Transportation Research Part C 64 (2016) 45–57

Contents lists available at ScienceDirect

Transportation Research Part C

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

Anomaly detection via a Gaussian Mixture Model for flight operation and safety monitoring

Lishuai Li [a,*], R. John Hansman [b], Rafael Palacios [c], Roy Welsch [d]

[a] Department of Systems Engineering and Engineering Management, City University of Hong Kong, Hong Kong Special Administrative Region
[b] Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[c] Institute for Research in Technology, Comillas Pontifical University, CL Alberto Aguilera 23, 28015 Madrid, Spain
[d] MIT Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

ARTICLE INFO

Article history:
Received 14 July 2015
Received in revised form 21 January 2016
Accepted 21 January 2016
Available online 10 February 2016

Keywords:
Flight safety
Flight data
Flight operations monitoring
Anomaly detection
Cluster analysis

ABSTRACT

Safety is key to civil aviation. To further improve its already respectable safety records, the airline industry is transitioning towards a proactive approach which anticipates and mitigates risks before incidents occur. This approach requires continuous monitoring and analysis of flight operations; however, modern aircraft systems have become increasingly complex to a degree that traditional analytical methods have reached their limits – the current methods in use can only detect 'hazardous' behaviors on a pre-defined list; they will miss important risks that are unlisted or unknown. This paper presents a novel approach to apply data mining in flight data analysis allowing airline safety experts to identify latent risks from daily operations without specifying what to look for in advance. In this approach, we apply a Gaussian Mixture Model (GMM) based clustering to digital flight data in order to detect flights with unusual data patterns. These flights may indicate an increased level of risks under the assumption that normal flights share common patterns, while anomalies do not. Safety experts can then review these flights in detail to identify risks, if any. Compared with other data-driven methods to monitor flight operations, this approach, referred to as ClusterAD-DataSample, can (1) better establish the norm by automatically recognizing multiple typical patterns of flight operations, and (2) pinpoint which part of a detected flight is abnormal. Evaluation of ClusterAD-DataSample was performed on two sets of A320 flight data of real-world airline operations; results showed that ClusterAD-DataSample was able to detect abnormal flights with elevated risks, which make it a promising tool for airline operators to identify early signs of safety degradation even if the criteria are unknown a priori.

1. Introduction

Historically, accident prevention efforts focused on accidents analysis: a particular event was investigated in detail; measures were developed to prevent the event from reoccurring. These efforts led a steady improvements of aviation safety over the past 60 years (Boeing Commercial Airplanes, 2014). However, safety remains a key element in air transportation, a colossal industry which moves 3.1 billion passengers a year (International Civil Aviation Organization, 2014). Mishaps still

* Corresponding author. Tel.: +852 3442 4726.
E-mail addresses: lishuai.li@cityu.edu.hk (L. Li), rjhans@mit.edu (R.J. Hansman), rafael.palacios@iit.upcomillas.es (R. Palacios), rwelsch@mit.edu (R. Welsch).

http://dx.doi.org/10.1016/j.trc.2016.01.007
0968-090X/

happen, and if they do, often make headline news globally. To further improve an already respectable safety performance, the aviation industry is transitioning towards a proactive approach. Different from previous ones, the proactive approach aims at continuously monitoring flight operations and identifying risks; mitigation measures are taken before incidents occur. Data, especially the use of sensor data, are the key to facilitate this transition.

Among others, data from the Flight Data Recorder (FDR) or Quick Assess Recorder (QAR) onboard every aircraft are the most promising. Thousands of technical parameters are recorded throughout a flight, airspeed, altitude, pitch, roll, engine parameters, etc. The dataset contains rich information about the aircraft system, the external environment, and pilot operations. Major airlines have implemented a Flight Operational Quality Assurance (FOQA) program or Flight Data Monitoring (FDM) program to archive and analyze these flight data from daily operations (Federal Aviation Administration, 2004). Current analysis tools are all based on the Exceedance Detection (ED) (Federal Aviation Administration, 2004), a method which can identify undesired events by checking if particular flight parameters exceed a pre-defined limit under specified conditions. The adjustment of ED is costly because parameter thresholds depend specifically on aircraft types, flight phases, airport conditions, and flying procedures, etc. A key limitation of ED is that one needs to pre-define what to look for in advance; and the pre-defined criteria, which are derived from finite historical incidents, can never predict future risks arising from infinite operational variances and emerging new technologies. A number of studies have discussed the limitations of a pre-defined list (Matthews et al., 2013; Tsuruta, 2009).

Recent studies focused on the development of data-driven methods that monitor operations continuously for road and rail transportation using in-vehicle data recorders or network sensors (Chang et al., 2008; Li et al., 2014; Shi and Abdel-Aty, 2015; Shichrur et al., 2014; Toledo et al., 2008; Zhang et al., 2011). Studies focusing on air transportation are sparse. The Morning Report software package was one of the earliest efforts made to detect anomalies from routine flight data as part of NASA's Aviation Performance Measurement System (APMS) (Amidan and Ferryman, 2005). The software models time series data of selected flight parameters using a quadratic equation. Each flight, mapped as a point, is described by the coefficients of the quadratic equations in the feature space. The distance between this point and the mean of the distribution in the feature space, is used to compute an "atypical score" for each flight. Some studies, the Inductive Monitoring System (IMS) software (Iverson, 2004) for instance, adopt a semi-supervised learning approach that summarizes the data distributions of typical system behaviors from a pre-sanitized training dataset. The typical data distributions are then compared with real-time operational data to detect abnormal behaviors. However, the IMS is limited in its ability to account for temporal patterns and it cannot function without a training dataset. Others adopt the unsupervised approach. The Sequence Miner algorithm focuses on discrete flight parameters to monitor pilot operations, such as cockpit switch flips (Budalakoti et al., 2008, 2006). The algorithm can discover abnormal sequences in the switch operations based on the Longest Common Subsequence (LCS) measures. To incorporate both discrete and continuous flight parameters in FDR data, Srivastava develops a statistical framework that discretizes the continuous flight parameters in pre-processing steps (Srivastava, 2005). Built on this framework, Das et al. develop the Multiple Kernel Anomaly Detection (MKAD) which combines both continuous and discrete parameters via kernel functions and applies one-class Support Vector Machine (SVM) for anomaly detection (Das et al., 2010). MKAD assumes there will always be a single, consistent data pattern for normal operations. This assumption does not hold in real practice. One example is that both Instrument Landing Systems (ILS) approaches and visual approaches are standard operations in landing; yet the procedures of these two approaches are different from each other, so are their data patterns. Further, how to characterize the temporal structure during various flight phases remains unresolved. Matthews et al. summarize the knowledge discovery pipeline for aviation data using these algorithms discussed above (Matthews et al., 2013). A common challenge exists for all the above methods: standards of norm are not easy to define in practice – real-world flight operations are too complex to be assumed to have one standard pattern, or to be represented by a limited set of training data.

A thorough literature review concluded that despite a vast number of techniques developed in cluster analysis and anomaly detection in general (Chandola et al., 2009; Hodge and Austin, 2004; Jain et al., 1999), no existing technique is directly applicable to solve the anomaly detection problem for flight operations.

In this paper, we developed a data mining-based approach (referred to as ClusterAD-DataSample) to support proactive safety management in air transportation. In this approach, we apply a Gaussian Mixture Model (GMM) based clustering to digital flight data in order to detect flights with unusual data patterns. These flights may indicate an increased level of risk under the assumption that normal flights share common patterns. ClusterAD-DataSample is built on another anomaly detection method, ClusterAD-Flight, developed by the authors (Li et al., 2015, 2011). However, ClusterAD-Flight can only detect abnormal flights during take-off or approach as a whole, rather than instantaneous abnormal data samples during a flight. Compared with other data-driven methods to monitor flight operations, ClusterAD-DataSample can (1) better establish the norm by automatically recognizing multiple typical patterns of flight operations, and (2) pinpoint which part of a detected flight is abnormal. With this method, airline safety experts will be better equipped to monitor flight operations by detecting flights with unusual data patterns, and locate abnormal behaviors from everyday operations even if the criteria for anomalies are unknown a prior.

In this paper, "anomaly" and "abnormal flights" means flights with unusual data patterns, which differs from "unsafe flights" and "risky flights." Flights with abnormal data patterns are of interest for detection, but they need to be reviewed by safety experts to determine whether they represent any safety risks.

2. Method

The method presented in this paper is referred to as ClusterAD-DataSample. The name stands for Cluster-Based Anomaly Detection, with the unit of analysis being the Data Sample. The method is based on the assumption that the majority of flights exhibit common patterns under routine operations; a few outliers that deviate from those common patterns are of interest to airline safety management. For example, a spike of irregularities during landing at an airport might be the symptom of a systemic problem with a runway environment, standard operating procedures, or pilot training programs. Since these common patterns are unknown a priori, the method deploys cluster analysis to identify common patterns in order to establish the norm before anomaly detection. A number of studies have shown that cluster analysis is an effective technique to identify the norms of operations in air transportation systems (Gariel et al., 2011; Jain et al., 1999; Li, 2013; Li et al., 2011).

The workflow of ClusterAD-DataSample is illustrated in Fig. 1, which consists of four steps: the first transforms the digital flight data into a form applicable for cluster analysis; the second is cluster analysis with a Gaussian Mixture Model (GMM), which characterizes typical behaviors of aircraft systems and temporal patterns of flight operations; the third detects anomalies based on the norm established in the previous step; and the forth is to have safety experts review these anomalies to identify any safety risks at stake.

2.1. Digital flight data

The input of ClusterAD-DataSample is the digital flight data of a fleet type archived by airlines. The flight data is recorded by an aircraft Digital Flight Data Recorder (DFDR) or Quick Access Recorder (QAR) and it includes thousands of flight parameters throughout a flight, such as airspeed, altitude, pitch, roll, and engine parameters.

2.2. Step 1: Transformation

The first step of ClusterAD-DataSample is to convert the original flight data, in the form of multivariate time-series, into vectors, a form that is applicable for cluster analysis, as illustrated in Fig. 2. Each flight is re-sampled at fixed intervals along a standard reference (e.g. time after applying takeoff power during the takeoff phase, distance to touchdown during the approach phase), to ensure data of different flights have equal length. Each sample of flight data from flight f at time t is represented by a vector, as in the form,

$$\mathbf{x}_t^f = [x_t^1, x_t^2, \ldots, x_t^m], \tag{1}$$

where x_t^m is the value of the mth flight parameter at time t.

Because different flight parameters have different ranges and units, we then normalize each flight parameter to have "zero mean and unit variance." As a result, each data sample of the original flight data corresponds to a normalized vector $\hat{\mathbf{x}}_t^f$. Now, the data is in a form that is applicable for cluster analysis.

2.3. Step 2: Cluster analysis

After the transformation step, we develop a clustering method based on a GMM to identify the norm of flight operations in order to detect the abnormal ones. The norm is characterized by clusters and the temporal distribution of these clusters.

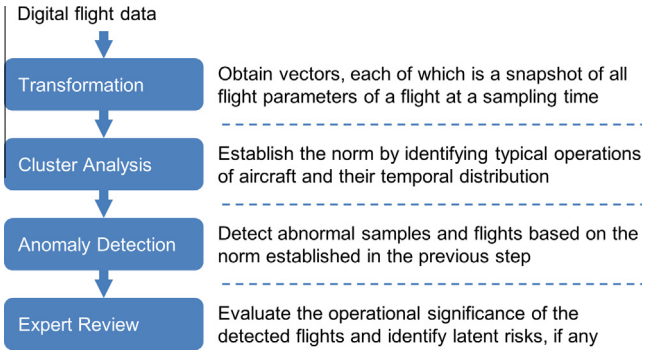

Fig. 1. ClusterAD-DataSample workflow.

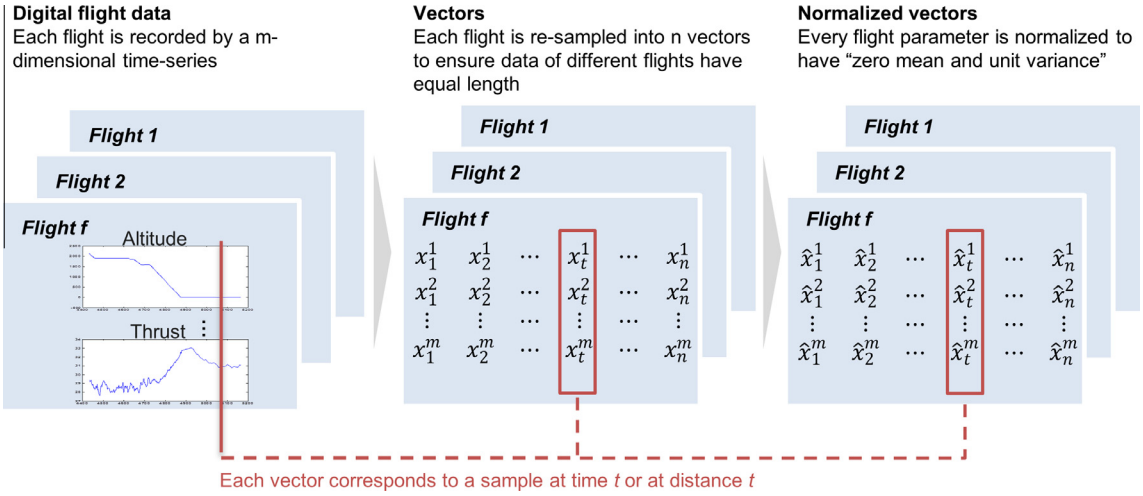

Fig. 2. Transformation step: flight data samples are converted into vectors for cluster analysis.

2.3.1. Identification of frequently observed operations

We use cluster analysis to identify frequently observed operations. In the hyperspace, the algorithm will automatically organize similar vectors into clusters. Each cluster represents a type of frequently observed operations of the aircraft system. These clusters will be used as reference to detect anomalies in the next step. A Gaussian Mixture Model (GMM) was the technique selected for cluster analysis. A GMM is a parametric probability density function represented as a weighted sum of Gaussian component densities (Bouman et al., 1997; Dempster et al., 1977; McLachlan and Basford, 1988; Reynolds, 2008). Each Gaussian component captures a type of frequently observed operations. Compared to other clustering techniques such as *K*-means, a GMM is able to give statistical inferences of the underlying distributions, which can be used to calculate the degree of abnormality of data samples in later steps. A GMM with *K* components is given by:

$$p(\mathbf{x}|\lambda) = \sum_{i=1}^{K} \omega_i g(\mathbf{x}|\mu_i, \Sigma_i) \tag{2}$$

where $\mathbf{x}$ is a collection of *M*-dimensional vectors, $\lambda = \{w_i, \mu_i, \Sigma_i\}$ are the GMM parameters, *K* is the number of Gaussians and $w_i,\ i = 1, \ldots, K$ are the mixture weights, which satisfy the constraint that $\sum_{i=1}^{K} \omega_i = 1$, μ_i is the mean vector and Σ_i is the covariance matrix of a Gaussian, and $g(\mathbf{x}|\mu_i, \Sigma_i)$ are the component Gaussian densities. Each component density is a *M*-variate Gaussian function with mean vector μ_i and covariance matrix Σ_i. It summarizes the probability density of vectors belonging to a cluster.

$$g(\mathbf{x}|\mu_i, \Sigma_i) = \frac{1}{\sqrt{(2\pi)^M |\Sigma_i|}} e^{-\frac{1}{2}(\mathbf{x}-\mu_i)' \Sigma_i^{-1} (\mathbf{x}-\mu_i)} \tag{3}$$

In order to build a GMM model, it is necessary to define attributes of the model before using the expectation–maximization (EM) algorithm to estimate the GMM parameters, including whether the covariance matrices are full or diagonal, whether the parameters among Gaussian components are shared or not, and how many Gaussian components are there in this model. These are often determined by the characteristics of an application problem, e.g. the amount of data available, the computational resources or time requirement, and the natural underlying patterns of the data.

In this particular application, it was decided to use diagonal covariance matrices, independent parameters among Gaussian components, and estimate the number of mixture components (*K*) by sensitivity analysis. The reasons are discussed below.

Covariance matrices are restricted to be diagonal in order to reduce computational complexity. Since the component Gaussians are acting together to model the overall vector, full covariance matrices are not necessary even when the flight parameters are not statistically independent. The linear combination of diagonal covariance Gaussians is capable of modeling the correlations between flight parameters. Reynolds (2008) shows that the performance attained by using a set of full covariance Gaussians can be equally obtained by using a larger set of diagonal covariance Gaussians.

The parameters among Gaussian components are not shared to maximize the goodness of fit. We expect that different Gaussian components would have different distributions to better capture the variety of aircraft operations.

The number of mixture components (*K*) is estimated by sensitivity analysis. Several values of *K* are tested, and the optimal one is chosen based on Bayesian Information Criterion (BIC) (Schwarz, 1978). BIC is a measure of fitness of a statistical model that has been widely used for model identification in time series and linear regression (Abraham and Box, 1979). BIC rewards

model accuracy and penalizes model complexity; therefore BIC selects the minimum value of K that attains adequate accuracy. An example of the process is illustrated in Fig. 6 and described in Section 3.1.2.

After a GMM configuration is defined, the parameters of the GMM ($\lambda = \{\omega_i, \mu_i, \Sigma_i\},\ i = 1, \ldots, K$) are obtained using the expectation–maximization (EM) algorithm (Dempster et al., 1977). It is the most popular and well-established method. The process starts with an initial model, then a new model is created, the process is repeated until improvements are no longer significant. We use 1×10^{-6} as the termination tolerance (ε) for the objective function value. The pseudo code is represented here (see Table 1).

2.3.2. Characterization of cluster temporal distribution

To assess whether a data sample is abnormal or not, we first need to know if it belongs to a type of frequently observed operations, namely a cluster, or not; if yes, we also need to know whether the cluster is appropriate or not considering the temporal sequence of clusters during normal operations.

Fig. 3 shows the process of estimating the distribution of clusters as a function of "distance to touchdown" during the approach phase. The subfigure on the right in Fig. 3 illustrates the observation frequency of clusters changes during approach phase, which is described by circle size and color density. A larger and darker circle indicates a higher frequency. As shown in Fig. 3, frequently observed clusters evolve from Cluster No. 1–10 to No. 30–35 as aircraft are getting closer to touchdown. It is also noted that Cluster 33 is the most frequently observed cluster at the time of touchdown in this dataset.

Table 1
Pseudo code of cluster analysis of flight data based on GMM.

Input:
Normalized vectors of digital flight data
The number of Gaussian components, K
Output:
GMM parameters, $\lambda = \{\omega_i, \mu_i, \Sigma_i\}$, $i = 1, 2, \ldots, K$, that best fit the input data
Algorithm:
1. Initialize GMM parameters λ using K-means result
2. E step. For each data sample, determine the *a posteriori* probability for each Gaussian component i using Eq. (4)

$$\Pr(i|\mathbf{x}_t, \lambda) = \frac{\omega_i g(\mathbf{x}_t|\mu_i, \Sigma_i)}{\sum_{j=1}^{K} \omega_j g(\mathbf{x}_t|\mu_j, \Sigma_j)} \tag{4}$$

where x_t is a normalized vector of digital flight data
3. M step. Update GMM parameters λ. For each Gaussian component, update parameters using Eqs. (5)–(7)

$$\omega_i^{new} = \frac{1}{N}\sum_{t=1}^{N} \Pr(i|\mathbf{x}_t, \lambda) \tag{5}$$

$$\mu_i^{new} = \frac{\sum_{t=1}^{N} \Pr(i|\mathbf{x}_t, \lambda)\mathbf{x}_t}{\sum_{t=1}^{N} \Pr(i|\mathbf{x}_t, \lambda)} \tag{6}$$

$$\sum_i^{new} = \frac{1}{\sum_{t=1}^{N} \Pr(i|\mathbf{x}_t, \lambda)} \sum_{t=1}^{N} \Pr(i|\mathbf{x}_t, \lambda)(\mathbf{x}_t - \mu_i^{new})(\mathbf{x}_t - \mu_i^{new})^T \tag{7}$$

where N is the total number of data samples in the dataset
4. Evaluate log likelihood

$$\ln(p(\mathbf{x}|\lambda)) = \sum_{t=1}^{N} \ln\left(\sum_{i=1}^{K} \omega_i g(\mathbf{x}_t|\mu_i, \Sigma_i)\right) \tag{8}$$

If likelihood converge (the difference between $\ln(p(x|\lambda^{new}))$ and $\ln(p(x|\lambda))$ is smaller than the termination tolerance, ε),
Stop;
Else,
$\lambda = \lambda^{new}$
go to Step 2.
5. Output λ

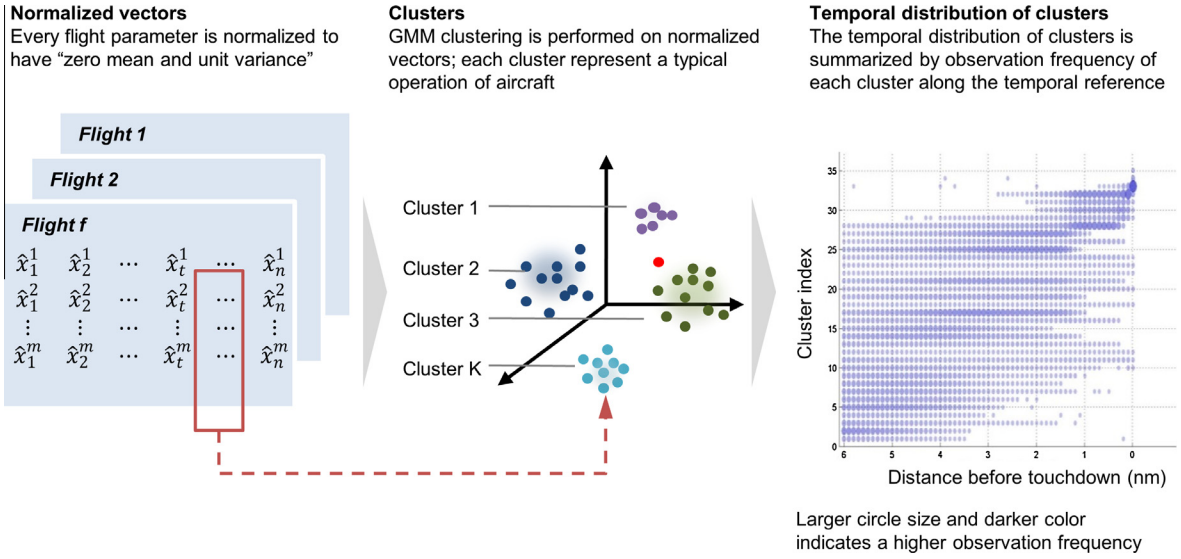

Fig. 3. Cluster analysis: identify typical operations and temporal distribution.

In this step, to calculate the relative appropriateness of Cluster q at time t, we use the ratio between the *a posteriori* probability for Gaussian component q and the sum of *a posteriori* probability for all Gaussian components given vectors sampled at time t from all flights. The formula is described in Eq. (11).

2.4. Step 3: Anomaly detection

The final procedure for detecting anomalies is to compute the probability density function (pdf) of each sample being normal, which is determined by (1) how likely it is to belong to one of the clusters, (2) the cluster is appropriate at that moment during a phase of flight. Mathematically, it is a sum of the pdf of a sample belonging to a cluster, weighted by the pdf of the cluster being appropriate, as described in Eq. (4),

$$p(\mathbf{x}_t^f \text{ is normal}) = \sum_{q=1}^{K} p(\mathbf{x}_t^f \text{ is from cluster } q) \cdot p(\text{cluster } q \text{ is appropriate at time } t) \tag{9}$$

where $\mathbf{x}_t^f$ is a M-dimensional vector at time t from flight f, q represents a cluster, F is the total number of flights in the dataset, and

$$p(\mathbf{x}_t^f \text{ is from cluster } q) = g(\mathbf{x}_t^f | \mu_q, \Sigma_q) \tag{10}$$

$$p(\text{cluster } q \text{ is appropriate at time } t) = \frac{\sum_{f=1}^{F} Pr(q|\mathbf{x}_t^f, \lambda)}{\sum_{i=1}^{K}\sum_{f=1}^{F} Pr(i|\mathbf{x}_t^f, \lambda)} \tag{11}$$

A probability density profile for every flight can be constructed after calculating the pdf of every data sample. Fig. 4 shows an example. In this graph, a high value indicates a data sample is normal (relative to other samples in the dataset); whereas lower values mean relatively abnormal. A logarithmic scale is used because the original data covers a very large range of

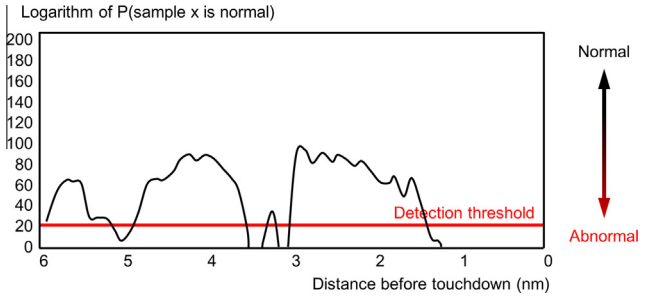

Fig. 4. Probability density profile of being "normal" for a flight during the approach phase.

values. Abnormal samples are detected by identifying samples with $p(\mathbf{x}_t^f$ is normal$)$ that is lower than a threshold. This threshold value is associated with the sensitivity of the detection system, and can be adjusted based on the distribution of all samples in a dataset. For example, Fig. 5 shows a distribution of $p(\mathbf{x}_t^f$ is normal$)$ for all samples in a dataset comprising 10,528 flights (Airbus 320) during the approach phase. If we choose the 1st percentile value in this distribution to be the detection threshold, the top 1% anomalies will be detected. Similar to the detection of abnormal samples, abnormal flights are detected based on the sum of $p(\mathbf{x}_t^f$ is normal$)$ over all samples during a given flight phase, and the detection threshold is set based on the pdf distribution of all flights in a dataset.

2.5. Step 4: Expert review

The last step is to let airline safety experts review the abnormal flights detected in the previous step in order to determine whether they indicate safety hazards or not. Safety experts review the original QAR data and refer to other sources of information, e.g. weather reports, procedure standards, to determine what is abnormal about a flight and whether the abnormality reveals any risks or not.

A practical issue of the expert review process is that the time required to review a flight is significant, since hundreds of flight parameters need to be examined manually. This practical issue is common among existing data-driven methods to monitor flight operations (Das et al., 2010; Li et al., 2015; Matthews et al., 2013). The advantage of ClusterAD-DataSample is that the probability profile generated in the previous step (see Fig. 4) can be used to shorten the time – which part of a flight is abnormal can be quickly located by looking at the graph, as shown in Figs. 9 and 10.

3. Evaluation studies

3.1. Testing ClusterAD-DataSample on real airline data

In order to show the implementation of the proposed method on a real-world data set, ClusterAD-DataSample was tested on a set of QAR data provided by an international airline.

3.1.1. Dataset

The data set contains 10,528 flights of Airbus 320s with the same engine configuration. Flight origins or destinations spread across 36 airports. Each flight's recording has 142 flight parameters (113 continuous and 29 discrete). To compare flights at different airports, position related flight parameters were first converted to values relative to the airport location. For instance, recorded altitude values such as "pressure altitude" and "density altitude" were transformed into relative altitudes like "height above takeoff" or "height above touchdown".

3.1.2. GMM of nominal operations

A GMM was built following the steps described in Section 2. Regarding the selection of K (number of mixture components), we found 35 to be the optimal value for this dataset as it gave the lowest BIC value, as shown in Fig. 6. Then, the parameters of the 35 Gaussian mixture components ($\lambda = \{w_i, \mu_i, \Sigma_i\},\ i = 1, \ldots, 35$) are obtained using the expectation–max imization (EM) algorithm.

After reviewing the parameter values, ten of the 35 mixture components were recognized as flight operations that are well known during approach phase (see Fig. 7). For example, flight parameters in Cluster 17 have values at listed in Table 2, thus this GMM component is identified as the flight operation "ILS (Instrument Landing System) approach". Other well-known flight operations identified include "Flare", "Touchdown", and "Thrust reverser deployment", etc. Fig. 7 shows all

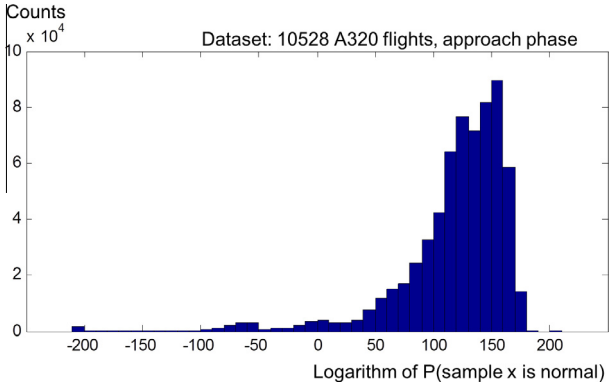

Fig. 5. Distribution of probability density functions of being "normal" for all samples of all flights.

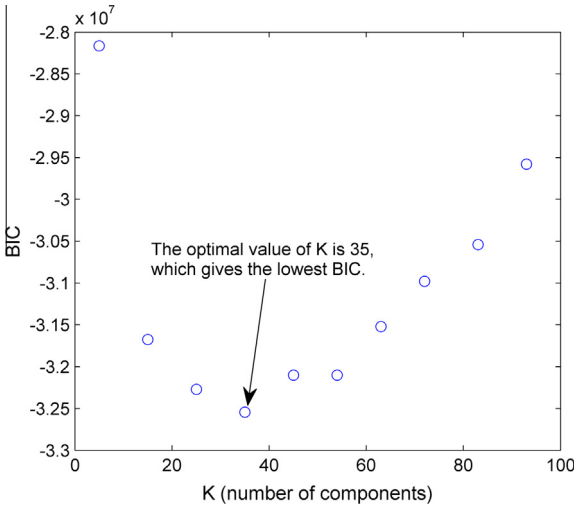

Fig. 6. Sensitivity analysis of number of components based on BIC.

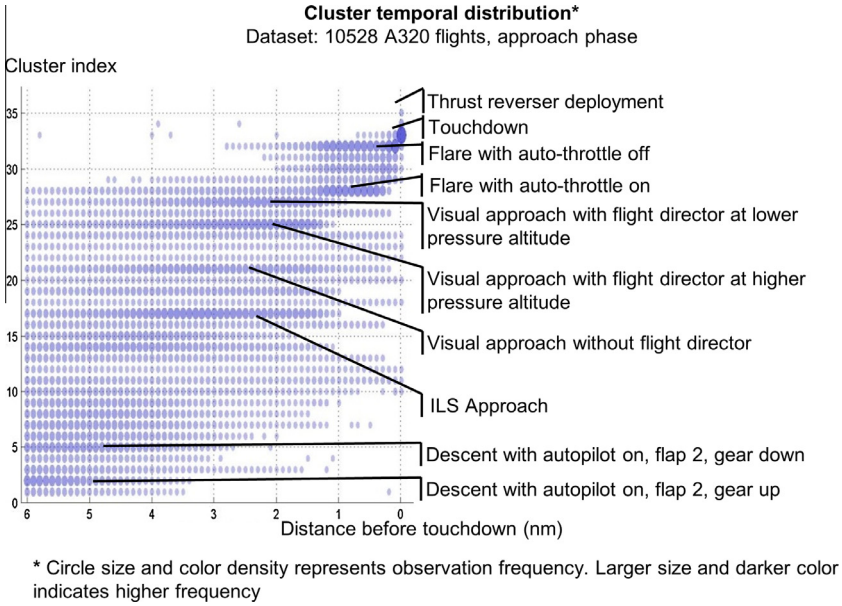

Fig. 7. Evolution of frequently observed clusters during the approach phase.

35 clusters during approach phase and how they activate and deactivate as a function of the distance to touchdown. Many flights started the approach phase from Cluster 2: "Descent with flap 2 and gear up" or Cluster 5: "Descent with flap 2 and gear down". Between 4 nm and 2 nm before touchdown, the most frequently observed operations were captured by Cluster 17: "ILS approach." The following well-known operations include Cluster 21: "Visual approach without flight director", Cluster 25: "Visual approach with flight director at lower pressure altitude" and Cluster 27: "Visual approach with flight director at higher pressure altitude". At the end of the approach phase, Cluster 33: "Touchdown" has significantly more observations than other clusters, as indicated by the large dark blue dot at 0 nm before touchdown.

3.1.3. Abnormal flights and expert review

Four airline pilots were recruited through the Air Line Pilots Association (ALPA) and professional pilot forums to review the abnormal flights detected by ClusterAD-DataSample. The four pilots have more than 48,600 h of flying experience combined, and they are either in charge of a FOQA program or have past experience with FOQA program.

The 0.5, 1, and 3 percentile values of the probability distribution were used as the detection thresholds to detect abnormal flights. Respectively, 53, 106, and 316 flights were detected as abnormal. The detection threshold can be adjusted according

Table 2
Flight parameters related to ILS approach in Cluster 17.

Parameter	Value
CAT 3 capability	True
Glideslope hold mode	On
Localizer hold mode	On
Approach type	ILS
Autopilot	Engaged

to users' preference. In practice, the detection thresholds will be determined with a good balance between review workloads and the likelihood of missed detections. Ideally, all these flights would be reviewed by domain experts. However, limited by the number of domain experts and time available in this study, an initial analysis was performed to reduce the number of flights that needed to be reviewed by each domain expert. Since there were a finite number of data patterns observed in these flights, we selected a flight to represent each pattern. After the initial analysis, twelve of the 53 flights detected by 0.5% threshold were selected for domain experts to review.

All flights under review were confirmed to exhibit some level of abnormal behaviors. However, whether the abnormal behaviors indicated safety hazards or not varies by flight and by expert. The perceived level of safety risks depended on experts' operational experience and personal opinion. The results are summarized in Table 3.

In this section, we describe two examples. The first example (Flight ID 1) shows that ClusterAD-DataSample is able to detect an unstabilized approach, which was rated as operationally abnormal and representing safety hazards by all four domain experts. Information of key flight parameters of this flight is shown in Fig. 8. This flight is commented as an unstabilized approach with high speed, late flap configuration and idle thrust until very short final by safety experts. Its behavior was away from "normal" in terms of expected stabilized approach criteria. One expert stated, "As a pilot and also as an operator I would consider this approach as requiring a go-around."

The second example (Flight ID 12) shows that ClusterAD-DataSample is capable of detecting instantaneous abnormal samples. Fig. 9 represents the probability profile of the abnormal flight (only the last 6 nm are represented to focus on the problematic area). There is an area corresponding to a distance between 2.5 and 2.8 nm in which the pdf of being normal is extremely low. Further inspection found that parameters related to the left engine had abnormal values, as shown in Fig. 10, and that was probably the root cause of the abnormality. In a situation like this, it may happen that neither parameter deviation was significant enough to trigger an alarm in traditional methods, so the small malfunction of engine or sensor may go unnoticed until further damage occurred. ClusterAD-DataSample is able to combine different sources of information so it can be more sensitive to such incipient failures.

3.2. Comparing with other methods for detecting known issues in flight data

In order to further assess ClusterAD-DataSample's performance, we compared ClusterAD-DataSample with the current industry standard method, Exceedance Detection (ED), and two other data-driven methods, MKAD and ClusterAD-Flight, given there is no "ground truth" for performance evaluation, or a benchmark dataset in which every flight is labeled 'normal or abnormal" in definitive terms. **ED** is a standard method widely used by airlines in FOQA programs. Therefore in this comparison study it was used as a baseline. It consists on checking if some flight parameters exceed predefined limits under certain conditions. If the limits are overpassed, the event is labeled in three alarm levels: Level 1 indicates minor deviation, Level 2 indicates moderate, and Level 3 indicates severe deviation from expected values. **MKAD** stands for Multiple Kernel Anomaly Detection. It is a method developed by Das et al. (2010) based on one-class Support Vector Machine (SVM). It is able

Table 3
Expert review results of abnormal flights detected by ClusterAD-DataSample.

Flight ID	Abnormal behaviors	Number of experts agree that the flight indicated safety hazards
1	Unstabilized approach, high energy approach	4
2	Unstabilized approach if IMC, large roll and idle thrust between 5 and 2.5 nm	1
3	Unstabilized approach until 2.5 nm, visual approach without flight director	2
4	ILS approach with manual takeover, slightly fast until 1.5 nm, one pneumatic system off	1
5	ILS approach from low intercept altitude, initially strong crosswind	0
6	ILS approach, autopilot and flight director recycle at 3 nm	0
7	Spikes in "brake pressure"	0
8	Instability in roll	0
9	Strong tailwind and late localizer intercept approach	2
10	Visual approach in cold weather	0
11	Strong, quartering tail wind, abnormal shift in center of gravity at 2 nm	1
12	Abnormal values in left engine	0

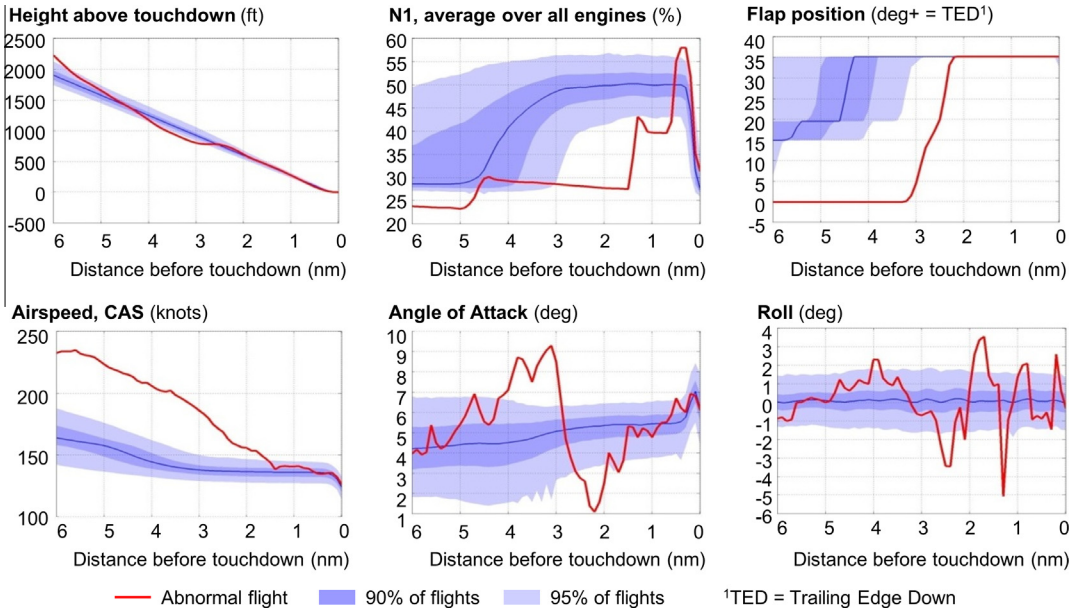

Fig. 8. Unstabilized approach detected by ClusterAD-DataSample.

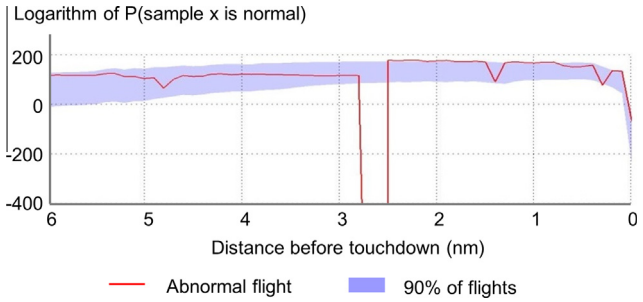

Fig. 9. Probability density profile of an abnormal flight: abnormal behaviors are detected by ClusterAD-DataSample between 2.5 and 2.8 nm before touchdown.

to combine information of different data types and to detect anomalies by analyzing various parameters simultaneously. **ClusterAD-Flight** is an anomaly detection method developed by Li et al. (2015, 2011)). This method uses clustering techniques to detect abnormal flights during phases of flight that have standard procedures associated and clear time anchors. It is effective in detecting anomalies during take-off and final approach.

The scope of this comparative study focused on testing the capability of data-driven methods (MKAD, ClusterAD-Flight, and ClusterAD-DataSample) in detecting known risky issues that have been specified in ED. The performance of detecting unknown issues (the ones that could not be detected by ED) was not evaluated in this study due to the lack of domain experts and a systematic approach in reviewing a large number of abnormal flights.

3.2.1. Dataset

The dataset contains 25,519 A320 flights landing at a typical European airport provided by another international airline. Average flight length is between 2 and 3 h. For each flight, 367 discrete and continuous parameters were recorded at a sampling frequency of 1 Hz.

3.2.2. Algorithm settings

All four detection algorithms can be set at different sensitivity levels, which may make results incomparable. In this study, three data-driven methods (ClusterAD-DataSample, MKAD, and ClusterAD-Flight) were compared against Level 3 ED Events using a series of detection thresholds: 1%, 3%, 5%, and 10%. The percentage controls the number of flights to be considered abnormal ranked by severity in each method.

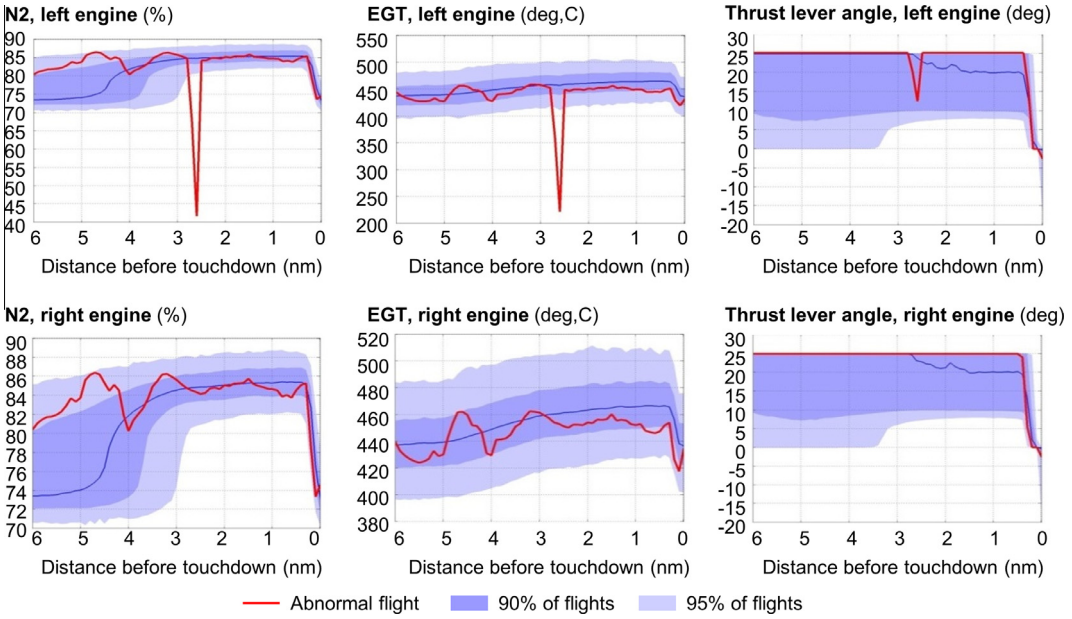

Fig. 10. Abnormal values in left engine detected by ClusterAD-DataSample.

3.2.3. Results

Fig. 11 shows the percentages of flights detected by MKAD, ClusterAD-Flight and ClusterAD-DataSample, which were also detected with ED Level 3. ClusterAD-DataSample detected more flights with ED Level 3 than ClusterAD-Flight and MKAD across all detection thresholds. For example, when the detection threshold was set to detect the top 10% abnormal flights, 70% of flights with ED Level 3 were identified by ClusterAD-DataSample, in contrast with only 12% identified by MKAD and 30% by ClusterAD-Flight.

3.3. Discussion

The initial testing of ClusterAD-DataSample on real-world airline data showed that the method can detect flights with abnormal operations; some may exhibit early signs of safety degradation. The comparative study also shows that the new method is able to detect unsafe events that have been defined in the industry standard method. Thus, ClusterAD-DataSample is a promising tool to help airline safety experts monitor everyday flights and identify the abnormal ones.

In these two evaluation studies, the detection threshold is artificially chosen for comparison between different algorithms. However, there are several methods to determine the threshold value in ClusterAD-DataSample. One option is to use the traditional outlier detection method based on boxplot (Tukey, 1977). We can consider values lower than Q1-1.5IQR (Inter Quartile Range) to be abnormal. However, the pdf distribution is always left skewed. Too many abnormal flights might be detected for airline analysts to review using this method. Another option is to set the threshold by the users (e.g. an airline operator) according to their preference and needs. They can adjust the detection thresholds to balance between review workloads and the likelihood of miss-detection.

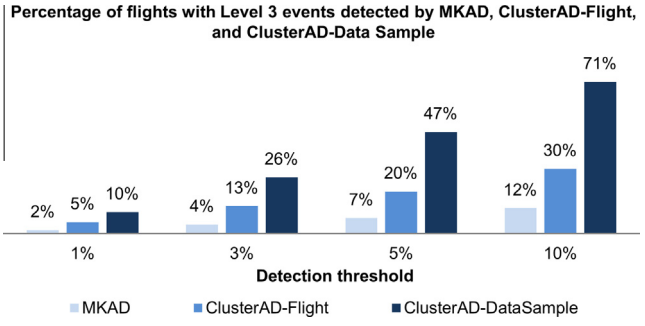

Fig. 11. Percentage of flights with Level 3 Exceedance Events Detected by MKAD, ClusterAD-Flight, and ClusterAD-DataSample.

Another limitation of this paper is that we did not provide an evaluation of the detection accuracy in a traditional sense, such as false negatives and false positives. The reason is there is a lack of "ground truth", or a benchmark dataset in which every flight is labeled 'normal or abnormal" in definitive terms. Incidents or accidents are extremely rare in modern airline operations, none of them can be found in available flight data sets. Individual flights are highly varied from one another and it is challenging to label it 'right or wrong'. New techniques are needed to evaluate the performance of a detection method that does not rely on pre-existing criteria but is capable of detecting unknown issues.

4. Conclusions

Digital flight data are collected by airlines from all aircraft on a regular basis. These data contain a large amount of information about daily operations that could be used to inform airlines for safety improvement. Yet the analysis of such data is challenging due to the increased complexity and variability in air transportation operations. We developed a new data-driven approach that can support safety experts to utilize digital flight data, better monitor flight operations and potentially improve airline safety. The new approach can automatically detect anomalous situations without extended initial tuning which are time consuming and expensive. Results of abnormal flight can inform further analysis by domain experts to identify risks, and to determine whether mitigation measures are needed to prevent accidents. The method was tested on real-world datasets provided by international airlines. Results show that ClusterAD-DataSample is able to detect anomalies that are operationally significant and may represent increased level of risks. Compared with other data-driven methods to detect anomalies in flight data, ClusterAD-DataSample performed better in detecting known unsafe events. Further study is needed to comprehensively evaluate its performance in detecting unknown issues.

Although ClusterAD-DataSample is specifically designed to monitor flight operations, the cluster-based method can potentially be adapted for operations and systems monitoring in other domains. A next step is to adopt similar approaches for drivers' behaviors analysis, urban traffic monitoring, aircraft and vehicle maintenance. A data transformation method that captures meaningful system behaviors for that domain is the key.

Acknowledgments

The authors would like to thank Ashok Srivastava and Santanu Das for their help on the comparative study of different anomaly detection methods. The authors would also like to thank Alan H. Midkiff and many other airline safety experts and pilots for their invaluable domain expertise and insightful discussions. This work was supported by the Federal Aviation Administration under the Joint University Project (JUP) FAA 11-G-016, the National Aeronautics and Space Administration (NASA) under Grant # NNA06CN23A, and City University of Hong Kong Start-up Grant # 7200418.

References

Abraham, B., Box, G.E.P., 1979. Bayesian analysis of some outlier problems in time series. Biometrika 66, 229–237.

Amidan, B.G., Ferryman, T.a., 2005. Atypical event and typical pattern detection within complex systems. In: 2005 IEEE Aerospace Conference. IEEE, Big Sky, MT, pp. 3620–3631. http://dx.doi.org/10.1109/AERO.2005.1559667.

Boeing Commercial Airplanes, 2014. Statistical Summary of Commercial Jet Airplane Accidents. Boeing Commercial Airplanes.

Bouman, C.A., Shapiro, M., Cook, G.W., Atkins, C.B., Cheng, H.G., Dy, J., Borman, S., 1997. CLUSTER: An Unsupervised Algorithm for Modeling Gaussian Mixtures.

Budalakoti, S., Srivastava, A.A.N., Otey, M.E.M., 2008. Anomaly detection and diagnosis algorithms for discrete symbol sequences with applications to airline safety. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. 39, 101–113. http://dx.doi.org/10.1109/TSMCC.2008.2007248.

Budalakoti, S., Srivastava, A.N., Akella, R., 2006. Discovering atypical flights in sequences of discrete flight parameters. In: Aerospace Conference, 2006 IEEE. IEEE, Big Sky, MT, pp. 1–8. http://dx.doi.org/10.1109/AERO.2006.1656109.

Chandola, V., Banerjee, A., Kumar, V., 2009. Outlier detection: a survey. ACM Comput. Surv., 1–72

Chang, T.H., Hsu, C.S., Wang, C., Yang, L.K., 2008. Onboard measurement and warning module for irregular vehicle behavior. IEEE Trans. Intell. Transp. Syst. 9, 501–513. http://dx.doi.org/10.1109/TITS.2008.928243.

Das, S., Matthews, B.L., Srivastava, A.N., Oza, N.C., 2010. Multiple kernel learning for heterogeneous anomaly detection: algorithm and aviation safety case study. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, Washington, DC, pp. 47–56. http://dx.doi.org/10.1145/1835804.1835813.

Dempster, A., Laird, N., Rubin, D., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. 39, 1–38.

Federal Aviation Administration, 2004. Advisory Circular. 120–82 Flight Operational Quality Assurance. Fed. Aviat. Adm. http://dx.doi.org/10.1177/004728757301200242.

Gariel, M., Srivastava, A.N., Feron, E., 2011. Trajectory clustering and an application to airspace monitoring. IEEE Trans. Intell. Transp. Syst. 12, 1511–1524. http://dx.doi.org/10.1109/TITS.2011.2160628.

Hodge, V., Austin, J., 2004. A survey of outlier detection methodologies. Artif. Intell. Rev. 22, 85–126. http://dx.doi.org/10.1023/B:AIRE.0000045502.10941.a9.

International Civil Aviation Organization, 2014. ICAO Safety Report 2014 Edition. International Civil Aviation Organization, Montreal, Canada.

Iverson, D.L., 2004. Inductive system health monitoring. In: Proceedings of the 2004 International Conference on Artificial Intelligence (IC-AI04). Las Vegas, NV.

Jain, A.K., Murty, M.N., Flynn, P.J., 1999. Data clustering: a review. ACM Comput. Surv. 31, 264–323. http://dx.doi.org/10.1145/331499.331504.

Li, H., Parikh, D., He, Q., Qian, B., Li, Z., Fang, D., Hampapur, A., 2014. Improving rail network velocity: a machine learning approach to predictive maintenance. Transp. Res. Part C Emerg. Technol. 45, 17–26. http://dx.doi.org/10.1016/j.trc.2014.04.013.

Li, L., 2013. Anomaly Detection in Airline Routine Operations Using Flight Data Recorder Data. Massachusetts Institute of Technology.

Li, L., Das, S., Hansman, R. John, Palacios, R., Srivastava, A.N., 2015. Analysis of flight data using clustering techniques for detecting abnormal operations. J. Aerosp. Inf. Syst. 1–12. http://dx.doi.org/10.2514/1.I010329.

Li, L., Gariel, M., Hansman, R.J., Palacios, R., 2011. Anomaly detection in onboard-recorded flight data using cluster analysis. In: Digital Avionics Systems Conference (DASC), 2011 IEEE/AIAA 30th. Seattle, WA. http://dx.doi.org/10.1109/DASC.2011.6096068.
Matthews, B.L., Das, S., Bhaduri, K., Das, K., Martin, R., Oza, N., 2013. Discovering anomalous aviation safety events using scalable data mining algorithms. J. Aerosp. Inf. Syst. 10, 467–475. http://dx.doi.org/10.2514/1.I010080.
McLachlan, G.J., Basford, K.E., 1988. Mixture Models: Inference and Applications to Clustering. Marcel Dekker, New York.
Reynolds, D.A., 2008. Gaussian Mixture Models, Encycl. Biometric Recognition. Springer, pp. 659–663. http://dx.doi.org/10.1007/978-0-387-73003-5_196.
Schwarz, G.E., 1978. Estimating the dimension of a model. Ann. Stat. 6, 461–464.
Shi, Q., Abdel-Aty, M., 2015. Big data applications in real-time traffic operation and safety monitoring and improvement on urban expressways. Transp. Res. Part C Emerg. Technol. 1, 1–15. http://dx.doi.org/10.1016/j.trc.2015.02.022.
Shichrur, R., Sarid, A., Ratzon, N.Z., 2014. Determining the sampling time frame for In-Vehicle Data Recorder measurement in assessing drivers. Transp. Res. Part C Emerg. Technol. 42, 99–106. http://dx.doi.org/10.1016/j.trc.2014.02.017.
Srivastava, A.N., 2005. Discovering system health anomalies using data mining techniques. In: Proceedings of the 2005 Joint Army Navy NASA Airforce Conference on Propulsion. Charleston, SC, pp. 1–11.
Toledo, T., Musicant, O., Lotan, T., 2008. In-vehicle data recorders for monitoring and feedback on drivers' behavior. Transp. Res. Part C Emerg. Technol. 16, 320–331. http://dx.doi.org/10.1016/j.trc.2008.01.001.
Tsuruta, G., 2009. The Analysis of Flight Operational Quality Assurance (FOQA) Data: Exploration of a Proposed List of Improved Safety Parameters. VDM Verlag Dr, Muller.
Tukey, J.W., 1977. Exploratory Data Analysis. Pearson.
Zhang, J., Wang, F.-Y., Wang, K., Lin, W.-H., Xu, X., Chen, C., 2011. Data-driven intelligent transportation systems: a survey. IEEE Trans. Intell. Transp. Syst. 12, 1624–1639. http://dx.doi.org/10.1109/TITS.2011.2158001.