Transportation Research Part C 129 (2021) 103262

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Modeling go-around occurrence using principal component logistic regression

Lu Dai *, Yulin Liu, Mark Hansen

Department of Civil and Environmental Engineering, University of California, Berkeley, CA 94720, USA

ARTICLE INFO

Keywords:
Go-around
Principal component regression
Factor analysis
Counterfactual analysis
Flight approach and landing

ABSTRACT

A go-around is an aborted approach of an aircraft. We model go-around occurrence using Principal Component Logistic Regression (PCLR). This entails go-around detection, feature engineering, and model estimation. As a case study, we consider John F. Kennedy (JFK) International Airport arrivals, and model go-around occurrence based on information available when the subject flight is five nautical miles from its landing runway threshold. The PCLR model is based on Principal Component Analysis (PCA) for analyzing data that suffer from multi-collinearity. The model provides a representation of the empirical relationship between go-around occurrence and Principal Components (PCs) covariates, which encompass flight approach features, aircraft characteristics, flight lead-trail spacing, surface operation, go-around clustering effect, airport and weather conditions. We use factor loading analysis to reveal the relationship between variables and the PCs they formed. Coefficient estimates of PCs are also transformed back to the scale of the original variables by matrix operations. A counterfactual analysis is employed to assess the importance of different features, and reveals that the stability of an approach, flight lead-trail spacing, departure traffic, and ceiling are the most salient factors affecting go-around occurrence.

1. Introduction

One of the primary goals of all air navigation service providers (ANSPs) is ensuring safety. Although the National Airspace System (NAS) is one of the safest and most efficient transportation infrastructures, growing air traffic demand and the implementation of autonomous NextGen technologies place strain on NAS safety and efficiency. In the past decades, 56% of the fatalities and 62% of the accidents occurred during the final approach and landing stages, accounting for only 16% of the airborne time (Boeing Commercial, 2018). Typically, the approach-and-landing accident is triggered by an unstabilized approach and is the consequence of a subsequent failure to initiate a go-around, which is an aborted landing process that is initiated when proceeding with the landing is considered to be unsafe.

A go-around can be initiated by either the pilot or the controller to abort the landing of an aircraft that is on the final approach due to conditions such as wind shear, runway incursion, and unstabilized approach. From 2012 to 2017, the average percentage of final approaches leading to go-arounds across the core 30 U.S. airports is 0.4% (FAA, 2019). While a go-around is a risk mitigation tool intended to increase flight safety, it is also an operational anomaly that significantly increases air traffic controller workload (Jou et al., 2013) and noise (Prats et al., 2010), while degrading airport throughput (Shortle and Sherry, 2013) and flight on-time performance (Blajev and Curtis, 2017). For enhancing both safety and efficiency of the aviation system, the effectiveness of collaboration between

* Corresponding author at: 107D McLaughlin Hall, University of California, Berkeley, CA 94720-1720, USA.
E-mail address: dailu@berkeley.edu (L. Dai).

https://doi.org/10.1016/j.trc.2021.103262
Received 5 August 2020; Received in revised form 1 May 2021; Accepted 7 June 2021
Available online 22 June 2021

pilots and air traffic controllers in making go-around decisions based on their anticipation of landing conditions is of great importance and practical significance. However, interviews suggest that the collaborative decisions on go-arounds are strongly influenced by individual experiences and mental states, as opposed to the collective knowledge about the complete picture of underlying reasons for the go-around procedure (Blajev and Curtis, 2017).

The purpose of this paper is to quantitatively assess the factors that contribute to go-around occurrences. The motivation of this paper is threefold. First of all, our study can provide flight crews, air traffic controllers, and other decision makers with better knowledge of the conditions in which a go-around is more likely to be executed. Second, quantifying the contributing factors of go-around occurrence can help identify countermeasures to reduce go-arounds, and more generally the conditions that give rise to them, which may be considered anomalous states that are inherently undesirable. Mitigation strategies can be developed to reduce the go-around occurrences through procedure modification, pilot training, and equipment design. Finally, our research may also inform efforts to develop a real-time tool that can identify, and perhaps remediate, situations in which there is a substantial risk of a go-around.

This paper builds upon the earlier work (Dai et al., 2019) to further understand how a mix of traffic and environmental conditions affect the go-around decisions, and quantify the underlying factor contributions on go-around occurrence based on actual flight data. In this study, we employ statistical models that estimating the impact of a large set of situational and environmental conditions on go-around occurrence. Toward this end, we first develop a go-around detection algorithm which has then been applied to the JFK arrival flight track dataset in 2018. Second, we develop a large set of features that may affect go-around occurrence, including aircraft characteristics (weight class, operating airline, landing runway), flight approach features (localizer deviation, speed, flight energy), the occurrence of other go-arounds at about the same time, in-trail separation features (loss of separation, speed difference, altitude difference), as well as features pertaining to surface operations, airport configuration, and local weather. Third, we build and compare both a standard logistic regression model and a principal component logistic regression model (PCLR) to establish statistical relations between the derived features and go-around occurrence. Lastly, we used the estimated PCLR model to construct counterfactual scenarios to estimate the contributions of different factors to go-around occurrence.

The remainder of this paper is organized as follows. In Section 2, we review previous work on go-arounds with emphasis on work that investigates how go-around decisions were made and factors that trigger the occurrence of go-arounds. Section 3 introduces the go-around detection algorithm. Section 4 describes our data sources and feature engineering methods. Section 5 presents the PCLR model and the results of counterfactual analysis. Section 6 offers the conclusions and discussions of the implication and limitations of the current study.

2. Literature review

The mainstream literature related to go-arounds has focused on the behavior and performance of pilots and controllers when a go-around occurs. From the pilot's point of view, Causse et al. (2013) found that the negative emotional consequences attached to the go-around – the uncertainty of a decision outcome and the reward/punishment – can temporarily jeopardize pilot decision making and cognitive functioning, while (Dehais et al., 2017) examined the errors in pilot's flying performance (e.g., flightpath deviations) and visual scanning behaviors during go-around execution. From the air traffic controller's point of view, (Jou et al., 2013) point out that controllers' failure to maintain situational awareness was the leading cause of Taiwan's go-around incidents in 2010. (Kennedy et al., 2010) found that controller age and expertise have significant impacts on aircraft landing decision making during a flight simulation task.

Another area of study is criteria that should be used by controllers and or pilots for deciding whether to initiate a go-around. The Flight Safety Foundation (Blajev and Curtis, 2017) developed surveys and interviews to identify four groups of factors that were most influential to the decision of a go-around: flight path profile, aircraft configuration, flight energy, and environmental conditions. (Campbell et al., 2018a; 2018b) developed go-around criteria in terms of airspeed, glideslope deviation, localizer deviation, and rate of descent, at different starting altitudes from which pilots cannot successfully recover from an unstabilized approach on full-flight simulators. They found that the airspeed and localizer deviations impact go-around occurrence the most. To further validate their proposed go-around criteria, (Campbell et al., 2019) collected objective simulation data and subjective post-simulation written questionnaires to check the stability of the 300-feet decision gate. Later (Zaal et al., 2019) evaluated the effects of environmental conditions on the proposed go-around criteria using statistical tests and decision tree analysis. Wind speed, visibility, and localizer deviation substantially affect the go-around decision making and perception of risk. The study suggests that certain environmental conditions might warrant altered decision thresholds of the go-around criteria.

While the criteria above suggest some of the factors that influence go-around occurrence, other research has addressed this question through statistical analysis of historical flight data. (Shepherd et al., 1997) presented an in-trail runway occupancy scenario event tree model that accounts for the risks associated with the go-arounds due to multiple runway occupancy and runway incursions. The go-around execution probability and the go-around failure probability are calculated. Surveillance track data is utilized in more recent research. (Sherry et al., 2013) detected go-arounds (termed as aborted approaches in the paper) using the cumulated change of aircraft heading angles. The go-around rate of the Chicago O'Hare International airport (ORD) airport was reported as 0.74% with some false positives resulting from procedure turns or other normal maneuvers that meet the quantitative criteria but are not aborted approaches. The go-around detection algorithm based on cumulative turn angle is a valuable contribution, but may not be fully representative. (Sherry et al., 2013) further reviewed 467 voluntary Aviation Safety Reporting System (ASRS) reports to identify underlying factors leading to aborted approaches. They classified the factors as airplane issues (48%), traffic separation issues (27%), weather (16%), runway issues (5%), and crew-ATC interaction issues (4%). These summary statistics afford valuable insights, but the ASRS reports are

voluntary and are made only when there is a perceived safety issue.

(Donavalli, 2016) identified go-arounds by simply checking whether the approaching aircraft crosses the end of the arrival runway. The go-around detection results were used in two-proportion Z tests to compare the go-around rates for different weather condition scenarios. High wind gust speeds and thunderstorms have a significant impact on go-around occurrence. However, possibly due to the limited variability of the 18-day data set, the Z-test did not find a significant visibility impact. The authors further developed a linear regression model that defines the proportion of daily go-arounds as the dependent variable, different weather factors as the independent variables. However, none of the variables were significant, suggesting that many factors other than weather conditions also contribute to the go-around occurrence. Without incorporating a wide range of situational conditions and environmental measures in the feature space, models would not be appropriate for making any concrete analysis and policy decisions.

(Deshmukh et al., 2019) identified go-arounds by looking at the violation of two linear regression lines, which are specified for the bounds of normal operations in terms of latitude, longitude, altitude, groundspeed, and aircraft energy. Each flight track was truncated into two parts. The detected go-around labels obtained from the last 5-timestep track, together with aircraft energy and separation features from the first 55-timestep track, were trained to classify whether a flight track is a go-around or not, given the first 55 timesteps. However, the paper neither considers the impacts of runway operation and weather impact, nor explains how the classification model could serve their purpose of identifying go-around precursors. In addition, given the truncation timestamp choice, a go-around could easily be initiated during the 55-timestep portion of the track, resulting in artificially high-performance evaluations. It is likely that the prediction performance would decline rapidly if the portion of the track used for performance was reduced.

While previous studies have yielded valuable insights about various causes of and contributors to go-around occurrence, none has developed a comprehensive, quantitative assessment of the relative importance of a wide range of factors that affect the likelihood of a go-around. That is the aim of the present paper. Toward this end, our paper makes several contributions. First, we designed and implemented a trajectory-based go-around detection algorithm and applied it to the JFK arrival flights in 2018. The go-around detection algorithm utilizes flight 4-D trajectory and employs multiple criteria based on theoretical and empirical analysis, rather than relying on a single criterion. Our detection method can be applied to any airport for which the surveillance track data are available. Second, we have fused multiple datasets in order to capture a wide range of factors that may contribute to the go-around occurrence. We have collected features from the dataset directly, but also used domain knowledge to derive features pertaining to approach stability, in-trail separation, and surface operations. Third, we have developed statistical relationships between go-around occurrence and the derived features using a principal component logistic regression (PCLR) model. This technique enables us to overcome the high dimensionality and multi-collinearity of the original data set while preserving the ability to assess the contribution of the original features to go-around occurrence.

3. Go-around detection algorithm

We did not find an algorithm that rigorously detects the go-around occurrence from flight track data in the open literature; current practice of go-around "detection" is mostly based on voluntary self-reports of controllers or pilots, which are typically unreliable and incomplete. Therefore, this paper proposes a scientific way of detecting go-around occurrence by analyzing historical flight trajectories.

According to (IATA, 2016), "A go-around begins when the crew aborts the descent to the planned landing runway during the approach phase; it ends after speed and configuration are established at a defined maneuvering altitude or to continue the climb for the purpose of the cruise." Fig. 1 compares the horizontal and vertical profiles between a normal flight approach (left) and a typical go-around flight (right). In these plots, the black curves show the altitude and blue curves show the horizontal distance away from the airport. A typical go-around flight would first decrease its altitude and distance to the landing runway threshold, then climb and fly

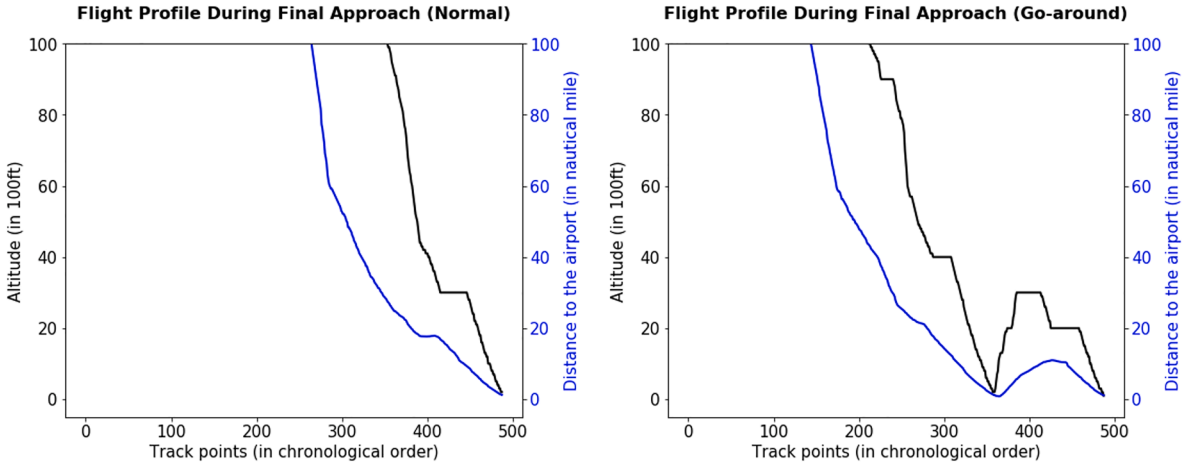

Fig. 1. Profiles of a normal landing flight (left) and a go-around flight (right).

away from the airport for another approach and landing. The detailed detection algorithm is presented in Table 1:

The trajectory-based go-around detection algorithm is designed to detect any aborted landing behaviors – including missed approaches and go-arounds – that are initiated within 10 nm of the airport, regardless of whether the flight overflies the runway, or whether it proceeds to the Missed Approach Points (MAPt). The definition of go-around in this study applies to both VFR aircraft and IFR aircraft, as long as the aircraft tries again for the landing.

In step 4, we did not directly use the landing runway information recorded in the given dataset due to a significant number of missing and incorrect records. For every two-minute analyzed segment extracted from Step 3, we calculate the distance from the extended centerlines of all the available runways (by configuration) using formulas (1), (2) and (3) in (Lee et al., 2007). Each track point votes for the closest Extended Runway Centerline (ERC) segment. The landing runway is the one that receives the most votes from track points in the vector. This runway is also used as one of the features in the statistical model.

We implemented our algorithm with consultations from subject matter experts and sensitivity analysis to determine parameters such as the altitude at the start of ascent h_{start} and the total altitude gain during the ascent Δh. We further applied it to the JFK arrival flights except for military flights, general aviation, and helicopters. Fig. 2 illustrates the JFK airport chart and the runway threshold of runway 22L as an example. We collected the coordinate information of the midpoint of all the runway threshold bars at JFK airport (marked as yellow for runway 22L in Fig. 2), and applied it as the input to the go-around detection algorithm described in Table 1. In this manuscript, the "runway threshold" specifically refers to the midpoint of the runway threshold bar.

Two of the detected go-around flight trajectories are visualized in Fig. 3. The flight in the left plot aborted the descent to the planned runway 22R, overflew the runway, returned to the landing queue, and finally landed on runway 22L. The flight in the right plot was planned to land on runway 4L, but it ended up landing on runway 22L (Note that this flight is counted as a go-around and its subsequent successful landing is not considered). In total, 433 go-arounds have been detected from July 1st to December 24th in 2018, which accounts for 0.43% of all JFK arrivals within the period. This statistic agrees with the FAA report, which indicates that the average percentage of go-around occurrence across the core 30 airports in the U.S. from 2012 to 2017 is 0.4% (FAA, 2019).

4. Feature engineering

To assess the impacts of variables on go-around occurrence, we consider only those features that can be evaluated before the initiation of a go-around. The available features thus depend on the distance of the go-around initiation point from the runway. In this study, we assume a go-around is initiated when its altitude starts increasing during the approach procedure based on our detection algorithm. We found that over 90% of detected go-arounds in our dataset occurred within 5 nm of the landing runway threshold. To obtain features that are proximate in time to go-around initiation, but without losing too many go-around observations, we choose 5 nm as the ***information cutoff gate***. Thus, a flight that initiated go-around at 4.5 nautical miles from the runway is included in the data set, while a flight that initiated a go-around at 5.1 nm from the runway is removed from the sample. In addition, we use the linear extrapolation technique to derive features at 5 nm from the runway threshold for each flight. In this study, we evaluate whether a flight initiates the go-around anywhere at [0, 5) nautical miles only based on the information available when this flight passes the 5 nm arc. The extrapolation guarantees that we do not include any information that cannot be known in the feature space when the aircraft is at

Table 1
Go-around detection algorithm.

Algorithm: Go-around Detection Algorithm
Procedure
INPUT: 4D flight track data (latitude, longitude, altitude, and time)
INITIALIZE: Coordinates of arrival runway thresholds
OUTPUT: Go-around labels and their related properties (execution time, execution altitude, etc.)
Step 1: Data preprocessing. Apply median filtering with a sliding window size of 10 entries to remove noise from the trajectory records. Remove incomplete trajectories (the altitude of the last track point is higher than 500 feet) and define landing endpoint (the rate of descent equals 0 feet) for complete trajectories.
Step 2: Altitude check. Piecewise linear regression is applied to identify points at which the slope of the altitude evolution curve is changed, as the black curve is shown in Fig. 1. Each flight trajectory is processed and must meet the following criteria to be considered as a go-around:
● The altitude at the start of ascent is no more than h_{start} (default value of 5500 feet);
● The total altitude gain during the ascent must not be less than Δh (default value of 400 feet).
Step 3: Define the analyzed segment. For flights that pass the altitude check in Step 2, the landing endpoint (defined in Step 1) is updated to the point at which the altitude starts increasing. Each aircraft's analyzed segment is a two-minute (T_{final}) trajectory segment ending at the landing endpoint.
Step 4: Calculate the landing runway. For every two-minute analyzed segment, calculate its landing runway.
Step 5: Distance check. For each track point of the two-minute analyzed flight trajectory segment, calculate the distance to the runway threshold markings of the corresponding landing runway obtained in Step 4. Piecewise linear regression is applied to identify points at which the slope of the distance to landing runway threshold evolution curve is changed, as the blue curve is shown in Fig. 1. Each flight trajectory is processed and must meet the following criteria to be considered as a go-around:
● When a go-around flight is within 1-nautical-mile range of the airport, its altitude does not exceed h_{1nm} (default value of 1500 feet);
● Go-around must occur within the 10-nautical-mile range of the airport, in order to distinguish go-arounds from aircraft being vectored or in holding patterns;
● The ascending segment of a go-around trajectory must intersect with a ten nautical-mile-radius cylinder centered at the airport.
Step 6: Multi-go-arounds. The two consecutive go-around procedures should be separated by at least 5 min (T_{multi}). The trajectory starting point for the second and subsequent flight trajectory segments is when the previous go-around trajectory segment starts ascending.
end procedure

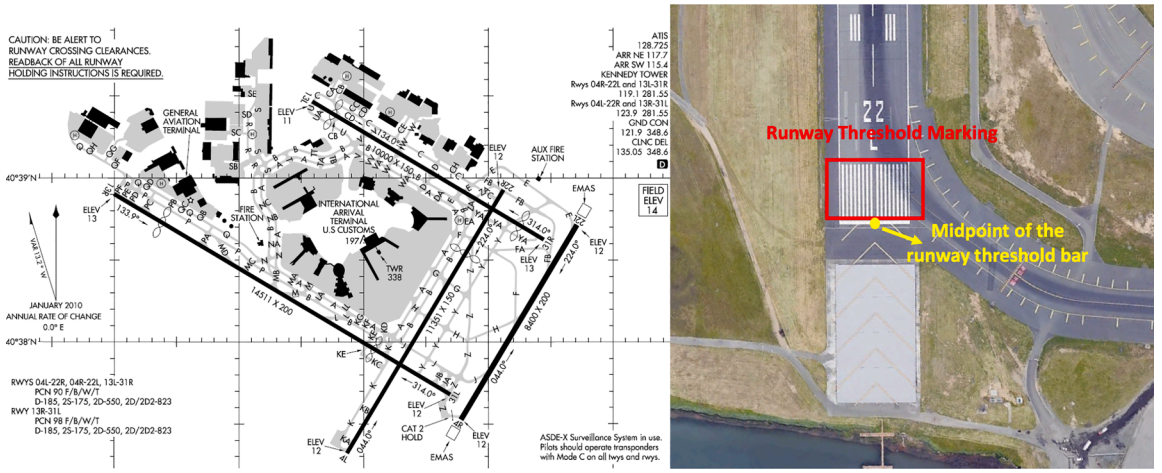

Fig. 2. JFK airport chart (left) and the runway threshold marking (right).

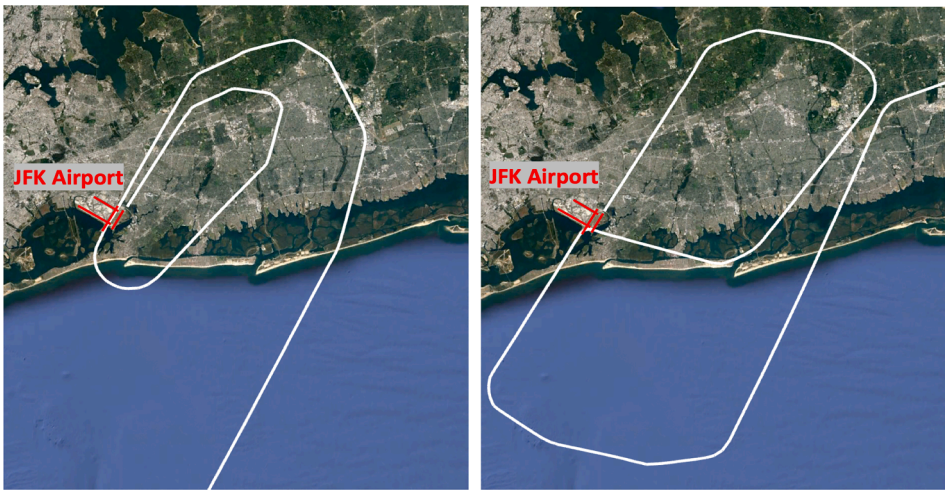

Fig. 3. Exampled go-around flight trajectories.

the 5 nm information cutoff gate. Notice that this distance-driven sampling strategy also guarantees that time-varying features are comparable for all the flights in the data set.

We derived six categories of features at the 5 nm information cutoff gate. These include: aircraft characteristics, flight approach features, in-trail separation features, airport and weather conditions, go-around clustering effect features, and surface operation features. Originally, 0.43% of JFK arrivals are detected as go-arounds within the period. After data preprocessing and matching flight trajectories with the 5 nm features, our final dataset has a total of 343 go-arounds initiated within 5 nm (5 nm is exclusive) from the landing runway threshold, which accounts for 0.34% of JFK arrivals between July and December of 2018. The following subsections summarize the process of feature extraction and its related functional modules. The variable code, variable description, and the summary statistics of all the features derived for flight *i* at 5 nm from the landing runway threshold can be found in Table 2 at the end of this section.

4.1. Data sources and meta features

We limit the scope of the study to JFK airport, and our datasets come from three sources ranging from July 1st to December 24th in 2018, except for four days with defective data.

The first dataset is retrieved from the Integrated Flight Format (IFF) and Reduced Data (RD) summary of the NASA Sherlock Data Warehouse, which are gathered from 76 FAA facilities and formatted by ATAC corporation. Fields of interest include flight summary (e.g., aircraft type, origin, destination), trajectory information (timestamp, latitude, longitude, altitude, ground speed, course, rate of climb, etc.), and landing information (e.g., runway threshold crossing time). Arrival trajectories have been filtered to 400 nautical miles centered on the analyzed airport for each flight. The RD summary and the IFF data have been further processed and merged on a

Table 2
Model variables and summary statistics.

Var. Category	Var. Code	Variable Description (when flight i is at 5 nm from the threshold)	Mean	Min	Max
Dependent Variable	GA	1 if flight *i* is detected as a go-around occurring within [0, 5) nm to the landing runway threshold, 0 otherwise.	0.0034	0	1
(I) Flight Characteristics	Airline^{+}	1 if flight *i* is operated by an international airline, 0 otherwise	0.21	0	1
	Body^{+}	1 if flight *i* is wide-body aircraft, 0 otherwise	0.24	0	1
	Runway^{+}	Dummy variable for calculated landing runway of flight*i*	–	0	1
	Daytime^{+}	1 if the observed time is between 6 am and 6 pm in local time, 0 otherwise	0.61	0	1
(II) Go-around Clustering Effect Features	GaGap	The minimal time interval between the approaching time of flight *i* and the initiation time of the latest go-around occurred in the past 24 h (in minutes)	712.62	5.37	1440
	GaCnt	The number of go-arounds occurred in the past 30 min	0.06	0	5
(III) Approach Features	Angle	Angle with the Extended Runway Centerline (in degree)	7.99	0.00	68.45
	AltDev	Absolute altitude deviation from 3-degree glideslope (in feet)	151.92	0.00	719.12
	Speed	Flight groundspeed (in knots)	163.19	77.28	277.85
	Energy	Kinetic energy height (in feet)	2841.45	1318.72	5370.49
(IV) In-trail Separation	LOS	The loss of separation between leading and the trailing flight *i* (in nautical miles)	0.09	0.00	2.61
	SpeedDiff	Groundspeed difference between leading and the trailing flight *i* (in knots)	20.92	−86.95	144.73
	AltDiff	The altitude difference between leading and trailing flight *i* (in 100 feet)	13.30	0.50	22.28
	NoLead^{+}	1 if there is no leading aircraft in front of flight *i* within 10-minute landing sequence, 0 otherwise	0.13	0	1
(V) Airport and Weather Condition Features	ArrQue	Difference between airport supplied arrival rate (AAR) and the number of intended landing aircraft (counts)	−2.04	−15	31
	DepQue	Difference between airport supplied departure rate (ADR) and the number of intended depart aircraft (counts)	1.24	−15	52
	RwyChange^{+}	1 if the used runway configuration is changed from the previous quarter hour, otherwise 0	0.08	0	1
	Wind	Wind speed where the headwind component is subtracted (in knots)	5.72	0	25.98
	Visibility_k	Discretized visibility (k = 1,2,3; intervals are [0, 3], (3, 5] and (5, 10] in miles)	–	0	10
	Ceiling_k	Discretized ceiling (k = 1, 2, 3, 4; intervals are [0, 5], (5, 10], (10, 30], (30, 100] (in 100 feet)	–	2	100
	IMC^{+}	1 for IMC, 0 for VMC	0.15	0	1
(VI)Surface Operation	$\widehat{ROB}$	Predicted runway occupancy buffer (in seconds)	30.60	−28.89	146.79
	RwyCnt	The number of aircraft and vehicles appearing on the landing runway (counts)	2.04	0	9

+ Variables are one-hot encoded.

daily basis for each flight arriving at JFK. One-hot encoding is applied to create dummy variables for wide-body aircraft (*Body*), international airliner (*Airline*), and the calculated landing runway (*Runway*). Specifically, one-hot encoding creates binary columns, indicating the presence of each possible value from the original categorical feature. This dataset is also used to derive flight approach features (discussed in Section 4.3) and in-trail separation features (discussed in Section 4.4).

The second dataset, airport surface detection equipment Model X (ASDE-X) data, allows us to determine the position of aircraft and ground support equipment in the airport surface area. Each record of raw surface track data contains the timestamp, latitude, longitude, altitude, and groundspeed. This dataset is used to derive surface operation features in Section 4.5.

The third dataset, which comes from the FAA aviation system performance metrics (ASPM) database, provides airport level configuration and weather information every quarter-hour. We use this dataset to derive the airport and weather features by matching each flight with the 15-minute interval in which the aircraft reaches 5 nm from the landing runway threshold. To capture the expected non-linearity of impacts of various visual conditions on go-around occurrence, the visibility variable (in statute miles) is discretized into three continuous subsections: [0, 3], (3, 5] and (5, 10]. Similarly, the ceiling variable (in 100 feet) is discretized into four continuous subsections: [0, 5], (5, 10], (10, 30], (30, 100]. For example, if the ceiling equals 600 feet, the discretized ceiling variables Ceiling_k are 5, 6, 0, 0. Headwind or tailwind speed, as well as crosswind speed, are calculated using trigonometric calculations with the information of wind speed (in knots), wind angle (in degrees), and landing runway configuration at the airport. When the wind is a headwind, the tailwind is set as zero, and vice versa. We subtract the arrival/departure rate (counts) from the arrival/departure demand (counts) to capture the airport traffic conditions. A negative sign in these demand-minus-capacity variables indicates the absence of an arrival queue. The Airport Arrival/Departure Rate (AAR/ADR) and the number of intended landing/departing aircraft (demand) are obtained directly from the ASPM dataset on a quarter-hourly basis. For a given flight, the change of runway configuration variable (*RwyChange*) is set to 1 if the used runway configuration during the observed period is different from the preceding 15-minute period, and 0 otherwise. As an additional indicator of operational traffic, we include daytime dummy variables if the observed aircraft reaches 5 nm information cutoff gate between 6 am and 6 pm in local time (*Daytime*). A dummy variable is also created to indicate the Instrument Meteorological Conditions (*IMC*), as opposed to Visual Meteorological Conditions (*VMC*).

4.2. Go-around clustering effect features

From the go-around detection results, we observed that go-arounds sometimes occur in clusters—that is several occur in a short time interval. To capture this effect, we calculate the time difference between when a given flight is at the 5 nm information cutoff gate and the initiation time of the latest go-around that occurred in the past 24 h. This minimum time difference (*GaGap*) among all other go-arounds in record is used as a temporal clustering feature. If no go-arounds occurred in the past 24 h for a given flight, we set the *GaGap* to 1440 (minutes). The *GaGap* variable only focuses on the effect from the previous go-around flight and such an effect weakens with time. As a second clustering feature, we include the number of go-arounds (except the given flight if it was a go-around) that occurred in the past 30 min in JFK airport when a given flight *i* was at 5 nm from the runway threshold (*GaCnt*). This variable measures the clustering effect in terms of quantity.

4.3. Flight approach features

As the (FAA, 2018) emphasized, "If not stabilized, go around". Continuation of an unstabilized approach to land may result in an aircraft arriving at the runway threshold too high, too fast, out of alignment with the runway centerline, incorrectly configured, or otherwise unable to land safely. Accordingly, we derive altitude deviation (*AltDev*), groundspeed (*Speed*), angle with the extended runway centerline (*Angle*), and Kinetic energy height (*Energy*) as flight approach features to capture potential instability indicators that may prompt a go-around.

Normally the optimum vertical profile to use during a landing approach is the standard 3-degree glideslope path (Kim et al., 2008), which requires that the aircraft descend at about 300 feet per nautical mile. A large deviation from the target descent rate indicates an unstable approach. Thus, we calculate the altitude deviation from the standard 3-degree glideslope path (*AltDev*) to capture the potential unstabilized approach risk:

$$AltDev_i = |h_i - 6076.12 \cdot 5 \cdot tan(3^\circ)| \tag{1}$$

where h_i is aircraft i's altitude in feet at the 5 nm information cutoff gate, 1 nm = 6076.12 feet.

When the aircraft i is at the 5 nm information cutoff gate, its groundspeed (*Speed*) after median filtering is extrapolated to capture the scenario where an aircraft approaches too fast. We applied the median filtering as a preprocessing step to remove out-of-range isolated noise in the trajectory data. The median filter is one kind of smoothing technique, which runs through the data entry by entry, replacing each entry with the median of neighboring entries. At this moment, the angle of horizontal deviation from the extended runway centerline (*Angle*) is also calculated using the flight latitude and longitude information to measure the misalignment of the standard approach path.

Aircraft energy management is of great importance during the approach procedure to maintain safety. An aircraft's energy state is the sum of potential energy and kinetic energy per unit weight (Amelink et al., 2005). However, the calculation requires information on aircraft mass, which depends on payload and fuel load data that are not available to the researchers. Thus, we use the energy height metric that can be defined as the hypothetical height Gluck et al. (2019), H_i, at which the aircraft i's potential energy ($m_i g H_i$) is equal to the total energy at its current state ($m_i g h_i + \frac{1}{2} m_i v_i^2$), so the metric is calculated as:

$$H_i = h_i + \frac{v_i^2}{2g} \tag{2}$$

where H_i is the kinetic energy height to be calculated (*Energy*, in feet), h_i and v_i are respectively the aircraft altitude and aircraft groundspeed when the flight i is at the 5 nm information cutoff gate, and g is the constant of gravitational acceleration. This metric can be calculated for each flight during the approach process to represent the aircraft energy-related risks using only the surveillance track data.

4.4. In-trail relationship features

Separation is defined as the distance, either horizontal or vertical, between two aircraft. Here we are interested in the distance between a given aircraft on approach the lead aircraft (if any) landing on the same runway. The minimum required horizontal separation in this situation depends on the relative weight class of two aircraft and meteorological conditions (visual or instrument). For instrument conditions, the separation is prescribed explicitly, while in good visibility the Loss of separation (LOS) occurs whenever the specified separation minima are breached. It is calculated as the difference between the minimum required separation and the actual separation between the lead-trail aircraft pair. We expect that a more significant loss of separation (in nautical miles) increases the probability of go-around. Therefore, to capture the separation effect, we derive four variables that are employed in the statistical models: *Loss of separation (LOS),* a dummy variable *NoLead* indicating the case where there was no leading aircraft for a given flight, the speed difference *(SpeedDiff)* and the altitude difference *(AltDiff)* between leading and trailing (subject) flight.

The algorithm for obtaining these variables requires three steps – finding leading and trailing aircraft pair, obtaining actual separations, speed difference, and altitude difference for the lead-trail aircraft pair, and finally calculating the loss of separation.

i. Group flights with the same (calculated) landing runway obtained from the go-around detection algorithm in Table 1, and sort them in chronological order based on the time that flights cross the runway threshold. For each group, we create a list of tuples where each tuple contains two consecutive aircraft that have been sorted. Within each tuple, if the runway threshold crossing time difference of the two aircraft is smaller than 10 min, then we define them as a lead-trail aircraft pair. Otherwise, we set a dummy variable *NoLead* to 1 for the trailing (subject) aircraft to indicate the case in which, for all practical purposes, there was no leading aircraft for a flight.

ii. For each trailing flight, we find the linearly extrapolated timestamp *t*at which the trailing (subject) flight is 5 nm to its landing runway. At the extrapolated timestamp *t*, we again extrapolate the locations (latitude, longitude, altitude) and groundspeed of both leading and trailing aircraft. The separation between these two extrapolated locations (in terms of latitude and longitude) is noted as S_t. We also calculate the speed difference *(SpeedDiff)* and altitude difference *(AltDiff)* between leading and trailing (subject) flight at this extrapolated timestamp *t*. To be specific, we subtract the extrapolated groundspeed/altitude of the leading aircraft from the extrapolated groundspeed/altitude of the trailing (subject) aircraft when the trailing flight is at the 5 nm information cutoff gate.

iii. Obtain the separation minima from FAA Wake Separation Standards (FAA, 1991) based on the weight class of leading and trailing aircraft under VMC (S_m^{VMC}) and IMC (S_m^{IMC}). When the trailing flight is at 5 nm to its landing runway, if the meteorological condition is recorded as "VMC" in the ASPM quarter-hour dataset, the standard separation minima is $S_m = S_m^{VMC}$ (e.g., 1.9 nm for the Large-Large lead-trail pairs), otherwise $S_m = S_m^{IMC}$(e.g., 3.0 nm for the Large-Large lead-trail pairs). Thus, the loss of separation *(LOS)* is $S_l = max(0, S_m - S_t)$, and is directly employed as a continuous variable in the model.

4.5. Surface operation features

In the case where pilots or controllers anticipate a runway incursion, a common practice would be to initiate a go-around (FAA, 2015). Therefore, we have derived two variables – predicted Runway Occupancy Buffer ($\widehat{ROB}$) and counts of objects (both aircraft and vehicles) on the runway *(RwyCnt)* - to serve as indicators of incursion risk and used them as features in our go-around model.

The *ROB* is defined as the time difference between the runway threshold crossing time of the trailing aircraft and the runway exit time of the leading aircraft. When a trailing aircraft reaches a certain point in the final approach phase (referring to 5 nm information cutoff gate in this paper), the $\widehat{ROB}$ is predicted using algorithms given by Dai and Hansen (Dai and Hansen, 2020). It captures the variations in the runway threshold interarrival time (Tošic et al., 1976), the landing occupancy time (Simpson, 1986), and the spacing buffers routinely applied by air traffic controllers (Ruiz et al., 2013). Note that we incorporate the predicted $\widehat{ROB}$ at the 5 nm information cutoff gate, instead of the observed ROB, in order to reflect the information available at the time of a go-around decision. For flights that do not have leading aircraft, this value is set 0 s, but with a dummy variable, *NoLead* added as described above.

The other incursion variable is the number of aircraft or ground vehicles on the runway *(RwyCnt)* when the subject flight is 5 nm from its landing runway threshold. We first define the Runway Safety Area (RSA) polygon bounded by holding position markings painted on the taxiway or runway surface (FAA, 2007; Dai and Hansen, 2020). When a trailing (subject) aircraft reaches the 5 nm information cutoff gate, we count the total number of arrivals, departures, and crossing aircraft/vehicles that are contained in the corresponding landing RSA polygon at that moment, using ASDE-X surface track data.

5. Principal component logistic regression model

In this section, we investigate how the derived features impact go-around occurrence, using a principal component logistic regression (PCLR) model. We found that almost 90% of detected go-arounds in our data set occurred within five nautical miles of the landing runway threshold. Therefore, we develop the PCLR model to predict go-around occurrence based only on features known when the subject flight reaches the 5 nm information cutoff gate. We first illustrate the detailed algorithm and estimation procedures of the PCLR model. The estimation results are then interpreted through factor loading analysis and reconstruction of coefficients back for the original variables. Lastly, we construct counterfactual scenarios to quantify factor contributions to go-arounds, based on the models estimated at the cutoff gate of 5 nm.

In the 5 nm models, the dependent variable *Y* is set to 1 if a flight is detected as a go-around occurring within [0, 5) nautical miles to its landing runway threshold, 0 otherwise. (Flights that initiate go-arounds more than 5 nm from the threshold are not considered.) Therefore, by applying the retrospective causal inference method (Holland and Rubin, 1987) to observational data, we can capture the statistical relations among the go-around occurrence and features described in Section 4.

5.1. Standard logit model

We firstly estimated a standard binary logistic regression model to relate go-around occurrence to contributing factors. The model specification is formulated as in Eq. (3) and Eq. (4), where $\mathscr{Y}$ is the log-odds function, $\boldsymbol{X}$ is a design matrix that contains all contributing factors introduced in Section 4, and β is the associated coefficient vector estimated by employing maximum likelihood estimation (MLE).

$$\mathscr{Y} = \boldsymbol{X} \cdot \boldsymbol{\beta} \tag{3}$$

The probability of an aircraft initiating go-around $P_r(y_i = 1|\boldsymbol{X})$ can be written as:

$$\Pr(y_i = 1|\mathbf{X}) = \frac{1}{1 + \exp(-\mathscr{Y})} \tag{4}$$

The estimation results are presented in Table 3. The majority of coefficients are not significant at a 5% confidence level, and many have unexpected signs. For example, the estimates for the visibility and ceiling variables suggest that flights landing at an airport with good visibility and ceiling conditions would have a higher probability of go-around, which is not plausible in practice. This is probably because many independent variables used in the model are highly correlated. As a result of this multi-collinearity, the standard logistic regression model fails to give us a proper understanding of the contributing effects. To remedy this problem, we employ decorrelation techniques.

5.2. Principal component logistic regression (PCLR) and interpretation

To handle the multi-collinearity problem, we apply Principal Component Analysis (PCA) to decorrelate and reduce the dimensionality of the original feature space. Instead of regressing the dependent variable on the explanatory variables directly, the principal components (PCs) formed by all the explanatory variables are used as covariates in the logistic regression model.

5.2.1. PCLR of mixed data

While PCA is a mature technique to decorrelate feature vectors, it must be adapted in our setting because our dataset contains a mixture of continuous and categorical variables. Specifically, the design matrix (feature space) for the 5-nautical-mile model contains 28 features vectors, 21 of which are continuous and seven are categorical, including *Runway* (8 levels), *RwyChange* (2 levels), *Daytime* (2 levels), *Airline* (2 levels), *Body* (2 levels), *MC* (2 levels) and *NoLead* (2 levels). Therefore, appropriate treatment of such mixed data types, especially the categorical variables, is required for PCA application. Accordingly, we adapted and applied the PCA-mixed algorithm introduced by (Chavent et al., 2017) to deal with our mixed set of variables. The detailed notations and algorithm are described as follow.

A. Notations

Let $\boldsymbol{X} = \left[\boldsymbol{X}_1^{n\times p_1}|\boldsymbol{X}_2^{n\times p_2}\right]$ denote the full design matrix, which is constructed by two submatrices $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. $\boldsymbol{X}_1$ contains solely continuous variables with dimension n by $p_1 (p_1 = 21)$, while $\boldsymbol{X}_2$ contains categorical variables with dimension n by $p_2 (p_2 = 7)$. We further denote $q_1, q_2, \cdots, q_{p_2}$ as the number of levels for each categorical variable (e.g., $q_1 = 8$ for the first categorical variable, which is *Runway*), and $m = \sum_{i=1}^{p_2} q_i$ as the total levels for all categorical variables. Notice that the elements in $\boldsymbol{X}_2$ are integers that range from 1 to the number of levels for each variable.

B. Design Matrix Preparation

Using the above notation, we first convert $\boldsymbol{X}_2$ to a complete disjunctive table (CDT) $\boldsymbol{Z}_2 = [z_1, z_2, \cdots, z_m] \in \mathbb{B}^{n\times m}$ by employing one-hot encoding. Then we center $\boldsymbol{Z}_2$ by respectively subtracting the mean of each column, denoted as $\boldsymbol{Z}_2^c$, and standardize $\boldsymbol{X}_1$ to zero mean and unit standard deviation, denoted as $\boldsymbol{X}_1^s$. Lastly, we combine $\boldsymbol{X}_1^s$ and $\boldsymbol{Z}_2^c$ to build a new design matrix $\boldsymbol{Z} = \left[\boldsymbol{X}_1^s|\boldsymbol{Z}_2^c\right]$. Notice that the rank of $\boldsymbol{Z}$ equals to $r = p_1 + m - p_2$.

C. Generalized Singular Value Decomposition (GSVD)

We first define a weighting matrix $\boldsymbol{M} = \begin{bmatrix} \mathbb{I}_{p_1} & 0 \\ 0 & \mathbf{W} \end{bmatrix}$, where $\mathbb{I}_{p_1}$ is an identity matrix with dimension p_1. $\boldsymbol{W} = [w_{ii}] \in \mathbb{R}^{m\times m}$ is a di-

Table 3
Standard logit model estimation results.

Variable	Estimate (std.)	Variable	Estimate (std.)	Variable	Estimate (std.)	Variable	Estimate (std.)
Constant	-5.252^{***} (0.847)	**Visibility$_1$**	−0.175 (0.169)	**ArrQue**	0.026^{**} (0.010)	**RwyChange**	−0.016 (0.206)
AltDev	0.258^{***} (0.035)	**Visibility$_2$**	0.017 (0.079)	**DepQue**	0.022^{**} (0.007)	**Rwy04R**	-1.330^{***}(0.255)
Speed	-0.022^{**} (0.007)	**Visibility$_3$**	−0.171* (0.070)	**RwyCnt**	−0.054 (0.037)	**Rwy13L**	-4.156^{***}(0.742)
Energy	0.003^{***} (0.000)	**Ceiling$_1$**	−0.282 (0.158)	$\widehat{\boldsymbol{ROB}}$	−0.005 (0.003)	**Rwy13R**	1.636^{**} (0.603)
Angle	0.032* (0.013)	**Ceiling$_2$**	0.010 (0.031)	**IMC**	0.493 (0.385)	**Rwy22L**	-0.996^{***}(0.253)
LOS	1.682^{***} (0.234)	**Ceiling$_3$**	0.011 (0.016)	**Daytime**	0.080 (0.124)	**Rwy22R**	-1.803^{***}(0.348)
SpeedDiff	0.001 (0.004)	**Ceiling$_4$**	-0.006^{**} (0.002)	**AirlineIntl**	0.425^{**} (0.162)	**Rwy31L**	-1.333^{***}(0.333)
AltDiff	-0.105^{***} (0.027)	**GaCnt**	0.546^{***} (0.088)	**BodyWide**	0.328* (0.158)	**Rwy31R**	-1.390^{***}(0.310)
Wind	0.045^{**} (0.014)	**GaGap**	−0.000 (0.000)	**NoLead**	0.305 (0.278)		
Log-likelihood	−1721.7			**Pseudo R-squared**	0.222		

Variables are significant at the 0.1% level***, 1% level**, 5% level*.

agonal matrix where $w_{ii} = \frac{n}{1^T \cdot z_i}$, and 1 is a vector of ones. Then we perform generalized singular value decomposition on the product of matrices $\boldsymbol{Z}$ and $\boldsymbol{M}$ (Eq. (6)), where $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal matrices, and Λ is a diagonal matrix that contains singular values sorted by their values.

$$\boldsymbol{Z} \cdot \boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^{\boldsymbol{T}} \tag{5}$$

Notice that in Eq. (6), $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_{p_1+m}]$ represents the principal component directions of the matrix $\boldsymbol{Z} \cdot \boldsymbol{M}$, and $\boldsymbol{\Lambda} = \text{diag}\{\sqrt{\sigma_1}, \sqrt{\sigma_2}, \cdots, \sqrt{\sigma_{p_1+m}}\}$where σ_i's are eigenvalues of $\boldsymbol{M}^{\mathbf{T}}\boldsymbol{Z}^{\mathbf{T}}\boldsymbol{Z}\boldsymbol{M}$. Thus, we can find the principal components of $\boldsymbol{Z} \cdot \boldsymbol{M}$, that is $\boldsymbol{F} \in \mathbb{R}^{n \times r}$, by using Eq. (6). $\boldsymbol{F}$ has the same rank as $\boldsymbol{Z}$.

$$\boldsymbol{F} = \boldsymbol{Z} \cdot \boldsymbol{M} \cdot \boldsymbol{V} \tag{6}$$

D. Derived Covariates

Common techniques to derive covariates from $\boldsymbol{F}$ include: (a) pick the top k columns with the sum of explained variances that exceeds some thresholds; (b) pick the top k columns with the smallest squared singular value (i.e., eigenvalue) exceeding some threshold; and (c) pick the top k columns with the sum of squared singular values exceeding some threshold. However, these variance-based or singular-value-based criterion might not always be optimal in predictive analytics. PCs with large variances are not necessarily the best predictors (Aguilera et al., 2006) as principal components with low explained variability could be highly correlated with the response variable. Therefore, the dependence between response and predictor variables must be taken into account.

In order to choose the number of principal components k big enough to account for the variance in the data as much as possible, and also reduce the dimensionality of the data, we have applied the Kaiser rule (Carter Hill et al., 1977; Longman et al., 1989; Kaiser, 1991) by iteratively selecting k PCs using the aforementioned criteria (b) with a threshold δ. When varying the threshold δ from 0.5 to 1.0 with a step of 0.1, we regress selected PCs with the response value Y using logistic regression, and record the model's adjusted pseudo R-squared. We determine the finalδ and k PCs with the best adjusted pseudo R-squared.

We denote the selected PCs as $\boldsymbol{F}_k$, and hereafter we use $\boldsymbol{F}_k$ as the final design matrix to conduct logistic regression analysis using the same technique described in Section 5.1. except that in Eq. (4), we use the feature vectors in $\boldsymbol{F}_k$ instead of $\boldsymbol{X}$.

E. Transformation of Estimated Coefficients

While the PCLR regime gives us estimates for principal components, we eventually desire estimated coefficients of the actual features derived in Section 4 (e.g., ceilings and runway fixed effects). Let α denote the coefficient vector for PCs, then the utility function (3) can be rewritten as:

$$\mathcal{V} = \boldsymbol{F}_{\mathbf{k}} \cdot \boldsymbol{\alpha} = (\mathbf{Z} \cdot \mathbf{M} \cdot \mathbf{V}_{\mathbf{k}}) \cdot \boldsymbol{\alpha} = \boldsymbol{Z} \cdot (\mathbf{M} \cdot \mathbf{V}_{\mathbf{k}} \cdot \boldsymbol{\alpha}) = \boldsymbol{Z} \cdot \boldsymbol{\beta} \tag{7}$$

where $\boldsymbol{V}_k$ is the first k columns of the matrix $\boldsymbol{V}$.

Given Eq. (8), the logistic regression model with respect to $\boldsymbol{F}_k$ can be equivalently expressed in matrix form with respect to the original feature space $\boldsymbol{Z}$. Thus, the associated coefficient vector is given by:

$$\boldsymbol{\beta} = \boldsymbol{M} \cdot \boldsymbol{V}_k \cdot \boldsymbol{\alpha} \tag{8}$$

5.2.2. Factor loading analysis

In addition to obtaining principal components (PCs) and their associated estimates, we are also interested in linking PCs to the original feature space in order to identify the variables that are primarily associated with any given PC. To do so, we use factor analysis to map PCs to groups of features quantitatively.

We first denote $\boldsymbol{L} = \begin{bmatrix} \boldsymbol{L}_1 \\ \boldsymbol{L}_2 \end{bmatrix}, \boldsymbol{L}_1 \in \mathbb{R}^{p_1 \times k}, \boldsymbol{L}_2 \in \mathbb{R}^{m \times k}$ as the loading matrix representing the variance in features explained by PCs. The formal definition is formulated as:

$$\boldsymbol{L} = \boldsymbol{M} \cdot \boldsymbol{V}_k \cdot \boldsymbol{\Lambda}_k \tag{9}$$

where $\boldsymbol{V}_k$ is the first k columns of matrix $\boldsymbol{V}$, and $\boldsymbol{\Lambda}_k$ is the k$^{\text{th}}$ order leading principal minors of $\boldsymbol{\Lambda}$. However, due to the fact that continuous variables and categorical variables are not on the same scale, and thus we cannot compare their explained variance by PCs directly from $\boldsymbol{L}$. We have derived a ***contribution matrix*** $\boldsymbol{C} = \begin{bmatrix} \boldsymbol{C}_1 \\ \boldsymbol{C}_2 \end{bmatrix} \in \mathbb{R}_+$ in which each element c_{ij} describes the contribution of the i^{th} feature to the j^{th} PC. Specifically, a larger value of c_{ij} indicates a higher contribution of the i^{th} feature to the j^{th} PC. Furthermore, the $\boldsymbol{C}$ matrix can be decoupled into two submatrices where $\boldsymbol{C}_1 \in \mathbb{R}_+^{p_1 \times k}$ and $\boldsymbol{C}_2 \in \mathbb{R}_+^{p_2 \times k}$ respectively correspond to the continuous and categorical contribution matrices. Eq. (11) – (12) illustrate the derivation of $\boldsymbol{C}$.

$$\boldsymbol{C}_1 = \mathbf{L}_1{}^{\circ}\mathbf{L}_1 \tag{10}$$

$$\boldsymbol{C}_2 = \mathbf{H} \cdot (\mathbf{L}_2{}^{\circ}\mathbf{L}_2) \tag{11}$$

where $\circ$ denotes element-wise multiplication. $\boldsymbol{H} \in \mathbb{R}^{p_2 \times m}$ is a block diagonal matrix in which the diagonal elements are vectors of the level frequency of the i^{th} categorical variable, and the off-diagonal elements are 0. $q_1, q_2, \cdots, q_{p_2}$ are the number of levels for each categorical variable (e.g., $q_1 = 8$ for variable *Runway*), and $m = \sum_{i=1}^{p_2} q_i$ is the total levels for all categorical variables.

$$\boldsymbol{H} = \begin{bmatrix} \left(\frac{1^{\mathrm{T}} \cdot z_{11}}{n} \quad \frac{1^{\mathrm{T}} \cdot z_{12}}{n} \cdots \quad \frac{1^{\mathrm{T}} \cdot z_{1(q_1)}}{n}\right) & 0 & \cdots & 0 \\ 0 & \left(\frac{1^{\mathrm{T}} \cdot z_{21}}{n} \quad \frac{1^{\mathrm{T}} \cdot z_{22}}{n} \cdots \quad \frac{1^{\mathrm{T}} \cdot z_{2(q_2)}}{n}\right) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \left(\frac{1^{\mathrm{T}} \cdot z_{p_2 1}}{n} \quad \frac{1^{\mathrm{T}} \cdot z_{p_2 2}}{n} \cdots \quad \frac{1^{\mathrm{T}} \cdot z_{p_2(q_{p_2})}}{n}\right) \end{bmatrix}$$

With such operations, the contribution of a categorical variable is the weighted sum of the squared loadings of its classes in $\boldsymbol{L}_2$ and therefore is equivalent to their correlation ratios. For the contribution of a continuous variable, on the other hand, the squared loading equals the squared correlation.

5.3. Estimation results

Using the above Kaiser rule, the first 17 principal components with $\delta = 0.8$ that explain 82% of the total variance have the best adjusted pseudo R-squared. In the end, we are left with 9 PCs after removing insignificant principal components from the estimated logistic regression model.

5.3.1. Factor analysis and model result

For the convenience of understanding the model estimated coefficients for each factor (PC), we first need to deploy the pattern matrix $\boldsymbol{L}$ (Eq. (10)) and the contribution matrix $\boldsymbol{C}$ (Eq. 11–12) from factor analysis to interpret the relationships between PCs and the original features. The pattern matrix $\boldsymbol{L}$ indicates the magnitude and sign of the correlation between the PCs and the original variables. The contribution matrix $\boldsymbol{C}$ transforms the continuous variable contributions and the categorical variable contributions on the same scale. It then describes the magnitude of the overall contribution of an individual variable to a PC.

The first and second columns of Table 4 are determined by finding which variable(s) – either continuous or categorical – make above-average contribution(s) to each PC based on contribution matrix $\boldsymbol{C}$. The average contribution is the value when all variables have the same contributions (i.e., 100% divided by the total number of variables). In the case of categorical variables, which may make an above-average contribution to more than one PC, we base the assignment on the maximum loading. We present the loading value – either positive or negative – of the i^{th} variable to the assigned j^{th} PC (from pattern matrix $\boldsymbol{L}$) in the third column. To save space, we only show the loading (l_{ij}) of the i^{th} variable to the assigned j^{th} PC in Table 4 rather than the whole pattern matrix $\boldsymbol{L}$ and the whole contribution matrix $\boldsymbol{C}$. Note that variables may have high loading values on just one or two PCs, or have a balanced spread with small loading values across more PCs. According to the allocation result, we assign semantic labels for each PC in the fourth column and use them to interpret the estimation results of the PCLR model in Table 5.

Turning to Table 5, we note that PC1 has a highly significant, positive coefficient estimate. From Table 4, we see that the visibility and ceiling variables are loaded in the opposite direction with PC1, while the IMC indicator variable is loaded in the same direction.

Table 4
Loadings of variables with above-average contributions for each PC.

PC	Related Variable (x)	Loading (l_{ij})	Semantic Labels	PC	Related Variable (x)	Loading (l_{ij})	Semantic Labels
1	Visibility_1	−0.715	Meteorological conditions	**5**	ArrQue	0.592	Traffic conditions
	Visibility_2	−0.817			DepQue	0.586	
	Visibility_3	−0.821			GaGap	−0.399	
	Ceiling_1	−0.643			GaCnt	0.210	
	Ceiling_2	−0.823			RwyCnt	0.307	
	Ceiling_3	−0.838			Rwy22R	1.726	
	Ceiling_4	−0.616		**6**	AirlineIntl	1.396	Aircraft characteristics
	IMC	2.000			BodyWide	1.293	
2	AltDiff	0.724	Lead-trail spacing	**7**	Rwy13R	0.975	Approach pattern
	Energy	0.719			Rwy31L	−1.136	
	$\widehat{ROB}$	−0.411		**8**	Wind	0.623	Wind and daytime effect
	SpeedDiff	0.220			Daytime	0.689	
	NoLead	−1.232		**13**	RwyChange	1.172	Approach pattern
4	Angle	0.618	Approach procedure features		Rwy13L	−1.415	
	AltDev	0.457			Rwy31R	0.606	
	Speed	0.430		**15**	LOS	0.812	Loss of separation
	Rwy04R	−0.781					
	Rwy22L	−0.682					

Table 5
PCLR model estimation results.

Dimension	Est./Std.	Dimension	Est./Std.	Dimension	Est./Std.
Constant	-6.560^{***}(0.088)				
PC1	0.252^{***}(0.018)	**PC5**	0.152^{***}(0.036)	**PC8**	0.177^{***}(0.046)
PC2	0.207^{***}(0.039)	**PC6**	0.407^{***}(0.040)	**PC13**	0.453^{***}(0.036)
PC4	0.149^{***}(0.043)	**PC7**	0.144^{***}(0.033)	**PC15**	0.180^{***}(0.022)

Variables are significant at the 0.1% level***, 1% level**, 5% level*.

Thus, the PC1 result indicates that adverse meteorological conditions (low visibility and ceiling, or IMC condition) increase the probability of the go-around occurrence.

The coefficient estimate of PC2 indicates that the threat of the lead and trail aircraft simultaneously occupying the runway increases the likelihood of a go-around. A small runway occupancy buffer ($\widehat{ROB}$), or the trailing aircraft approaching with high energy (*Energy*), or the trailing aircraft chasing too close to its leading flight (*SpeedDiff*) increases this threat, while the absence of a lead aircraft (*NoLead*) clearly reduces it. Similarly, the positive coefficient on PC15 implies that a higher loss of separation compared to FAA standards increases the probability of go-around.

The PC4 captures the effect of approach procedure deviations. The positive coefficients PC4 in Table 5 implies that flights that are deviated from the optimum approach procedure (3-degree glideslope, runway alignment, proper speed control) are more likely to initiate go-arounds. The landing runway indicators 04R and 22L are also captured in this PC, perhaps indicating that approaches to these two runways are correlated with the other PC4 covariates. It suggests that flights landing on runway 04R/22L would be less likely to initiate go-arounds than other runways, perhaps because 04R/22L has the most advanced landing aids of the JFK runways. As reported by (The Port Authority of New York & New Jersey, 2020), Runway 4R is a Category IIIB ILS runway, permitting landings with as little as 600 feet of visibility; Runway 22L is equipped with a Precision Approach Path Indicator (PAPI) and allows landings down to visibility of less than a half-mile (2640 feet), while other runways at JFK require more than half-mile visibility for landing. These technologies make it easier for pilots to land, alleviate the operational risks, and avert go-arounds.

The arrival queue, departure queue, go-around clustering effect, and the number of objects occupying the runway during the landing process are captured by PC5, which has a positive impact on the go-around occurrence. This may indicate that increased controller workload or pressure to maximize throughput increases the go-around probability. PC5 also picks up the clustering effect whereby go-arounds are more likely in the time period surrounding a given go-around.

The PC6 captures the aircraft characteristics – fixed effects of international airliners and wide-body aircraft, which is found a significant positive impact on go-around occurrence. The daytime operations and strong winds (PC8) increases the likelihood of go-around occurrences, as does a change of runway configuration (PC13). This could reflect how the configuration change interrupts traffic patterns, increasing pilot and controller workload.

Besides the landing runway 04R/22L loaded by PC4, the fixed effects of other landing runway variables for capturing different approach patterns are loaded in different PCs – 22R in PC5, 13R/31L in PC7, and 13L/31R in PC13. All of these PCs are statistically significant which suggests that, all else equal, this is a greater proclivity toward go-arounds in certain landing runways, but further investigation is required to interpret the relationship between runway configuration and go-around occurrence.

Table 6
Reconstructed coefficients of original variables.

Variable	Reconstructed Coef. β (std.)	Variable	Reconstructed Coef. β (std.)	Variable	Reconstructed Coef. β (std.)	Variable	Reconstructed Coef. β (std.)
Constant	-6.560^{***} (0.088)	**Visibility$_1$**	-0.164^{***} (0.030)	**ArrQue**	0.016^{***} (0.003)	**RwyChange**	0.562^{***} (0.004)
AltDev	0.015^{***} (0.006)	**Visibility$_2$**	-0.061^{***} (0.005)	**DepQue**	0.013^{***} (0.002)	**Rwy04R**	-0.036 (0.029)
Speed	0.010^{***} (0.001)	**Visibility$_3$**	-0.052^{***} (0.005)	**RwyCnt**	0.011 (0.008)	**Rwy13L**	-0.209^{**} (0.073)
Energy	0.001^{***} (0.000)	**Ceiling$_1$**	-0.190^{***} (0.038)	$\widehat{ROB}$	-0.005^{***} (0.001)	**Rwy13R**	2.758^{***} (0.180)
Angle	0.001^{**} (0.000)	**Ceiling$_2$**	-0.040^{***} (0.003)	**IMC**	0.244^{***} (0.006)	**Rwy22L**	-0.094^{**} (0.035)
LOS	2.101^{***} (0.215)	**Ceiling$_3$**	-0.008^{***} (0.001)	**Daytime**	0.201^{***} (0.023)	**Rwy22R**	0.383^{***} (0.068)
SpeedDiff	0.004^{***} (0.000)	**Ceiling$_4$**	-0.001^{**} (0.000)	**AirlineIntl**	0.463^{***} (0.014)	**Rwy31L**	-0.420^{***} (0.049)
AltDiff	0.023^{***} (0.005)	**GaCnt**	0.549^{***} (0.046)	**BodyWide**	0.445^{***} (0.015)	**Rwy31R**	0.107 (0.070)
Wind	0.058^{***} (0.007)	**GaGap**	-0.001^{***} (0.000)	**NoLead**	-0.473^{***} (0.005)		

Variables are significant at the 0.1% level***, 1% level**, 5% level*.

5.3.2. Transformation of coefficients

In the above section, we interpret how the original derived features impact go-around occurrence using the assigned semantic labels for each PC based on factor analysis. This section further quantifies the impacts of the original derived features by reconstructing their estimates using Eq. (8). The results are presented in Table 6. The coefficient estimates of the original features are based on the assumption that the features affect go-around occurrence through their effects on the factors included in the model. Compared to the standard logit model estimation results in Table 3, the coefficients in Table 6 are quite different, and the standard errors of the coefficient estimates are much lower. Nearly all the coefficients become statistically significant at the 1% confidence level and have expected signs. The PCLR model removes collinearity without eliminating any of the original variables and reduces the variance of the estimated coefficients. The majority of the estimates (β) are consistent with the discussions in the factor loading analysis. Note, however, that individual coefficient estimates are biased and the main value of the PCLR method is in the estimates of the latent variable coefficients. Practice (Mason and Gunst, 1985) has shown that strong collinearity induces the conditions under which the PCLR method is beneficial, in that the PCLR allows for minor bias for the sake of substantially smaller variance and improved model interpretability (Jolliffe, 1986). We here plot the coefficients of different visibility, and the ceiling discretized variables in Fig. 4. The green bar represents visibility (in statute miles), and the blue bar represents the ceiling (in 100 feet). We observe that go around occurrence is more sensitive to visibility and ceiling variation when these values are less than 3 statute miles and less than 500 feet, respectively, and that this sensitivity declines markedly as these conditions improve.

5.3.3. Counterfactual analysis

In this section, we directly measure the contributions of different factors to the go-around occurrence by conducting a counterfactual analysis. Each counterfactual scenario is constructed by setting a particular feature to its "best" value while leaving the other features unchanged. For example, model estimates suggested that the ceiling has a negative effect on go-around occurrence. To construct the counterfactual scenario for ceiling, we set the ceiling to 10,000 feet for each observation in the data set. Based on this assumption we reset the values of the various ceiling-related valuables, while leaving all other values unchanged. Then, we use the estimated PCLR model to predict the corresponding go-around probability for each flight. The variable contribution is calculated by measuring the percentage reduction between the baseline go-around rate and the expected go-around rate (Eq. (12)). Note that we assume individual feature does not directly impact go-around occurrence but via their effects on the PCA factor scores F.

$$\%reduction = \frac{P_{GA} - E(P_{GA}|X^{'})}{P_{GA}} \times 100\% \tag{12}$$

where $E(P_{GA}|X^{'})$is the expected go-around rate given the counterfactual input $X^{'}$; P_{GA} is the baseline go-around rate.

Table 7 reports the scenario value for each variable, that is the value found in the data that, based on the sign of its coefficient, would minimize go-around occurrence (labeled "Expected GA%"), and the percentage reduction in go-around occurrence relative to the observed baseline of 0.343%.

The relative importance of the variables on the reduction of go-around probability is shown in Fig. 5, where each row represents one variable. The length of the color bar indicates the variable contribution in percentage. The stability of an approach, flight lead-trail spacing, departing traffic and airport ceiling are the most important factors of go-around occurrence. If the flight aligns the extended runway centerline properly at 5 nm, strictly follows the 3-degree glideslope, and maintains the effective speed control and energy management, the go-around rate would potentially decline about 30%. For aircraft forming in-trail relationships, the go-around rate would also decline by about 28% by maintaining appropriate following speed and keeping safe spatial separation, while the absence of a lead aircraft (*NoLead*) would result in a 19% drop of the go-around rate. Managing and optimizing the departure queue seems to have a more substantial contribution to decreasing go-around rates (25%) than reducing the arrival queue (16%). The go-around rate would drop by more than 25% if the airport ceiling were set to its ideal scenario value, and 21% if there were high visibility. If all the flights in the observation dataset are narrow-body aircraft or operated by domestic airliners, the go-around rate decreases by 25% and 9%, respectively. The wind speed effect contributes 20% to the reduction of go-around occurrence. Ensuring that there are no aircraft or vehicles on the runway safety area when a flight is 5 nm from its landing runway threshold, would result in a 9% reduction of go-around occurrence. Finally, eliminating the clustering effect would reduce the go-around occurrence by 5%.

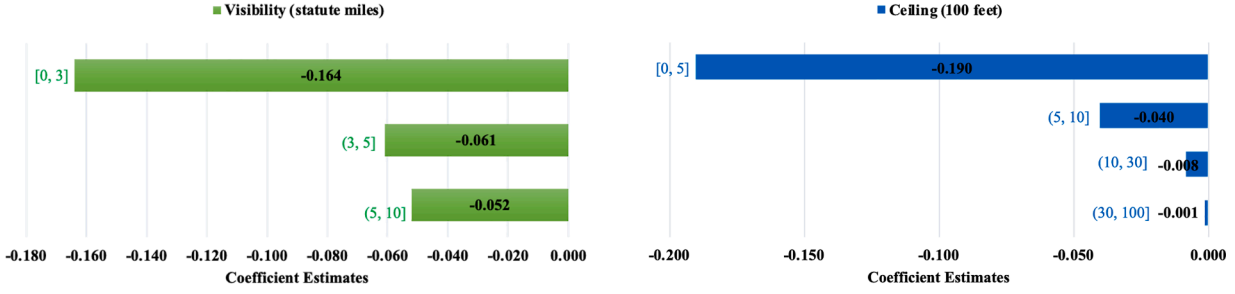

Fig. 4. Bar plot of visibility and ceiling effects.

Table 7
Counterfactual analysis results.

Baseline go-around rate: 0.343%			
Variable	**Scenario value**	**Expected GA%**	**% Reduction in GA's under scenario**
AltDev	0	0.240%	29.97%
Speed	77.28		
Energy	1318.72		
Angle	0		
LOS	0	0.248%	27.59%
SpeedDiff	−86.95		
DepQue	−15	0.256%	25.46%
Ceiling$_1$	5	0.257%	25.12%
Ceiling$_2$	10		
Ceiling$_3$	30		
Ceiling$_4$	100		
MC	VMC	0.258%	24.86%
Body	Narrow	0.259%	24.60%
Visibility$_1$	3	0.272%	20.62%
Visibility$_2$	5		
Visibility$_3$	10		
Wind	0	0.273%	20.47%
NoLead	1	0.276%	19.47%
$\widehat{ROB}$	146.79	0.277%	19.39%
ArrQue	−15	0.290%	15.51%
RwyChange	0	0.306%	10.86%
GaGap	1440	0.307%	10.41%
GaCnt	0		
AirlineIntl	0	0.310%	9.52%
RwyCnt	0	0.313%	8.88%

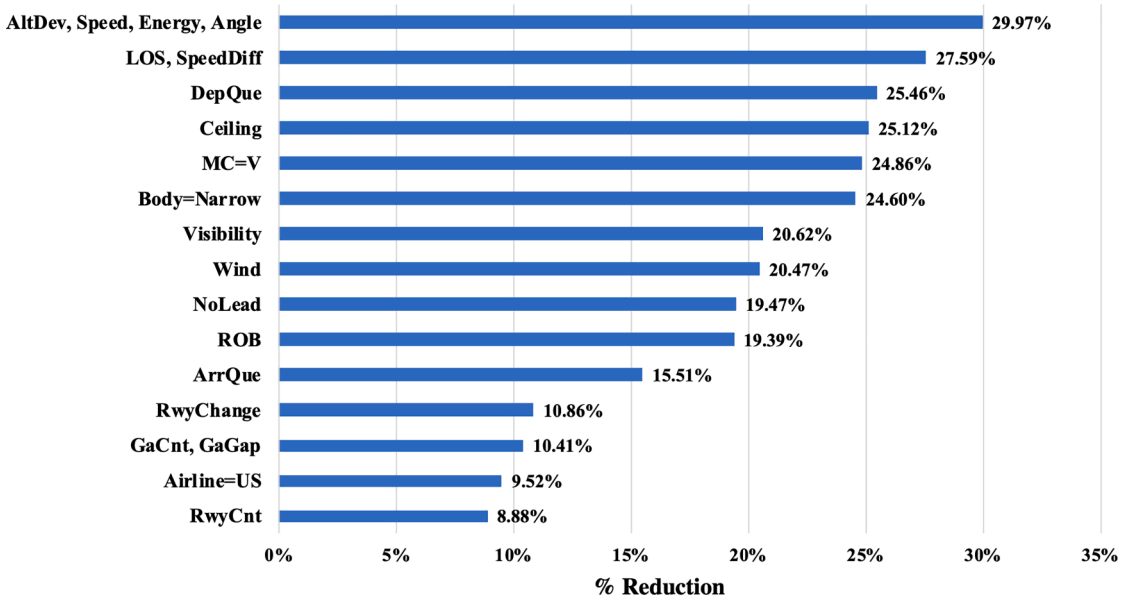

Fig. 5. Relative variable contribution in reducing go-around.

6. Conclusions and future work

In this paper, we develop the principal component logistic regression model to quantify the contribution of a wide range of factors to go-around occurrence. Specifically, we have designed a trajectory-based go-around detection algorithm and applied it to JFK arrival flights in 2018. Multiple datasets have been fused to capture features that may influence flight approach procedures and therefore help understand the causes of go-around occurrence. In developing the feature engineering, we have collected features from the dataset directly, and used domain knowledge to derive features, such as loss of separation and runway occupancy buffer. We then established statistical relationships between go-around occurrence with those derived features and estimated their effects using PCLR model and

factor loading analysis. Lastly, we quantify the contribution of various features to go-around occurrence through counterfactual analysis. Conclusions are in line with research using full-flight simulator trials (Campbell et al., 2018a; 2018b), interviews with ATC controllers and pilots (Blajev and Curtis, 2017), and a realized trajectory dataset (Wang et al., 2016).

As far as the authors know, this is the first work to detect, model, and interpret go-around occurrence from surveillance data, considering a broad set of environmental and operational variables. This enables us to assess the relative importance of a wide range of factors in determining go-around probability. We find that there is no single dominant factor. Factors in the top tier of importance include the state of the subject aircraft, its separation and speed difference from the aircraft in front, and factors related to visibility, cloud ceiling, and the subject aircraft type. Among these factors the first two are, in principle, subject to improvement through pilot and controller training, and thus inviting targets for initiatives to reduce go-arounds. These conclusions must be qualified by the strong assumption that individual features influence go-around occurrence via their contributions to factors. Larger data sets are required to overcome the multi-collinearity between features so that feature coefficients can be estimated directly rather than through principal components.

In addition to the scientific contribution of this paper, it has a variety of practical applications. This research could lead to a real-time monitoring tool that can anticipate, and perhaps remediate, situations in which there is a substantial risk of go-arounds. The model can also supply tactical instructions for controllers and pilots about the probability of go-arounds under varying conditions during approach procedures. It would be helpful for decision support monitoring and prediction-based alerting in advance to improve flight approach safety and airport efficiency. Our results can also inform strategies to reduce go-arounds by identifying the most salient contributing factors, some of which may be mitigated. Also, by summarizing historical patterns of go-around occurrence, our study can augment the limited individual experience of air traffic controllers and pilots, and this informs their judgment about whether a go-around is warranted.

Several improvements can be built upon the presented work. As noted above, one important direction is to overcome multi-collinearity among different features, presumably by employing a larger data set across multiple airports. Another interesting extension is to develop models at other distances and explore the evolutionary impacts of feature contributions on go-around occurrence. In this manuscript, the PCLR model is based on the features that are available when a flight is at 5 nm from its landing runway threshold. The algorithms and the estimation procedures can be applied to features spaces defined at different information cutoff gates from 10 nm to 1 nm. Thirdly, our methods can be extended to other types of atypical flight approach events, such as the unstabilized approach, short approach, dogleg approach, etc. Lastly, other contributing factors, such as crew-controller communications (Dai et al., 2018) and commercial pressure to maintain flight schedules, may be considered in future work.

CRediT authorship contribution statement

Lu Dai: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Visualization, Writing - original draft. **Yulin Liu:** Conceptualization, Methodology, Writing - review & editing. **Mark Hansen:** Conceptualization, Methodology, Writing - review & editing, Supervision, Project administration.

Acknowledgments

The authors would like to thank ATAC corporation and Metron Scientific Solution, Inc. for data collection and supporting analysis of flight anomalies.

Funding: This work was supported by the National Aeronautics and Space Administration (NASA) [grant number NNA16BE49C].

Declaration of Interest

None.

References

Aguilera, A.M., Escabias, M., Valderrama, M.J., 2006. Using principal components for estimating logistic regression with high-dimensional multicollinear data. Comput. Stat. Data Anal. 50 (8), 1905–1924.

Amelink, M.H.J., Mulder, M., van Paassen, M.M., Flach, J., 2005. Theoretical foundations for a total energy-based perspective flight-path display. The Int. J. Aviat. Psychol. 15 (3), 205–231.

Blajev, T., Curtis, C.W., 2017. Go-around decision-making and execution project. Flight Safety Foundation.

Boeing Commercial, A., 2018. Statistical summary of commercial jet airplane accidents.

Campbell, A.M., Zaal, P.M.T., Schroeder, J.A., Shah, S.R., 2018. Development of possible go-around criteria for transport aircraft. Aviation Technology, Integration, and Operations Conference, Atlanta, Georgia, AIAA AVIATION Forum.

Campbell, A., Zaal, P., Shah, S., Schroeder, J.A., 2019. Pilot Evaluation of Proposed Go-Around Criteria for Transport Aircraft. AIAA Aviation 2019 Forum, American Institute of Aeronautics and Astronautics.

Campbell, A., Zaal, P., Schroeder, J.A., Shah, S., 2018a. Development of Possible Go-Around Criteria for Transport Aircraft. 2018 Aviation Technology, Integration, and Operations Conference. American Institute of Aeronautics and Astronautics.

Carter Hill, R., Fomby, T.B., Johnson, S.R., 1977. Component selection norms for principal components regression. Communicat Statist - Theory Methods 6 (4), 309–334.

Chavent, M., Kuentz-Simonet, V., Labenne, A., Saracco, J., 2017. Multivariate Analysis of Mixed Data: The R Package PCAmixdata.

Dai, L., Hansen, M., 2020. Real-time Prediction of Runway Occupancy Buffer. International Conference on Artificial Intelligence and Data Analytics for Air Transportation, Singapore, IEEE Xplore Digital Library and Scopus.

Dai, L., Liu, Y., Hansen, M., 2018. In Search of the Upper Limit to Air Traffic Control Communication. International Conference for Research in Air Transportation, Barcelona, Spain, 2018.
Dai, L., Liu, Y., Hansen, M., 2019. Modeling Go-around Occurrence. Thirteenth USA/Europe Air Traffic Management Research and Development Seminar (ATM2019), Vienna, Austria.
Dehais, F., Behrend, J., Peysakhovich, V., Causse, M., Wickens, C.D., 2017. Pilot flying and pilot monitoring's aircraft state awareness during go-around execution in aviation: a behavioral and eye tracking study. Int. J. Aerospace Psychol. 27 (1–2), 15–28.
Deshmukh, R., Sun, D., Hwang, I., 2019. Data-driven precursor detection algorithm for terminal airspace operations. Thirteenth USA/Europe Air Traffic Management Research and Development Seminar (ATM2019), Vienna, Austria.
Donavalli, B., 2016. Impact of weather factors on go-around occurrence. The University of Texas at Arlington, Master of Science.
FAA, 1991. Airport Capacity and Delay Analyses. FAA.
FAA, 2007. Runway Safety Area Improvements in the United States. International Civil Aviation Organization.
FAA, 2015. Runway Safety - Runway Incursions. Retrieved 07-29-2019, 2015.
FAA, 2018. Stabilized Approach and Go-around.
FAA, 2019. Air traffic by the numbers.
Gluck, J., Tyagi, A., Grushin, A., Miller, D., Voronin, S., Nanda, J., Oza, N.C., 2019. Too fast, too low, and too close: improved real time safety assurance of the national airspace using Long Short Term Memory. AIAA Scitech 2019 Forum.
Holland, P.W., Rubin, D.B., 1987. Causal inference in retrospective studies. ETS Research Report Series 1987 (1), 203–231.
IATA, 2016. Unstable Approaches: Risk Mitigation Policies, Procedures and Best Practices. International Air Transport Association.
Jolliffe, I.T., 1986. Principal components in regression analysis. Principal component analysis, Springer: 129-155.
Jou, R.-C., Kuo, C.-W., Tang, M.-L., 2013. A study of job stress and turnover tendency among air traffic controllers: The mediating effects of job satisfaction. Transport. Res. E: Logist. Transport. Rev. 57, 95–104.
Kaiser, H.F., 1991. Coefficient alpha for a principal component and the Kaiser-Guttman rule. Psychol. Rep. 68 (3), 855–858.
Kennedy, Q., Taylor, J.L., Reade, G., Yesavage, J.A., 2010. Age and expertise effects in aviation decision making and flight control in a flight simulator. Aviat. Space Environ. Med. 81 (5), 489–497.
Kim, J., Palmisano, S.A., Ash, A., Allison, R.S., 2008. Pilot gaze and glideslope control. ACM Trans. Appl. Percept. 7 (3), 1–18.
Lee, J.-G., Han, J., Whang, K.-Y., 2007. Trajectory clustering: a partition-and-group framework. Proceedings of the 2007 ACM SIGMOD international conference on Management of data, Beijing, China, ACM.
Longman, R.S., Cota, A.A., Holden, R.R., Fekken, G.C., 1989. A Regression equation for the parallel analysis criterion in principal components analysis: mean and 95th percentile eigenvalues. Multivar. Behav. Res. 24 (1), 59–69.
Mason, R.L., Gunst, R.F., 1985. Selecting principal components in regression. Statistics & Probability Letters 3 (6), 299–301.
Prats, X., Puig, V., Quevedo, J., Nejjari, F., 2010. Multi-objective optimisation for aircraft departure trajectories minimising noise annoyance. Transport. Res. C: Emerg. Technol. 18 (6), 975–989.
Ruiz, S., Piera, M.A., Del Pozo, I., 2013. A medium term conflict detection and resolution system for terminal maneuvering area based on spatial data structures and 4D trajectories. Transport. Res. C: Emerg. Technol. 26, 396–417.
Shepherd, R., Cassell, R., Thapa, R., Lee, D., 1997. A reduced aircraft separation risk assessment model.
Sherry, L., Wang, Z., Kerkoub Kourdali, H., Shortle, J., 2013. Big data analysis of irregular operations: Aborted approaches and their underlying factors.
Shortle, J., Sherry, L., 2013. A Model for Investigating the Interaction Between Go-Arounds and Runway Throughput. 2013 Aviation Technology, Integration, and Operations Conference, American Institute of Aeronautics and Astronautics.
Simpson, R.W., 1986. Potential impacts of advanced technologies on the ATC capacity on high-density terminal areas [microform] / Robert W. Simpson, Amedeo R. Odoni, and Francisco Salas-Roche. [Washington, D.C.] : [Springfield, Va, National Aeronautics and Space Administration, Scientific and Technical Information Branch ; For sale by the National Technical Information Service].
The Port Authority of New York & New Jersey (2020). Runways at John F. Kennedy International Airport. T. P. A. o. N. Y. N. Jersey.
Tošic, V., Horonjeff, R., 1976. Effect of multiple path approach procedures on runway landing capacity. Transp. Res. 10 (5), 319–329.
Wang, Z., Sherry, L., Shortle, J., 2016. Feasibility of using historical flight track data to nowcast unstable approaches. 2016 Integrated Communications Navigation and Surveillance (ICNS).
Zaal, P., Campbell, A., Schroeder, J.A., Shah, S., 2019. Validation of Proposed Go-Around Criteria Under Various Environmental Conditions. AIAA Aviation 2019 Forum, American Institute of Aeronautics and Astronautics.