Thirteenth USA/Europe Air Traffic Management Research and Development Seminar (ATM2019)

Modelling Go-Around Occurrence

Lu Dai, Yulin Liu, Mark Hansen
Department of Civil and Environmental Engineering
University of California, Berkeley
Berkeley, California, United States
dailu@berkeley.edu
liuyulin101@berkeley.edu
mhansen@ce.berkeley.edu

***Abstract*—Go-around is an aborted landing of an aircraft that is on final approach. In this work, we model the impact of separation, airport condition, weather condition, and trajectory performance on go-around occurrence. A trajectory-based go-around detection algorithm has been developed and applied to the last three-quarter of JFK arrival flights in 2018. Principal component regression (PCR) model, with a retrospective causal inference design, has been estimated and further been used in counterfactual scenarios to reveal the causal correlations between factors of interest and go-around occurrence. Our results suggest that airport ceiling and aircraft speed control are the two most salient factors in causing go-arounds.**

Keywords-component; go-around; causal analysis; safety; logistic regression; principal component regression

I. INTRODUCTION

NextGen in the United States and SESAR in Europe are intended to improve Air Traffic Management (ATM) system performance through satellite-based navigation and digital data communications. Although the National Airspace System (NAS) is one of the safest and efficient transportation infrastructures, growing air traffic demand and the implementation of autonomous NextGen technologies create risks to NAS safety and efficiency. Boeing analyzed worldwide commercial flights from 1959 to 2017 and found that the number of fatalities per year has remained fairly stable. From 2008 through 2017, 44% of the fatalities and 49% of the accidents occurred during the final approach and landing stages, comparing to 23% of fatalities and 11% of accidents at the cruise stage [1].

A common procedure, called go-around, is initiated by either the pilot or the controller to abort the landing of an aircraft that is on final approach under certain conditions [2]. Those conditions include unstabilized approach, sudden change of weather condition (e.g., wind shear), obstructions on the runway, and aircraft overshoot. From 2012 to 2017, the average percent of go-arounds reported by FAA across all Core 30 airports in US, is 0.3% [2]. While go-around as a risk mitigation tool ensures aircrafts safety in the ATM, it is an operational anomaly that degrades the system efficiency significantly. First of all, although pre-trained, go-around itself is a challenging maneuver. Studies have been conducted to evaluate the performance of flight crews during an unexpected go-around maneuver using survey data [3] and flight simulator [4]. Second, the outcome of go-around can be hazardous. About 10% of go arounds result in exceeding aircraft performance limits, or fuel emergencies [3]. Lastly, go-around leads to flight inefficiency, increasing air traffic controller workload and passenger delay, and decreasing airport throughput.

In this paper, we estimate models to predict the go-around occurrence given factors of interest, and further quantitatively understand how these factors contribute to the go-around. To be more specific, this paper first proposes an algorithm to detect go-around occurrence and applies it to the last three quarters of JFK arrival flight track dataset in 2018. Second, we derive a set of contributing factors, which includes aircraft separation, traffic volume, runway configurations, weather, and trajectory performance. Then we build principal component regression (PCR) model to use those factors to predict go-around occurrence. Lastly, we use the estimated model to construct counterfactual scenarios to estimate the contributions of different factors. The results of our analysis 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. 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. Finally, by summarizing historical patterns of go-around occurrence, our study can augment the limited individual experience of air traffic controllers [5].

Current literature about go-arounds considers several aspects, including go-around decision-making policy, the performance of go-around maneuvers [6] and how to optimize the go-around operations [7]. Flight Safety Foundation [3] developed a psychological survey to examine flight crew go-around decision-making and the outcome of go-arounds. New stabilized approach and go-around guidelines was proposed based on the survey results. Four groups of factors are considered when designing the stable approach criteria: flight path profile (vertical and lateral), configuration (flaps, gear and speed brakes), flight energy (rate of descent, speed and thrust), and environmental conditions (runway length, runway

This research is sponsored by NASA.

condition, weather). Ref. [4] added a 3D full-flight pilot simulation data to help with the development of go-around criteria. They found that the go-around decision point occurs above the 100-feet gate, and suggested a 300-feet go-around decision gate. They also found that go-around occurrence is mostly impacted by the reference speed and localizer deviations. Both studies provided insightful guidance for what approach variables flight crew usually consider in deciding whether to execute a go-around. Owing to the limitation of survey and simulation data, additional analysis is needed to investigate the interaction effect with other traffic under different environmental conditions.

Recently, aviation researchers have begun applying machine learning techniques to detect, understand and predict go-arounds using historical flight data. Karboviak et al [8] use several airplane parameters to design a go-around detection tool to classify approach type for General Aviation flights, with an accuracy of 98% tested on 100 student flights. Bro [9] collected 2000 hours of General Aviation training flight data and trained a neural network to predict whether an approach is a landing or go-around event. Low error rates were achieved but the cause of go-around event is unknowable. Manikandan et al [10] used real commercial flight trajectory data, landing airport data, features derived from radar track and weather data, and the corresponding closing rates to the nearest aircraft to developed an algorithm aiming to discover precursors to go-around events. They firstly used Markov Decision Process framework and Inverse Reinforcement Learning (IRL) to generate an expert's nominal time series trajectory. A precursor score was then defined to evaluate a given instant of a time series by comparing its utility with the expert's nominal trajectory. Precursors were identified in an unsupervised manner, and the threshold for determining whether a precursor is strong, was somewhat unreliable.

There is little work analyzing the potential causes of go-arounds. Wang et al [11] built a simulation-based method and a heuristic method to extract unstabilized approach related features from one-month historical commercial flight track data and procedure data (runway elevation, glide path angle, etc.). The results quantified the deviation portion of each unstabilized approach feature violating the stabilized approach criteria. The groundspeed change, speed at Final Approach Fix (FAF), aircraft weight class and rate of descent were discussed. With the same datasets, Ref. [12] used logistic regression model to predict unstable approaches. This work reveals potential causal factors of go-arounds, but fails to consider weather conditions, traffic and other potentially relevant features.

While substantial literature can be found on go-round decision making policy, the performance of go-around operations, and detection and prediction of go-arounds, there is comparatively little work on the causes of go-arounds. Previous work has yet to fully comprehensively investigate features related to procedure execution, traffic separation, weather, airport conditions and trajectory performance (Section III). Our work focuses on advancing the knowledge of go-around incidence, and quantifying the contribution of these features on go-around occurrence. This study also developed an algorithm for detecting go-arounds and thus introducing a profile of go-around occurrence for the analyzed airport. Analysis of causal factors of go-around occurrence will inform further prediction work and deeper exploration in the field of aviation safety, thus improving flight efficiency and safety.

The remainder of this paper is organized as follows. In Section II, we introduce the data sources and how we utilize these datasets for go-around detection and deriving potential factors. Section III describes the derivation of different categories of features. Section IV presents our methodology framework and the model results. In Section V, we conduct a counterfactual analysis to quantify the effect of different factors. Section VI offers the conclusions and discusses the implication and limitation of this study.

II. DATA AND GO-AROUND DETECTION

A. Data Sources

We collect data from two different sources from April 1st to December 24th in 2018 at the JFK airport. After data cleaning and matching, there are on average 525 arrival flights per day in the analyzed airport within the analysis period.

1) Sherlock data warehourse

The first dataset is retrieved from the Integrated Flight Format (IFF) of the Sherlock Data Warehouse, which is a platform for reliable ATM data collection, archiving, processing, query, and delivery [13]. IFF dataset records flight summary (time, aircraft ID, aircraft type, origin, destination, operation type), flight plan (route, Estimated Time of Arrival (ETA), etc.), and track points (latitude, longitude, altitude, ground speed, course, rate of climb, etc.), which are gathered from 76 FAA facilities and formatted by ATAC corporation. Arrival trajectories has been filtered to 400 nautical miles centered on the analyzed airport for each flight.

Reduced Data (RD) summary, which includes the information of landing runway and time, is also used in this study. The RD summary and the IFF data have been processed and merged on a daily basis for each flight arriving at JFK.

2) Airport information

The second dataset, which comes from the Aviation System Performance Metrics (ASPM), provides airport information for each quarter hour, including meteorological conditions (i.e., IMC and VMC), ceiling (in feet), visibility (in statute miles), wind speed (in knots), wind angle (degree), arrival demand (counts) and airport runway configuration, which will be used to derive features for our analysis (Section III).

B. Detection of Go-Around Occurrence

A go-around occurs when a flight aborts the landing on final approach, firstly decreasing its altitude and distance to the airport, then climbing and flying away from the airport for another approach and landing, as shown in the right subplot in Figure 1. Our go-around algorithm is presented in the following steps.

(1) Query the track point data from IFF dataset for a given flight, and extract the 4D trajectory (time, latitude, longitude, altitude), as shown in Figure 2.

(2) Piecewise linear regression is applied to identify points at which the slope of the altitude evolution curve is changed.

(3) Each flight trajectory will be 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 a default value of 5500 feet.
- The total altitude gain during the ascent must not be less than a default value of 400 feet.

(4) Define the final approach trajectory endpoint of each aircraft. For flights that meet the criteria in (3), the final approach trajectory endpoint is the point at which the altitude starts increasing; for other flights, it is the landing point from RD files.

(5) Identify the final approach trajectory segment of each aircraft, which is a five-minute (T_{final}) trajectory segment ending at the final approach endpoint defined in Step (4).

(6) For every 5-minute flight trajectory segment, calculate its distance with the all available runways (by configuration) using formulas (1), (2) and (3) in [14]. Each track point votes for the closest runway line segment.

(7) Identify actual landing runway for each flight using the most voted runway from step (6)[1]. For each track point of a given flight, calculate the distance to the touchdown zone markings of the corresponding landing runway.

(8) Piecewise linear regression is applied to identify points at which the slope of the distance evolution curve is changed.

(9) Each flight trajectory will be 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 will not exceed a default value of 1500 feet.
- Go-around must occur within the 10-nautical-mile range of the airport, to distinguish go-arounds from aircrafts being vectored or in holding patterns.
- The ascending segment of a go-around trajectory must intersect with a 10 nautical-mile-radius cylinder centered at the airport.

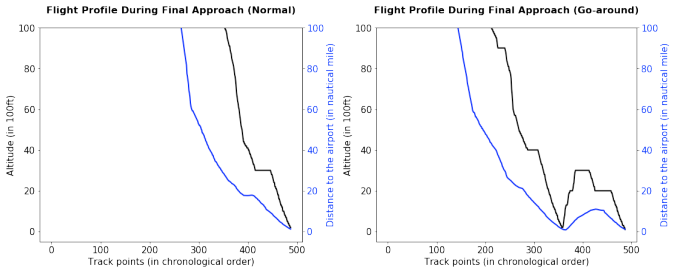

Figure 1. Profiles of normal landing flight (left) and go-around flight (right)

1. For non-go-around flights which have recorded landing runway in the RD files, we use the recorded runway directly. For other flights, which are go-around flights or miss records in the RD files, the approach landing runway is the one that receives most votes from track points in the vector.

2. SME: Michael Hanowsky *(Leigh Fisher Consulting)*, William Dunlay *(WJDunlay Consulting)*, and the authors *(University of California, Berkeley)*.

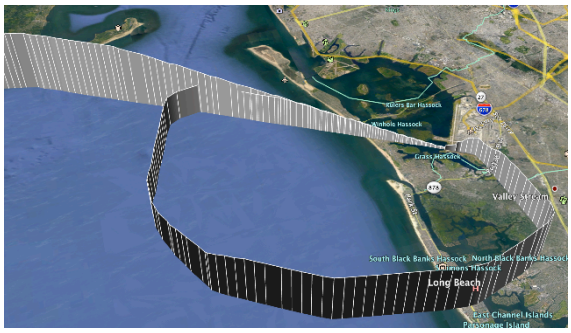

Figure 2. 4D trajectory visualization of go-around flight

We implemented our algorithm and applied it to all JFK arrival flights except military flights, general aviation and helicopters, and detected 691 go-arounds from April 1st to December 24th in 2018. We further validated those go-arounds by visualization inspection from Subject-Matter Expert (SME)[2] consultations. We compare the descending profiles of two aircrafts – a detected go-around flight and a non-go-around flight – in Figure 1. , the altitude and distance to the airport of the left normal flight have been declining during the whole final approach process. However, the go-around flights on the right first descend, then ascents, and descents again for both the altitude profile and the distance to airport profile. We also present the 4D trajectory of the go-around flight as shown in Figure 2.

III. Feature Engineering

Using data collected in Section II.A, observations are obtained for every flight operating during the study period with five categories of features. The derivations of these features are described in this section.

A. In-trail Relationship Features

Separation is defined as the distance, either horizontal or vertical, between the leading and trailing aircraft. The minimum required separation depends on the relative weight class (size) of two aircrafts and the operational environment conditions (visibility). Loss of separation occurs whenever specified separation minima are breached. Therefore, to capture the separation causal effect, four variables are derived from the 4D trajectory information:

1) Loss of separation: This variable calculates the difference between the minimum required separation (FAA standard) and the minimum actual separation between the leading and trailing aircraft pair. We expect larger magnitude of loss of separation (in nautical miles) increases the probability of go-arounds. The algorithm for calculating the loss of separation has two steps – finding leading and trailing aircraft pair and calculating separations for the aircraft pair. The detailed algorithm is illustrated as below.

(1) Group flights with the same (matched) landing runway, and sort them in chronological order.

(2) For each group in step (1), we create a list of tuples where each tuple contains two consecutive aircrafts that have

been sorted. Within each tuple, if the endpoint time difference of the two aircrafts is smaller than 10 minutes, then we define them as a leading-trailing aircraft pair. Otherwise we remove the trailing flight from the tuple.

(3) For each flight, we use Discrete Fourier Transform (DFT) [19] to find the extrapolated timestamps at which the analyzed (trailing) flight reaches certain distance to its landing runway ($dist_i$, i = 1, 2, …, 10, in nautical mile).

(4) We use DFT again to extrapolate the locations (latitude, longitude, altitude) of both leading and trailing aircrafts at the extrapolated timestamps found in (3).

(5) The distance between every two extrapolated points is calculated. An example is shown in Figure 3. For the analyzed (trailing) flight labeled in blue with 10-point extrapolation, 10 separations ($dist_i$, i = 1, 2, … ,10), which are the black two-way arrow in Figure 3. will be calculated $S_a = dist_i$.

(6) Obtain the separation minima S_t from FAA Wake Separation Standards based on the weight class of leading and trailing aircrafts. Thus, the loss of separation is $S_l = max\,(0, S_t - S_a)$, and will be directly employed as continuous variables in the model.

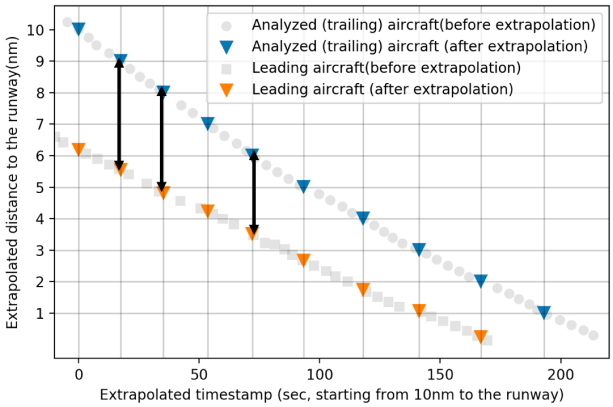

Figure 3. A synthetic example for extrapolation

2) Speed difference: Following the step (3) in calculating the loss of separation, we also obtain the extrapolated speed of leading and trailing aircrafts at the same extrpolated timestamps. The numerical value of speed difference will be directly employed as continuous variables in the model.

3) Chasing: A dummy variable is created based on the extrapolated speeds, and equals to 1 when the analyzed (trailing) flight ground speed is greater than the leading flight ground speed, indicating the chasing senario during final approach.

4) No leading: We create a dummy variable where 1 represents no leading aircraft and 0 otherwise.

B. Airport and Weather Features

We expected that runway configuration change, arrival traffic, airport capacity, visibility, ceiling, and wind condition could also trigger a go-around.

1) Arrival demand: This variable comes directly from the ASPM dataset to represent the number of aircrafts intending to arrive for the observed 15-minute interval.

2) Arrival rate: This variable comes directly from the ASPM dataset to represent the Airport supplied Arrival Rate (AAR) for capacity.

3) The change of runway configuration: Runway configurations are recorded every 15 minutes in the ASPM airport information data. For a given flight, this variable is set to 1 if the used runway configuration during the observed time period is different from either the preceding 15-minute period, or the succeeding 15-minute period, and 0 otherwise.

4) Wind condition: We consider four types of wind speeds – headwind, tailwind, crosswind, and variable wind. Since headwind is favored during approach, we subtract the headwind component from the original wind speed using trigonometric calculations with the information of wind speed, wind angle and flight corresponding landing runway. For each flight, if the wind direction record at the trajectory's endpoint time is "VRB", then this variable is set to the wind speed record. Otherwise, the variable wind speed is set to wind speed subtracting the headwind component.

5) Visibility: Visibility ranges from 0 to 10 statute miles in the ASPM dataset. We discretized the visibility variable into three categories: [0, 3], [3, 5] and [5, 10].

6) Ceiling: Ceiling ranges from 0 to 999 in hundreds of feet in the ASPM dataset. Similar to the visibility variable, we discretized the ceiling variable into 4 categories: [0, 5), [5, 10), [10, 30), [30, 999].

7) Meteorological condition: Two types of meteorological conditions are considered – VMC and IMC. We convert this variable into a dummy variable where 1 represents VMC condition and 0 otherwise.

C. Clustering Effect Features

From the go-around detection results, we observed that a go-around was more likely to occur when leading aircrafts initiated go-arounds. Therefore, two variables were derived to capture such effect – *closest go-around time* and *go-around count*. As shown in Figure 4., for each flight (yellow star), the closest go-around time is the minimal time interval between its approaching time and initiation time of all (other) go-around flights (T, in hours). The go-around count is the number of go-around flights (excluding any go-around by the subject flight) within a one-hour time period of the flight of interest. The variable value is equal to 3 in Figure 4., no matter whether the starred flight is go-around or not.

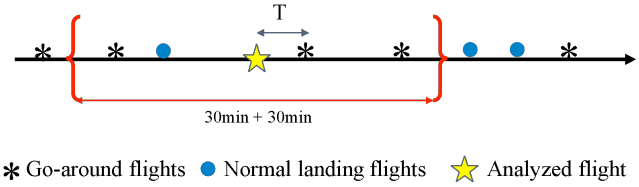

Figure 4. Clustering effect features

D. *Trajectory Performance Features*

Figure 5. visualizes some of the metrics which are derived from the trajectory information to measure the trajectory performance. The solid blue line is the Extended Runway Centerline (ERC) of 31R. When a flight, represented as red dot in the figure, intercepts the distance arc (e.g. 5nm), the flight altitude (in 100 feet), ground speed (in knots), perpendicular distance to ERC (in nautical miles), angle with ERC (in degree), glideslope angle (in degree, see Figure 6) are calculated at this moment.

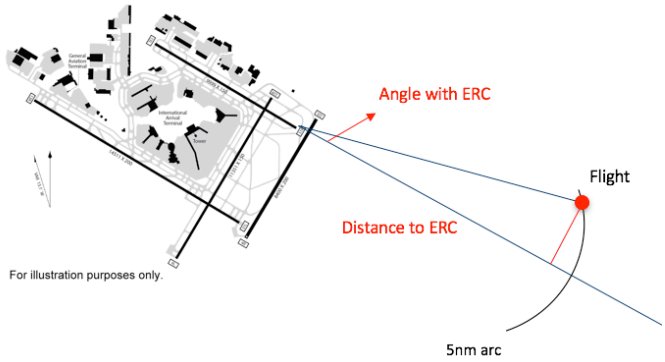

Figure 5. Diagram of trajectory performance features

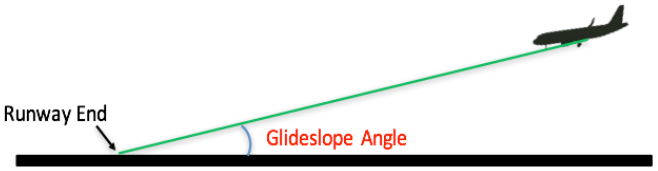

Figure 6. Diagram of glideslope angle

We also calculate the flight kinetic energy during the approach process. However, the traditional kinetic energy measurement requires information on aircraft mass. The fuel consumption profiles are unavailable at this time. Thus, we use the energy height metric [22], which is a function of altitude and ground speed, to identify the energy state of each flight.

$$H_{energy} = \mathrm{h}_i^d + \frac{(v_i^d)^2}{2g} \text{ (feet)}$$

where, H_{energy} is the kinetic energy height in feet, h_i^d is altitude of aircraft i at d-nautical miles away from the landing runway touch down markings, v_i^d is the ground speed of aircraft i at d-nautical miles away from the landing runway touch down markings, g is the gravitational acceleration. This metric only uses surveillance data.

E. *Flight-specific Features*

Flight approach performance relies on pilot experience. Although flight crew information is not available for this study, we categories airline as international and domestic carriers. We expect that pilots work for domestic (U.S.) airlines are more experienced with landing in JFK airport. Aircraft type is also considered by categorizing as narrow body and wide body.

1) Airline: 1 if the analyzed flight is operated by US carriers and 0 otherwise.

2) Aircraft Type: 1 if the analyzed flight is wide body aircraft and 0 otherwise.

3) Landing runway: There are 8 runways in JFK airport – 04L, 04R, 22L, 22R, 13L, 13R, 31L, 31R. This variable is set to 1 if its corresponding runway is the landing runway of analyzed aircraft. The purpose is to capture different landing patterns for different runways.

Model must take consideration of which of the features can be evaluated at a certain time prior to the go-around occurs. In this study, we assume a go-around is executed when its altitude starts increasing during the approach based on our detection algorithm. All the features – in-trail relationship features, airport and weather condition features, clustering effect features, trajectory performance features and flight-specific features, are derived and recorded at the moment when the analyzed flight is at 5 nautical miles away from the touchdown zone markings of its landing runway. Extrapolation technique is also applied to make sure that no information from future is included in the 5nm feature space. After preprocessing and matching flight trajectories with the 5nm features, our final dataset has a total of 489 go-around occurrences within 0-5 nm (5nm is exclusive) away from the airport over 140, 807 observations. The summary of variables is presented in TABLE I.

IV. Methodology and Model Results

A. *Retrospective Causal Inference Model*

There are three types of causal inference methods – randomized control method, prospective method, and retrospective method [15]. While the first two methods require randomized experimental data, the last method can be applied to observational data. Therefore, in this study, we use the retrospective method to investigate the causal relations among the go-around occurrence and features that described in TABLE I.

B. *Logistic Regression Model*

A binary logistic regression model was estimated to relate whether a flight initiates a go-around or not with contributing factors in TABLE I. In this study, we first estimate a vanilla logit model with model specification (1), where V is the utility function, x_i's are all contributing factors introduced in TABLE I. , and β_i's are the associated coefficients. Thus, the probability of an aircraft initiating go-around can be written as in equation (2), where $P_r(y_i = 1|X)$ indicates the probability of go-around.

$$V = \sum_{i \in \mathcal{I}} \beta_i \cdot x_i \quad (1)$$

$$\Pr(y_i = 1|X) = \frac{1}{1+\exp(-V)} \quad (2)$$

TABLE I. MODEL VARIABLES

Variable Code	Variable Description (per flight)	
	Dependent Variable	***Category***
GA	1 if flight is detected as a go-around occurring within [0, 5) nautical miles to the runway touchdown marking zone, 0 otherwise.	Detected
	Independent Variable	***Category***
Loss_of_sep	The loss of separation at 5nm to the runway touchdown (in nautical miles)	In-trail relationship features
Speed_diff	The difference of gound speed between leading and trailing aircrafts at 5nm to the runway touchdown (in knots)	
Chasing	1 if the analyzed (trailing) flight ground speed is greater than the leaidng flight ground speed at 5nm to the runway touchdown, 0 otherwise	
No_leading	1 if there is no leading aircraft, 0 otherwise	
Arr_demand	The number of intended landing aircrafts during the observed 15-minute interval (counts)	Airport and weather features
Arr_rate	Airport supplied Arrival Rate for capacity (counts)	
Rwy_change	1 if the runway configuration is changed during the observed 15-minute interval, otherwise 0	
Sub-wind	Wind speed where the headwind component is subtracted (in knots)	
Visibility_i	Airport visibility (i = 1,2,3; intervals are [0, 3), [3, 5), [5, 10] in miles)	
Ceiling_i	Airport ceiling (i =1, 2, 3, 4; intervals are [0, 5), [5, 10), [10, 30), [30, 999] (in 100 feet)	
MC	"V" for VMC, "I" for IMC operations	
Closest_time	The minimal time interval between the approaching time of this flight and the initiation time of all (other) go-arounds (in hours)	Clustering effect features
GA_occur_ct	The number of go-around flights within 1-hour time period of the flight of interest (counts)	
Alt	Flight altitude when flight is at 5nm to the runway touchdown (in 100 feet)	Trajectory performance features
Horiz_dist	Penperdicular distance to the ERC when flight is at 5nm to the runway touchdown (in nm)	
Speed	Flight ground speed when flight is at 5nm to the runway touchdown (in knots)	
Engergy	Kinetic energy height when flight is at 5nm to the runway touchown (in feet)	
AC_airline	1 if flight is operated by international airline, 0 if flight is operated by US carriers	Flight-specific features
AC_Type	1 if aircrafts is wide body, 0 otherwise	
Landing_rwy	Dummy variable for landing runway	

The estimation results are presented in TABLE II. Model I. The majority of coefficients are not significant at 0.05 level, and most of which have unexpected signs. For example, the estimates for the visibility and ceiling variables suggest that flights land on an airport with good visibility and ceiling conditions would have a higher probability of go-around, which is not plausible in practice. By carefully examining the error structures, we find that many independent variables used in the model are highly correlated (a.k.a., multicollinearity). Therefore, the vanilla logistic regression model fails to give us a proper understanding of the causal effect, and techniques need to be conducted to de-correlate the original feature vectors.

C. Principal Component Regression

To handle the multicollinearity problem, we apply principal component analysis (PCA) to de-correlate and reduce the dimensionality of the original feature space. To be more specific, instead of regressing the dependent variable on the explanatory variables directly, the principal components of the explanatory variables were used as regressors [16].

After the preprocessing and mapping procedures shown in Section III, our feature space ends up with a dimension of (140807, 25), with a categorical data matrix Z_1 (140807, 7) and a numerical data matrix Z_2 (140807, 18). The categorical variables are *Rwy_change* (2 classes), *Chasing* (2 classes), *AC_airline* (2 classes), *AC_body* (2 classes), *Landing_rwy* (8 classes), *MC* (2 classes) and *No_leading* (2 classes), we therefore have m = 20 categories. Since the dataset is a mixture of categorical and quantitative variables, an appropriate treatment of mixed data types, especially the categorical variables, is required. In this study, we applied the PCA-mixed algorithm introduced by [17] to handle such problem (Step (1), (2), (3)). The detailed algorithm is given as follows.

- Standardize the numerical data matrix Z_1^s and center the categorical data matrix Z_2^c.
- Build a numerical data matrix $Z = (Z_1^s | Z_2^c)$ of dimension (140807, 18+20), a diagonal matrix N of the weights of the rows, and a diagonal matrix M of the weights of the columns. Apply the generalized singular value decomposition (GSVD) of Z with the diagonal metrics of the weights N and M

$$Z = U\Lambda V^t \quad (3)$$

- The set of factor scores for rows are computed as

$$F = U\Lambda. \quad (4)$$

- The columns of F are the principal components that serve as inputs to the model.
- Select the first 19 principal components which explain 90.4% of the total variance. The number of principal components was chosen based on the rule that eigenvalue of principal components is greater than 0.6. The relationship between selected principal components F' and the independent variables X can be expressed as

$$F' = XW, \quad (5)$$

Where W is a weight matrix of ratio explained of independent variables by the selected principal components.

TABLE II. MODEL ESTIMATION RESULTS

Observations: 140, 807 Go-arounds: 489	**Dependent Variable: Go-around occurrence** Parameter Estimate [significant level] ***(Standard Error)***			
	Model I: Logistic model	***Model II: Logistic principal component regression model***		
Variable Code	***Coef.***	***PC coef. (γ)***	***PC Loading (α)***	***Variable coef. (β)***
Constant	-4.857*** (0.370)	-6.386*** (0.068)		
Arr_rate	-0.121 (0.070)	**PC1** -0.250*** (0.013)	0.49	-0.032
Visibility_1	-0.046 (0.060)		0.75	-0.100
Visibility_2	0.029 (0.074)		0.83	-0.066
Visibility_3	-0.204** (0.079)		0.83	-0.062
Ceiling_1	-0.069 (0.048)		0.69	-0.075
Ceiling_2	0.048 (0.060)		0.85	-0.023
Ceiling_3	0.154 (0.150)		0.85	-0.006
Ceiling_4	-0.126 (0.079)		0.64	-0.001
MC=V	-0.767* (0.343)		0.34	-0.025
Speed	-0.120 (0.219)	**PC2** 0.111*** (0.026)	0.85	0.002
Energy	0.287 (0.200)		0.84	0.0001
Speed_diff	0.033 (0.072)		0.75	0.002
Chasing	-0.020 (0.155)		0.49	0.045
No_leading	0.269 (0.184)		-0.90	0.257
Horiz_dist	0.040 (0.295)	**PC3**	-0.77	-0.028
Rwy_13L	-1.629* (0.706)		-1.51	-0.140
Alt	0.383*** (0.050)	**PC4**	0.44	0.156
AC_airline	0.391** (0.134)	**PC5** 0.131*** (0.032)	1.20	0.376
AC_body	0.308* (0.131)		1.11	0.422
Rwy_31R	-1.168*** (0.239)	**PC6** 0.342*** (0.033)	-1.00	-0.100
Rwy_22L	-0.943*** (0.180)	**PC7**	0.84	-0.113
Arr_demand	0.096* (0.040)	**PC8** 0.377*** (0.034)	0.55	0.023
Sub_wind	0.203*** (0.044)		0.45	0.048
Rwy_22R	-0.781** (0.260)	**PC9**	1.00	0.494
Rwy_change	0.099 (0.166)	**PC10** 0.328*** (0.018)	1.00	0.299
Loss_of_sep	0.166*** (0.019)	**PC12** 0.102*** (0.029)	0.49	1.258
Rwy_04L	0.746*** (0.168)		-1.00	-1.057
Rwy_31L	-1.198*** (0.305)		1.00	-0.144
Closest_time	-0.238*** (0.071)	**PC19** -0.281*** (0.040)	0.40	-0.0009
GA_occur_ct	0.215*** (0.019)		-0.43	0.392
Log Likelihood	-2670.0	-2686.1		

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

- Regress the observed outcomes of go-around occurrence y on the selected principal components F', using maximum likelihood estimation to get a vector of principal components coefficients γ in the logistic regression model (Shown as "PC coef." In TABLE II.). The 3rd, 4th, 7th, 9th, 11th, 13th, 14th, 15th, 16th, 17th, 18th principal components covariates are not significant at 0.05 level, thus removed from the model.
- Transform the vector of principal components coefficients γ back to the scale of the actual covariates, using the eigenvectors corresponding to the selected principal components. The final estimated coefficients of the actual covariates β will be obtained by

$$\beta = W^T \gamma, \tag{6}$$

since the regression is

$$y = F'\gamma = X(W^T\gamma) = X\beta. \tag{7}$$

- Each original variable will be only associated with one principal component according to the PC loadings (α), which makes the interpretation easier. The PC loadings (α), the final estimates of the actual covariates (β) and the principal components coefficients (γ) of the model are shown in TABLE II.

D. Estimation Results

The estimation results for the principal component regression are summarized as Model II in TABLE II. Some insignificant principal components are removed from the table. Original independent variables are presented in the order of its loaded principal component. The sign of principal component loadings (α) shows the direction (principal component) that the original independent variable aligned with, and the absolute values of loadings show the correlation coefficient between the original variables and their aligned principal components. For example, principal component 2 (PC2) captures the negative effects of all visibility variables.

We first observe that the vast majority of the principal components' coefficients (γ) are significant. We then use step (6) and (7) from Section IV.C to find the principal component that best explains each causal factor (α) and further calculate its coefficient in the original feature space (β).

For the in-trail relationship features, the loss of separation variable *(Loss_of_sep)* is loaded in the same direction with the 12th principal component with a correlation of 0.49. The coefficient of the 12th principal component is highly significant with positive sign and positive loading, indicating positive effect of the loss of separation on go-around occurrence. The effects of the aircraft ground speed (*Speed*) and the chasing

(*Chasing*) scenario are both captured in principal components 2. Too high or too low flight ground speed is associated with go-around occurrence. The chasing scenario during final approach increases the probability of go-arounds.

For the weather features, we are interested in the signs and magnitudes of transformed coefficients (β) for airport visibility and ceiling. By looking at the coefficients of different visibility/ceiling categories, we observe that the increase of visibility and ceiling has negative impact on go-around occurrence, and this effect is diminishing when the weather condition is good in itself. Higher wind speed (*Sub_wind*) increases the likelihood of go-around occurrence. The changes of airport runway configuration *(Rwy_change*) has a significant positive impact on go-around occurrence, which could be caused by the change of approach pattern, interrupted landing procedure, and increased crew workload. The probability of go-around occurrence also increases when the airport has high arrival demand (*Arr_demand)*, low airport capacity (*Arr_rate*) or under IMC conditions *(MC=I)*.

For the clustering effect feature, the estimate of the time duration to the closest go-around *(Closet time)* is negative, and the estimate of the number of other go-arounds within 1-hour time window *(GA_occur_ct)* is positive. They are both significant. This implies that go-arounds tend to occur in clusters and that this behavior cannot be fully explained by the other variables in the model.

For the trajectory performance features, flights with high altitude (*Alt*), high ground speed (*Speed*), and high kinetic energy height (*Energy)* at 5nm to the runway touchdown zone would more likely to have go-arounds.

For the flight-specific features, flights operated by non-domestic carriers (*AC_airline*) are more likely to have go-arounds, which could be caused by language issue or the flight crew are not familiar with the airport conditions. Wide body (*AC_body*) aircrafts are more likely to have go-arounds.

V. COUNTERFACTUAL ANALYSIS

To further quantify the contributions for different factors considered in the Model II, counterfactual scenarios are constructed, in which each selected factor is set to the best condition at one time. In our final dataset, there are 489 go-arounds over 140, 807 operations, the baseline go-around rate is 0.347%. The scenario value is set to minimize the probability of go-around occurrence, which we refer as the best condition. Model estimates will be plugged in to predict the corresponding go-around probability for each flight, and the percentage reduction between the baseline go-around rate and the expected go-around rate are calculated, which indicate the factor contribution. The percentage reduction is formulated as

$$\%reduction = \frac{P_{G\text{-}A} - E[P_{G\text{-}A}|x=c]}{P_{G\text{-}A}} \quad (8)$$

Where $P_{G\text{-}A}$ is the expected go-around rate given the current system inputs (0.347%); $E[P_{G\text{-}A}|x = c]$ is the expected go-around rate given the factor x is set to the best condition value c.

TABLE III. COUNTERFACTUAL ANALYSIS RESULTS

Variables	Baseline Go-around Rate: 0.347%			
	Baseline mean	***Scenario value***	***Expected G-A%***	***%reduction***
Ceiling_1	4.935	5	0.278%	19.88%
Ceiling_2	9.254	10		
Ceiling_3	25.003	30		
Ceiling_4	65.254	100		
Speed	162.004	162.004	0.307%	11.53%
Energy	5267.500	5267.500	0.309%	10.95%
Horiz_dist	0.940	0	0.314%	9.51%
Rwy_change	0.081	0	0.320%	7.81%
Chasing	0.582	0	0.328%	5.48%
Arr_rate	12.698	15	0.334%	3.75%
Visibility_1	2.920	3	0.335%	3.46%
Visibility_2	4.681	5		
Visibility_3	4.481	10		
No_leading	0.056	0	0.335%	3.46%
Speed_diff	18.140	0	0.346%	0.29%
Arr_demand	10.587	0	0.346%	0.29%

TABLE III. reports the average value of each feature in the current dataset (baseline mean), the scenario value that is set to minimize the likelihood of go-arounds (scenario value), the expected go-around rates calculated with the counterfactual dataset, and the percentage of reduction for continuous variables. To save space, we only report the variables that can be changed and have contribution to the reduction of go-around rates.

Figure 7 reflects the relative importance of each variable on the reduction of go-around probability. The airport ceiling and aircraft speed control are the two most important factors of go-around occurrence. Go arounds would decrease by 20% if airport ceiling was set to their scenario values. Improving visibility and ceiling indicates a clear view for landing, would reduce go-around occurrence by about 23%. In real operation environment, go-arounds could be initiated because the flight crew fail to have a clear visual condition on the runway, and the landing environment is not ideal.

The aircraft speed control, which can be represented by the numerical aircraft speed value (Speed) and the Kinetic Energy Height (Energy), is also important in reducing go-around occurrence. Stable speed control could reduce go-arounds up to 22% (11.53% + 10.95%). Go-around would decline about 9.5% if the aircraft is properly aligned with the ERC at 5 nautical miles away from its landing runway end.

If there is no change of runway configuration in the analyzed airport, go-around rate would potentially decline about 7.8%. When the airport capacity is high, go-arounds would decline about 3.75%.

Keeping a safe separation and comfortable following speed with the leading aircraft could reduce go-arounds up to 6%. While in the absence of leading aircraft, go-arounds would decline about 3.5%. The airspace traffic has minimal effects on the go-around reduction.

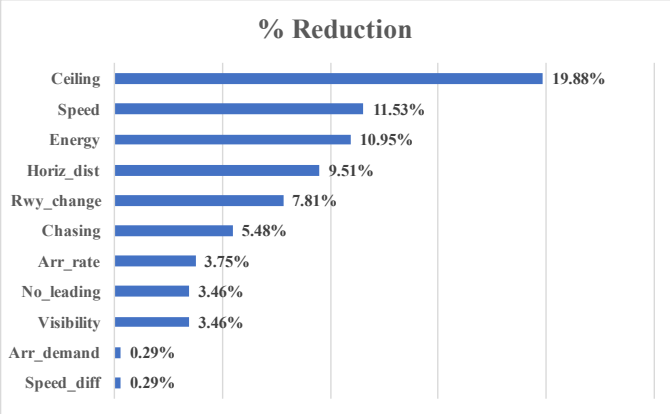

Figure 7. The estimated effect of the loss of projected separation

VI. CONCLUSIONS

This paper presents a modeling framework that incorporates go-around detection, separation analysis, and principal component regression (PCR) techniques to quantitatively understand what factors in ATM system may cause the go-around occurrence. We applied our go-around detection algorithm on the JFK arrival flights for the last three quarters in 2018, and detected 691 go-arounds (489 go-arounds occurred within 0 to 5 nautical miles to the airport).

Our study is not only in line with research using full-flight simulator trials [4] that speed and localizer deviation have strong influences on perceived risk and pilot's go-around decision, but also capturing the effects of environmental and runway variables on go-around decision. Interviews [20] with ATC controllers and pilots state that go-arounds often occur when the aircraft is not properly aligned or wind shear warnings [21]. However, their focus was on managing a singular problem, like descent rate or airspeed control. In this study, most aspects of features including in-trail relationship, airport and weather condition, trajectory performance, flight-specific features and go-around clustering effect have been derived from various data sources and used as causal factors in PCR modeling. The estimation results of the PCR model suggest that factors such as the airport visibility and ceiling, flight perpendicular distance to the ERC, clustering effect, traffic volume, airport capacity, wind condition and loss of separations significantly increase the probability of go-around occurrence. Counterfactual analysis that based on the results of PCR has been conducted to quantify what factors are most central in causing go-arounds, and the results suggest that the airport ceiling and aircraft speed control are the two greatest contributing factors to go-around occurrence.

This study presents a first step in advancing the knowledge of what factors are most salient in triggering go-arounds. Our work seeks to provide the evidence-based information of go-around occurrence, thus helps improve system performance for safety alerting and conflict resolution. A real-time monitoring tool is possible for identifying and accessing substantial risk of go-arounds for decision making. ATM system would have benefit from this study by having better understanding of go-around occurrence, so that precursory mitigating actions can be initiated to prevent the safety and efficiency degradations.

Lastly, we point out several limitations and future work of the current study. The retrospective causal inference provides relative weak evidence for a causal link because of the difficulty in controlling all causal factors. Although we have captured most of the features in ATM system that are important to go-around occurrence, some variables relating to airport surface operation, such as runway occupancy time and runway incursion, are unavailable at the present time.

Further research should progress along several lines. First of all, our PCR regression model could be improved by incorporating a broader range of features. Second, our framework could be extended to analyze other types of flight anomalies for a larger number of airports. In turn, instead of using airport capacity as a causal factor of go-around occurrence, study could be extended to analyze the operational impact on runway capacity and airport efficiency. Lastly, models that focus more explicitly on real-time prediction as opposed to causal inference could be investigated. These models must take greater consideration of which of the features can be evaluated at a certain time prior to the go-around, and whether these features can identify flights with sufficiently high go-around risk to warrant remedial actions.

ACKNOWLEDGMENT

The authors wish to thank NASA for funding support, ATAC corporation and Metron Scientific Solution, Inc. for data collection and supporting analysis of flight anomalies.

REFERENCES

[1] Boeing Commercial Airplanes, "Statistical summary of commercial jet airplane accidents," Worldwide Operations, 2018.

[2] FAA, "Air traffic by the numbers," FAA Report, 2018.

[3] T. Blajev and W. Curtis, "Go-Around Decision-Making and Execution Project,", Final Report to Flight Safety Foundation, March 2017.

[4] A. Campbell, P. Zaal, and J. Schroeder, S. Shah, "Development of possible go-around criteria for transport aircraft," 2018 Aviation Technology, Integration and Operations Conference, pp. 3198.

[5] G. Sreeta, Y. Liu, M. Hansen and A. Pozdnukhov, "Identifying similar days for air traffic management," Journal of Air Transport Management, 65, pp. 144-155, October 2017.

[6] Ö. Goteman,. and S. Dekker, "Flight crew and aircraft performance during RNAV approaches: Studying the effects of throwing new technology at an old problem," Human Factors and Aerospace Safety, pp. 147-164, Routledge, 2018

[7] H. Baomar and P. J. Bentley, "Autonomous landing and go-around of airliners under severe weather conditions using Artificial Neural Networks," IEEE Workshop on Research, Education and Development of Unmanned Aerial Systems (RED-UAS), pp. 162-167, October, 2017.

[8] K. Karboviak, S. Clachar, T. Desell, M. Dusenbury, W. Hedrick, J. Higgins, J. Walberg and B. Wild, "Classifying aircraft approach type in the national general aviation flight information database," International Conference on Computational Science, Springer, Cham, pp. 456-469, 2018.

[9] J. Bro, "FDM machine learning: an investigation into the utility of neural networks as a predictive analytic tool for go-around decision making," Journal of Applied Sciences and Arts, 1, no.3, pp. 3, December 2017.

[10] V. M. Janakiraman, B. Matthews and N. Oza, "Discovery of precursors to adverse events using time series data," Proceedings of the 2016 SIAM International Conference on Data Mining, Society for Industrial and Applied Mathematics, pp. 639-647, June 2016.

[11] Z. Wang, L. Sherry and J. Shortle, "Airspace risk management using surveillance track data: Stabilized approaches," IEEE 2015 Integrated Communication, Navigation and Surveillance Conference (ICNS), pp. W3-1, April 2015.

[12] Z. Wang, L. Sherry and J. Shortle, "Feasibility of using historical flight track data to nowcast unstable approaches," IEEE 2016 Integrated Communications Navigation and Surveillance Conferences (ICNS), pp. 4C1-1, April 2016.

[13] NASA Ames, USRA/Crown/FRA, SGT and ATAC, "Sherlock Data Warehouse," in press, June 2018.

[14] J. G. Lee, J. Han and K. Y. Whang, "Trajectory clustering: a partition-and-group framework," Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data, ACM, pp. 593-604, June 2017.

[15] P. W. Holland and D. B. Rubin, "Causal inference in retrospective studies," ETS Research Report Series, no. 1, pp. 203-231, June 1987.

[16] W. F. Massy, "Principal components regression in exploratory statistical research," Journal of the American Statistical Association, 60, no. 309, pp. 234-256, March 1965.

[17] M. Chavent, V. Kuentz, A. Labenne, B. Liquet and J. Saracco, "Package 'PCAmixdata'," The R Project for Statistical Computing, October 2017.

[18] H. F. Kaiser, "The varimax criterion for analytic rotation in factor analysis," Psychometrika, 23, no. 3, pp. 187-200, September 1958.

[19] M. D. Sacchi, T. J. Ulrych, and C. J. Walker. "Interpolation and extrapolation using a high-resolution discrete Fourier transform.", IEEE Transactions on Signal Processing, 46.1, pp. 31-38, 1998.

[20] S. Martin, "Why calling 'go-around' is an action, not a decision point", in press, March 2019.

[21] Aviation Interviews, "United Airlines Study Guide Questions", in press, 2016.

[22] J. Gluck, A. Tyagi, A. Grushin, D. Miller, S. Voronin, J. Nanda and N. C. Oza, "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, pp. 0400, January, 2019.

AUTHOR BIOGRAPHIES

Lu Dai is a Ph.D. student of Civil and Environmental Engineering at the University of California, Berkeley. Her research interests are in air traffic management, aviation safety, air transportation system performance, and machine learning algorithms. She received her M.S in Transportation Engineering in 2018 from University of California, Berkeley.

Yulin Liu is a Ph.D. candidate of Civil and Environmental Engineering at the University of California, Berkeley. His research interests include air traffic flow management, trajectory prediction and optimization, machine learning algorithms, and deep reinforcement learning in ATM. He received his B.S. degree in Civil Engineering in 2015 from Tsinghua University.

Mark Hansen is a Professor of Civil and Environmental Engineering at University of California, Berkeley, and co-Director of the National Center of Excellence for Aviation Operations Research (NEXTOR-II). Dr. Hansen received his Ph.D. in Engineering Science in 1988 from University of California, Berkeley.