Lundtorp Olsen, N., Markussen, B., & Lau Rakêt, L. (2018). Simultaneous inference for misaligned multivariate functional data. arXiv, arXiv: 1606.03295. Retrieved from https://adsabs.harvard.edu/abs/2016arXiv160603295L/abstract
Molnar, C. (2020). Interpretable machine learning. Retrieved from https://christophm.github.io/interpretable-ml-book/
Mugtussids, I. B. (2000). Flight Data Processing Techniques to Identify Unusual Events. Virginia Tech.
Oehling, J., & Barry, D. J. (2019). Using machine learning methods in airline flight data monitoring to generate new operational safety knowledge from existing data. Safety Science, 114, 89–104. https://doi.org/10.1016/j.ssci.2018.12.018
Pallant, J. (2011). SPSS survival manual: A step by step guide to data analysis using IBM SPSS. Retrieved from http://books.google.se/books?id=vHLlmgEACAAJ&hl=&source=gbs_api
Permanasari, Y., Harahap, E. H., & Ali, E. P. (2019). Speech recognition using dynamic time warping (DTW). Journal of Physics: Conference Series, 1366, 012091.
Rasmussen, J., Peijtersen, A. M., & Goodstein, L. P. (1994). Cognitive systems engineering. Wiley-Interscience. Retrieved from https://www.amazon.com/Cognitive-Systems-Engineering-Jens-Rasmussen/dp/0471011983
Ratanamahatana, C. A., & Keogh, E. (2005). Three myths about dynamic time warping data mining. In Proceedings of the 2005 SIAM International Conference on Data Mining, Philadelphia, PA: Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611972757.50
Rodgers, J. L. (2010). The epistemology of mathematical and statistical modeling: a quiet methodological revolution. American Psychologist, 65(1), 1.
Schilling, M. F. (1990). The longest run of heads. College Mathematics Journal, 21(3), 196–207. Retrieved from https://www.tandfonline.com/doi/pdf/10.1080/07468342.1990.11973306
Šimundič, A.-M. (2009). Measures of diagnostic accuracy: Basic definitions. EJIfcc, 19(4), 203. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3975285
Stolzer, A. J., Halford, C. D., & Goglia, J. J. (2015). Safety management systems in aviation (Second ed.). Surrey, UK: Ashgate Publishing. Retrieved from http://books.google.se/books?id=6nVxCGAAQBAJ&hl=&source=gbs_api
Ulffengren, P., McDonald, N., Kontogiannis, T., Malakis, S., Cacciabue, C., Cassani, M., … Frau, G. (2013). Requirements for proactive safety performance for operations at organizational & systemic levels (D2.2). Retrieved from https://cordis.europa.eu/project/id/314822/reporting
Wagner, D. C. S., & Barker, K. (2014). Statistical methods for modeling the risk of runway excursions. Journal of Risk Research, 17(7), 885–901. https://doi.org/10.1080/13669877.2013.822913
Walker, G. H. (2017). Redefining the incidents to learn from: Safety science insights acquired on the journey from black boxes to flight data monitoring. Safety Science, 99, 14–22. https://doi.org/10.1016/j.ssci.2017.05.010
Wang, L., Ren, Y., & Wu, C. (2018). Effects of flare operation on landing safety: A study based on ANOVA of real flight data. Safety Science, 102, 14–25. https://doi.org/10.1016/j.ssci.2017.09.027
Wang, L., Wu, C., Sun, R., & Cui, Z. (2014). An analysis of hard landing incidents based on flight QAR data.