



## ANOMALY DETECTION IN AIRLINE ROUTINE OPERATIONS USING FLIGHT DATA RECORDER DATA

Lishuai Li and R. John Hansman

*This report is based on the Doctoral Dissertation of Lishuai Li submitted to the Department of Aeronautics and Astronautics in partial fulfillment of the requirements for the degree of Doctor of Philosophy at the Massachusetts Institute of Technology.*

*The work presented in this report was also conducted in collaboration with the members of the Doctoral Committee:*

*Prof. R. John Hansman (Chair)  
Prof. Roy Welsch  
Prof. Rafael Palacios  
Prof. Julie Shah*

Report No. ICAT-2013-4  
June 2013

MIT International Center for Air Transportation (ICAT)  
Department of Aeronautics & Astronautics  
Massachusetts Institute of Technology  
Cambridge, MA 02139 USA

*[Page Intentionally Left Blank]*

# **ANOMALY DETECTION IN AIRLINE ROUTINE OPERATIONS USING FLIGHT DATA RECORDER DATA**

by

Lishuai Li and R. John Hansman

## **Abstract**

In order to improve safety in current air carrier operations, there is a growing emphasis on proactive safety management systems. These systems identify and mitigate risks before accidents occur. This thesis develops a new anomaly detection approach using routine operational data to support proactive safety management. The research applies cluster analysis to detect abnormal flights based on Flight Data Recorder (FDR) data. Results from cluster analysis are provided to domain experts to verify operational significance of such anomalies and associated safety hazards. Compared with existing methods, the cluster-based approach is capable of identifying new types of anomalies that were previously unaccounted for. It can help airlines detect early signs of performance deviation, identify safety degradation, deploy predictive maintenance, and train staff accordingly.

The first part of the detection approach employs data-mining algorithms to identify flights of interest from FDR data. These data are transformed into a high-dimensional space for cluster analysis, where normal patterns are identified in clusters while anomalies are detected as outliers. Two cluster-based anomaly detection algorithms were developed to explore different transformation techniques: ClusterAD-Flight and ClusterAD-Data Sample.

The second part of the detection approach is domain expert review. The review process is to determine whether detected anomalies are operationally significant and whether they represent safety risks. Several data visualization tools were developed to support the review process which can be otherwise labor-intensive: the Flight Parameter Plots can present raw FDR data in informative graphics; The Flight Abnormality Visualization can help domain experts quickly locate the source of such anomalies.

A number of evaluation studies were conducted using airline FDR data. ClusterAD-Flight and ClusterAD-Data Sample were compared with Exceedance Detection, the current method in use by airlines, and MKAD, another anomaly detection algorithm developed at NASA, using a dataset of 25519 A320 flights. An evaluation of the entire detection approach was conducted with domain experts using a dataset of 10,528 A320 flights. Results showed that both cluster-based detection algorithms were able to identify operationally significant anomalies that beyond the capacities of current methods. Also, domain experts confirmed that the data visualization tools were effective in supporting the review process.

## Acknowledgments

The work was supported by the Federal Aviation Administration under the Joint University Project (JUP) FAA 11-G-016 and the National Aeronautics and Space Administration (NASA) under Grant # NNA06CN23A.

# Contents

Chapter 1	Introduction .....	7
1.1	Current Aviation Safety Level and Proactive Safety Management .....	7
1.2	Current Use of Flight Data Recorder Data .....	9
1.2.1	Flight Data Recorder .....	9
1.2.2	Flight Operational Quality Assurance (FOQA) Program.....	10
1.3	Research Objectives.....	13
1.4	Research Overview .....	13
Chapter 2	Literature Review .....	15
2.1	Flight Data Monitoring.....	15
2.1.1	Flight Data Monitoring Process .....	15
2.1.2	Flight Data Analysis Tools .....	17
2.2	Anomaly Detection.....	20
2.2.1	Anomaly Detection Categorization.....	20
2.2.2	Anomaly Detection in This Research.....	22
2.3	General Anomaly Detection Techniques.....	23
2.3.1	Statistical Anomaly Detection Approaches.....	23
2.3.2	Classification-based Anomaly Detection Approaches .....	24
2.3.3	Cluster-based Anomaly Detection Approaches .....	25
2.4	Anomaly Detection Techniques for Time Series .....	28
2.4.1	Data-based Approaches.....	29
2.4.2	Model-based Approaches.....	30
2.5	Anomaly Detection in Aviation .....	31
Chapter 3	Cluster-based Anomaly Detection Algorithms.....	34
3.1	Challenges in Detecting Anomalies in FDR Data .....	34
3.2	Concept of Cluster-based Anomaly Detection Algorithms.....	35
3.3	Pattern-based Anomalies and Instantaneous Anomalies.....	37
3.4	ClusterAD-Flight .....	38
3.4.1	Data Transformation.....	39
3.4.2	Dimension Reduction .....	40
3.4.3	Cluster Analysis.....	43
3.5	Initial Testing of ClusterAD-Flight .....	44
3.5.1	Dataset and Data Preparation.....	45
3.5.2	Results Overview .....	46
3.5.3	Abnormal Behaviors in Flights Detected .....	48
3.5.4	Nominal Data Patterns from Clusters.....	54
3.6	ClusterAD-Data Sample .....	57
3.6.1	Identification of Nominal Modes.....	58

3.6.2	Characterization of Mode Distribution.....	61
3.6.3	Detecting Anomalies Based on Nominal Modes .....	63
3.7	Initial Testing of ClusterAD-Data Sample.....	65
3.7.1	Dataset and Data Preparation.....	65
3.7.2	Optimal Number of Nominal Modes .....	66
3.7.3	Nominal Modes.....	67
3.7.4	Anomaly Detection.....	69
3.8	Summary.....	73
Chapter 4	Expert Review and Data Representation.....	74
4.1	Expert Review Process and Practical Challenges .....	74
4.2	Data Visualization Tools for Expert Review .....	76
4.2.1	Flight Parameter Plot .....	77
4.2.2	Flight Abnormality Visualization .....	82
Chapter 5	Evaluation Studies.....	85
5.1	Overview of Evaluation Studies.....	85
5.1.1	Evaluation Challenges .....	85
5.1.2	Overview of Evaluation Studies.....	85
5.1.3	Overview of FDR Datasets .....	88
5.2	Evaluation Study I: Comparison of ClusterAD-Flight, MKAD and Exceedance Detection ..	89
5.2.1	Background: Evaluation Challenges in Anomaly Detection.....	89
5.2.2	Comparison Design and Setup.....	90
5.2.3	Results Overview .....	93
5.2.4	Comparison between ClusterAD-Flight and MKAD.....	95
5.2.5	Comparison with Exceedance Detection .....	107
5.2.6	Study Summary.....	110
5.3	Evaluation Study II: Comparison of ClusterAD-Flight, ClusterAD-Data Sample, and MKAD in Detecting Exceedances.....	110
5.3.1	Comparison Design and Setup.....	110
5.3.2	Results.....	111
5.3.3	Study Summary.....	112
5.4	Evaluation Study III: Evaluation of ClusterAD algorithms and Data Visualization Tools with Domain Experts.....	112
5.4.1	Evaluation Design .....	113
5.4.2	Apparatus, Participants, and Procedure .....	119
5.4.3	Results on Perceived Operational Significance of Abnormal Flights .....	121
5.4.4	Results on Data Visualization Tools .....	135
5.4.5	Study Summary.....	137
Chapter 6	Conclusions and Future Work.....	139
6.1	Conclusions .....	139
6.2	Recommendations for Future Work .....	141
Reference.....		142

# Chapter 1

## Introduction

### 1.1 Current Aviation Safety Level and Proactive Safety Management

Aviation safety has been improving steadily over the past 50 years. The annual accident rate and fatal accident rate decreased significantly as shown in Figure 1.1 and Figure 1.2 (Boeing Commercial Airplanes, 2012.) In recent years, the accident rate has been stable. As an industry which receives extensive public attention, the pressure on safety is always present.

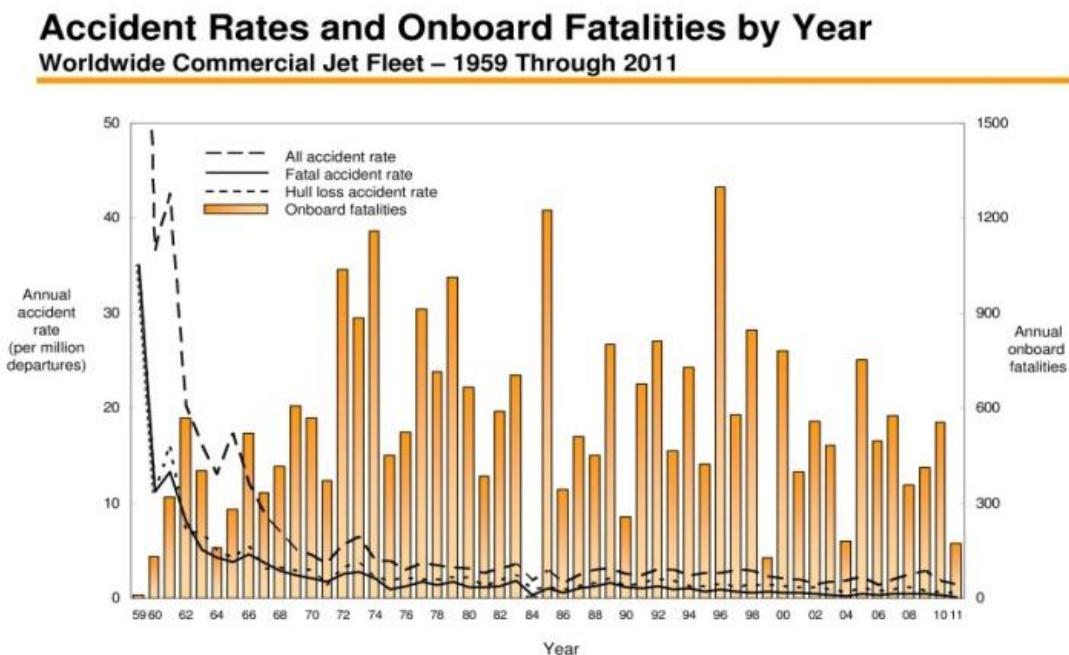


Figure 1.1 Accident Rates and Onboard Fatalities by Year (Boeing Commercial Airplanes, 2012)

## U.S. and Canadian Operators Accident Rates by Year

**Fatal Accidents – Worldwide Commercial Jet Fleet – 1959 Through 2011**

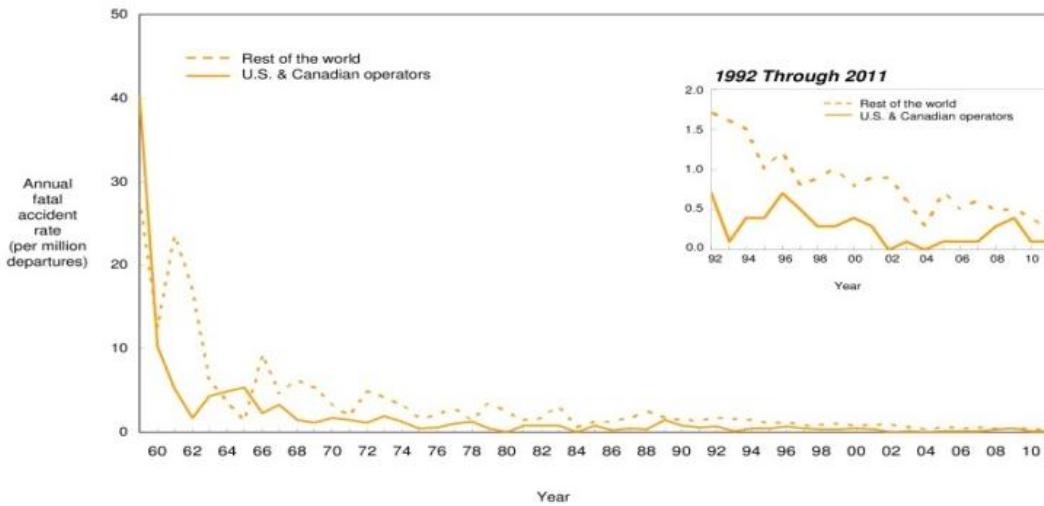


Figure 1.2 U.S. and Canadian Operators Accident Rates by Year (Boeing Commercial Airplanes, 2012)

In the past, airline safety has been improved by identifying problems after accidents. Accidents trigger the development and implementation of mitigation strategies (Logan, 2008). Further improvement of safety requires a proactive approach, in which potential hazards are identified and corrective actions are taken before accidents even occur.

In recent years, the airline industry has been making several efforts towards a more proactive safety management system. The Federal Aviation Administration (FAA) participants in three major voluntary programs (Federal Aviation Administration, 2010a)

- Aviation Safety Action Program (ASAP) - A joint FAA/industry program that allows aviation employees to self-report safety violations to air carriers and FAA free from legal or disciplinary consequences.
- Flight Operational Quality Assurance (FOQA) - A program for the routine collection and analysis of digital flight data generated during aircraft operations.
- Advanced Qualification Program (AQP) - A voluntary alternative to traditional pilot training regulations that replaces programmed hours with proficiency-based training. AQP incorporates data-driven processes enabling air carriers to refine training based on identified individual needs.

Along with the implementation of voluntary safety programs, the airline industry has also developed a systems approach to manage risks via a systematic, explicit, and comprehensive process. The 2006 Advisory Circular (AC120-92) (Federal Aviation Administration, 2006), titled Introduction to Safety Management Systems for Air Operators, introduced the concept of a Safety Management System (SMS) to aviation service providers. The 2010 Advisory Circular (AC120-92A) (Federal Aviation Administration, 2010b) provided a Framework for Safety Management Systems development by aviation service providers.

The implementation of the voluntary safety programs has made progress. Meanwhile, SMS plans have been developed further along with other safety initiatives (Federal Aviation Administration, 2013). All these efforts identify detailed information on hazards, processes, and precursor events, in order to support risk assessment and mitigation actions. They converge in one direction: proactive safety management.

## 1.2 Current Use of Flight Data Recorder Data

### 1.2.1 Flight Data Recorder

Every aircraft is equipped with a Flight Data Recorder (FDR). Historically, it was used for accident investigations only. A FDR records flight parameters onboard during an entire flight. Typical flight parameters include altitude, airspeed, accelerations, thrust, engine pressures and temperatures, control surfaces etc. An example of the FDR data is shown in Figure 1.3. It displays a few flight parameters recorded in the last 110 seconds of the Cogan Air Flight 3407 before impact (NTSB, 2009).

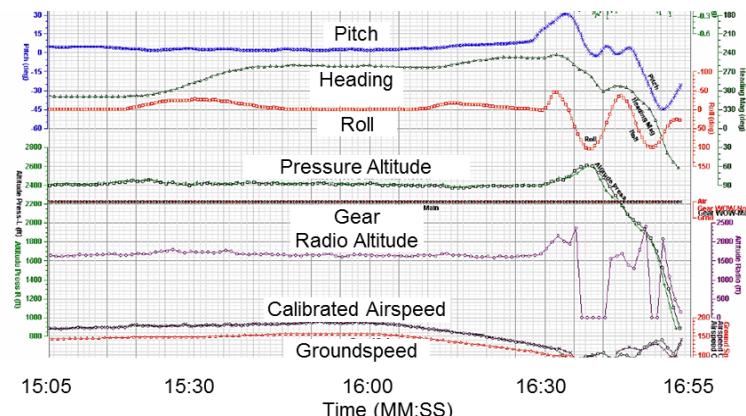


Figure 1.3 Example of Flight Data Recorder Data (NTSB, 2009)

Today, the recording capacity of FDR is significantly increased. The number of flight parameters being recorded increased significantly as shown in Table 1.1. The specification of flight parameters recorded and sampling rates vary by the type of recorder and the configuration requirements of the airline. FAA defined the minimum requirements of what flight parameters to be recorded and minimum sampling rates (Federal Aviation Administration, 2011) for US airlines. However, the recording capability on modern airplanes is much larger than the minimum requirements, as it was shown in Table 1.1. The FDR on Boeing 787 can record approximately 2000 flight parameters for 50 hours (Rosenkrans, 2008). The sampling rates vary by the nature of the flight parameter. For example, the vertical acceleration is recorded at 8 Hz, while the outside air temperature is recorded at 0.5 Hz.

Table 1.1 Evolution of FDR (N. A. H. Campbell, 2007)

Aircraft	Time into Service	FDR Type	Number of parameters	FDR data capacity
<b>Boeing 707</b>	1958	Analogue	5	Mechanical limit of ~10 parameter
<b>Airbus 330</b>	1993	Digital	280	128 words per second (wps)
<b>Embraer 170</b>	2004	Digital Combi-recorder <sup>1</sup>	774	256 wps
<b>Airbus 380</b>	2007	Digital	> 1000	1024 wps
<b>Boeing 787</b> (first flight)	2009	Digital EAFR <sup>2</sup>	> 1000	Ethernet system

As FDR data is digitalized, it is also easier to access the data. FDR data can be downloaded periodically during routine operations. Therefore, it is possible to utilize FDR data from routine operations before accidents happen.

## 1.2.2 Flight Operational Quality Assurance (FOQA) Program

### Program Overview

The Flight Operations Quality Assurance (FOQA) program, also called the Flight Data Monitoring (FDM) in Europe, aims to use detailed flight data recorded during daily flights to improve airline operations and safety. The general process of a FOQA program consists of data recording, downloading, and analysis. The recording of the data happens during every flight,

---

<sup>1</sup> The combi-recorder stores both cockpit audio and flight data

<sup>2</sup> EA FR: Enhanced Airborne Flight Recorder

measuring various flight parameters by airplane sensors and recording onboard using FDR. Data are downloaded and stored in a large database at the airline through manual retrieval or wireless transmission. For manual retrieval, the downloading process is performed once in several days or weeks when the aircraft park at a station or maintenance base. For wireless transmission, the downloading can be completed within 15 minutes after aircraft landing (Teledyne Controls, 2011). Downloaded data are analyzed by FOQA analysts to evaluate daily operations.

## Data Analysis

Currently, analysis of FDR data is conducted by using a number of special purpose software programs. Details of each program may vary, yet data analysis is using two primary approaches: the Exceedance Detection approach and the Statistical Analysis approach (Federal Aviation Administration, 2004).

**Exceedance Detection.** The exceedance detection approach detects pre-defined operationally undesired events. It monitors if particular flight parameters exceed the predefined limits under certain conditions. The list of flight parameters and the limits of those parameters need to be specified by safety specialists in advance. Usually, the watch list coincide with the airline's standard operating procedures, such as the pitch at takeoff, the speed at takeoff climb, the time of flap retraction, etc. A few examples are shown in Table 1.2 (Federal Aviation Administration, 2004).

Table 1.2 Examples of Exceedance Event Parameters and Definitions (Federal Aviation Administration, 2004)

Event	Parameters	Basic Event Definition	Event Description
<b>Pitch High at Takeoff</b>	Air/Ground Switch, Pitch	Air/Ground = Ground, Pitch > x degrees	An event that measures pitch at takeoff in relation to the angle required to strike the tail of the aircraft
<b>Takeoff Climb Speed High</b>	CAS, Gross Weight, HAT	HAT > x feet, HAA < x feet, CAS > V2 + x knots	An event to detect climb speed higher than desired during the Takeoff Phase of flight
<b>Early Flap Retraction</b>	HAT, Flap Position	HAT < x feet, Flap Position < Flap Position in the preceding sample	An event to detect any flap movement from the takeoff position prior to reaching the altitude at which flap retraction should begin

<b>Approach Speed High</b>	Gross Weight, CAS, HAT, Flaps	HAT > 1000 feet, HAT < 3000 feet, CAS > $V_{FE} - x$ knots HAT < 1000 feet, CAS > $V_{REF} + x$ knots	An event to detect operation on approach that is in excess of its computed final approach speed
<b>Excessive Power Increase</b>	HAT, $N_1$	HAT < 500 feet, $\Delta$ of $N_1 > x$	An event to detect an excessive power increase during final phase of approach
<b>Operation Below Glideslope</b>	Glide Slope Deviation Low, HAT	Glide Slope Deviation Low > $x$ dots, HAT < $x$ feet	An event to detect deviation below glideslope

**Statistical Analysis.** In the statistical analysis approach, distributions of certain flight parameters are plotted to examine a particular flight operation. Compared with the Exceedance Detection, a carrier can gain a more complete picture of the operation based on the distribution of all flights using the Statistical Analysis. Figure 1.4 provides an example of the distribution analysis on “altitude at which landing flap is set”.

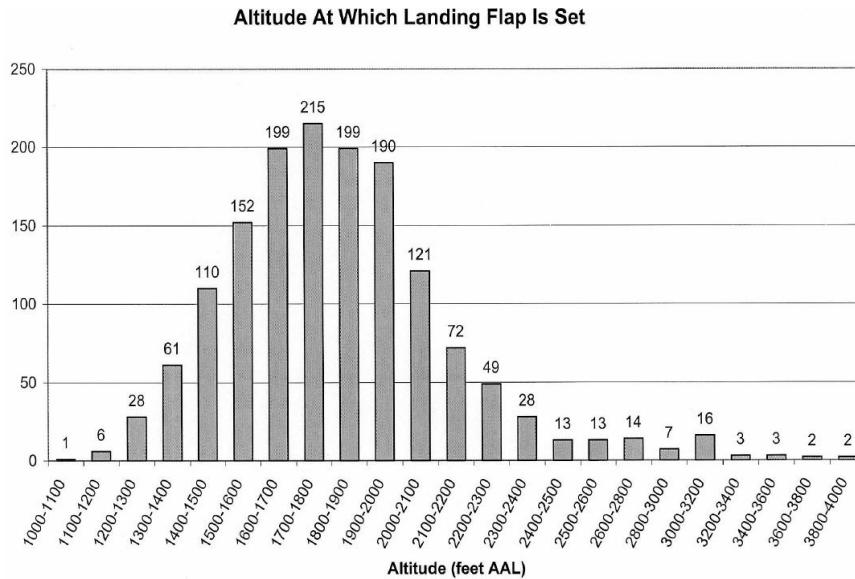


Figure 1.4 Distribution Analysis of ”Altitude at Which Landing Flap Is Set” (N. Campbell, 2003)

The current FOQA data analysis performs well on known issues. However, it is incapable in identifying the unknowns. Both approaches need a pre-defined watch list of key parameters under certain operational conditions; the Exceedance Detection requires the thresholds of key parameters to be precisely defined in advance. As a result, only known issues are examined; emerging risks will remain undetected.

## 1.3 Research Objectives

This research develops an innovative approach to identify emerging risks from routine airline operations using FDR data, which combines the strength of data mining and the knowledge of domain experts: data mining techniques are applied to detect abnormal flights from FDR data, and then abnormal flights are referred to domain experts for in-depth analysis to identify emerging risks.

The objectives of this research are:

- 1) Develop algorithms to detect abnormal flights from FDR data without pre-specification of parameter limits.

The objective of the algorithms is to identify abnormal flights from FDR data without pre-specification of detection scope. The detection is independent of prior knowledge of safe operations or specification of risks. The algorithms will detect flights based on data patterns, rather than testing a limited set of hypotheses, possibly exposing unknown issues.

- 2) Develop data visualizations to support expert reviewing abnormal flights.

The flights detected by data mining method are not always operationally significant. They need to be referred to domain experts for an in-depth review. Data visualization tools are developed to help experts interpret results and quickly identify what is abnormal about flights detected by the algorithms.

## 1.4 Research Overview

In order to identify unknown issues from routine flights' FDR data, this thesis proposed a new approach that uses data mining techniques to process FDR data efficiently, and it relies on domain experts to interpret the results and operational implications. Cluster analysis is used to detect flights of potential interest from FDR data, and these flights are referred to domain experts for in-depth analysis via data visualization tools developed in this thesis. An overview of the approach developed in this thesis is illustrated in Figure 1.5.

In the first step of the approach, cluster-based methods are used to detect abnormal flights from a large set of data. Since operations of commercial airplanes are highly standardized, a

majority of these flights demonstrate similar patterns in their FDR data. Assuming the majority are safe, flights with uncommon data patterns indicate abnormal situations, which might indicate increased level of risk. Therefore, flights with abnormal data patterns are labeled as potentially interesting flights for further inspection. Two cluster-based anomaly detection (ClusterAD) algorithms were developed: ClusterAD-Flight and ClusterAD-Data Sample.

The second step of the approach relies on domain experts to review the flights detected by anomaly detection algorithms. The algorithms are able to detect flights with abnormal data patterns, but these flights are not necessarily abnormal by operational standards. The system relies on domain experts to review flight details to draw conclusions based on their operational experience and knowledge. Challenges of presenting raw FDR data to domain experts are identified during the research. In response, this thesis developed data visualization tools to support the review process of domain experts.

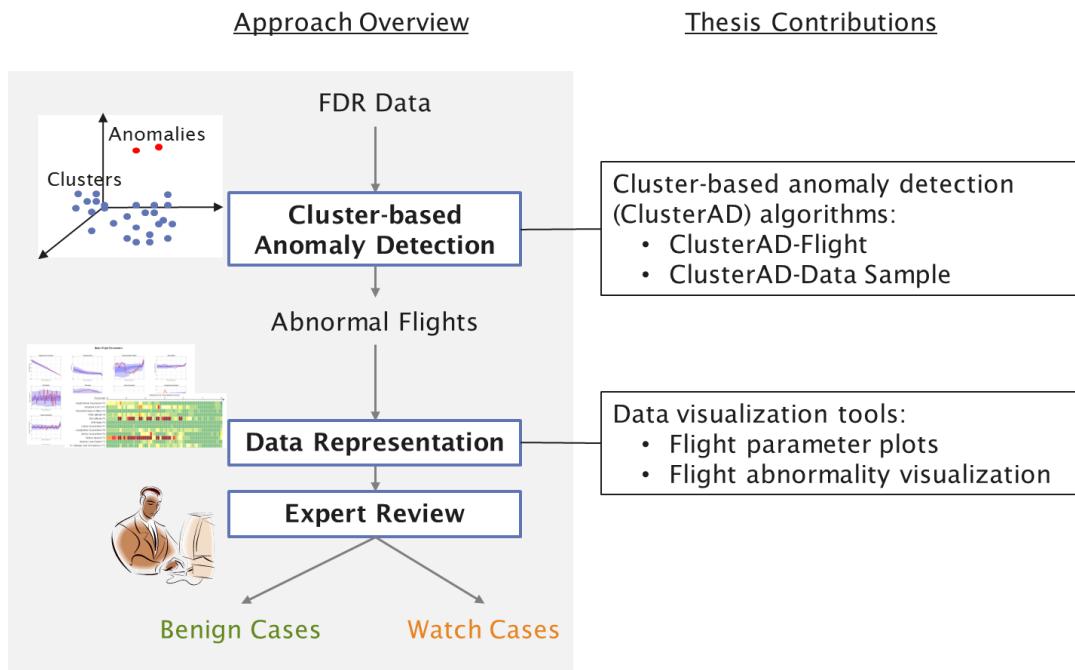


Figure 1.5 Research Overview

# Chapter 2

## Literature Review

### 2.1 Flight Data Monitoring

Many airlines collect and analyze flight data of routine flights. The process is generally referred as flight data monitoring, which involves data acquisition, transmission, storage and analysis, which are described in detail in this section. By reviewing a number of software tools for flight data analysis, a benchmark of current flight data analysis methods was established. Improvement opportunities were identified from the literature review, which motivated this research.

#### 2.1.1 Flight Data Monitoring Process

Major steps in the current process of flight data monitoring include data acquisition, transmission, and analysis, as illustrated in Figure 2.1. The equipment and technology to support each step is described below.

**Data acquisition and recording.** Flight data are measured by hundreds to thousands of sensors on an aircraft during an entire flight. They are collected and converted into a single stream of digital format by the Flight Data Interface Management Unit (FDIMU), also named as Flight Data Acquisition Unit (FDAU), Flight Data and Management Systems (FDAMS), or other variations depending on the manufacturer and the technology used. The converted single stream of data is then sent to the Flight Data Recorder (FDR) and the Quick Access Recorder (QAR) if equipped. The FDR, known as the black box, is a crash-survivable unit that stores flight data mainly for accident investigations. The QAR is similar to FDR, but have no crash survivability case. It is normally located in the avionics bay beneath the flight deck for easier access, and records data in tape, diskette, magneto-optical, or PCMCIA media for convenient data download.

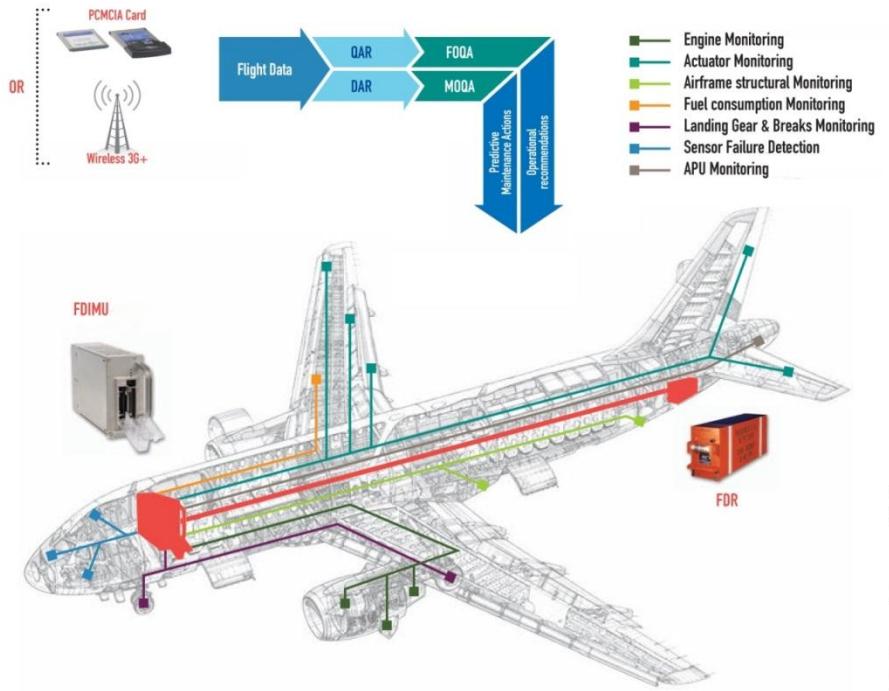


Figure 2.1 Flight Data Monitoring Process (Sagem, 2013)

**Data transfer.** Data stored in QAR or Digital FDR are periodically downloaded to ground stations by electronic transmission, wireless transmission, or manual data retrieval. The download period coincides with the recording memory capability of the media, or meets the operator's need for timely analysis of flight data. Using wireless transmission can reduce data delivery delays, from several days or weeks when using manual data retrieval, to 10-15 minutes after an aircraft is landed (Teledyne Controls, 2011).

**Data analysis.** Once the data are downloaded from QAR or Digital FDR, they are forwarded to the Ground Data Replay and Analysis Station (GDRAS), which transforms raw digital flight data into engineering values for processing and analysis. It also consists of flight data analysis, statistical reporting, flight animation, and other analytic tools based on the configuration, which is set by objective and scope of flight data monitoring programs at the operator, such as the FOQA or Maintenance Operational Quality Assurance (MOQA) program. Software tools for data analysis are discussed in detail in Section 2.1.2. After the analysis, flight data are archived for further analysis in the future.

## 2.1.2 Flight Data Analysis Tools

Data analysis is the core of flight data monitoring. A number of commercial software tools are available to perform various types of analysis on flight data. They are often addressed as Ground Data Replay and Analysis Station (GDRAS) software, FOQA software, or MOQA software, most of which have one or more of the following functions – Analysis, Animation, and Reporting

**Analysis.** All current commercial software packages analyze flight data based on a pre-defined library of undesired events. The techniques used include exceedance detection, statistical analysis, and trend analysis. **Exceedance Detection** identifies pre-defined undesired operational events by checking if particular flight parameters exceed the predefined limits under certain conditions. The list of flight parameters to watch and the limits of those parameters need to be specified by safety specialists in advance. The watch list is usually chosen to coincide with the airline's standard operating procedures, such as the pitch at takeoff, the speed at takeoff climb, the time of flap retraction, etc. **Statistical Analysis** creates profiles of flight, maintenance, or engineering operational data. Airlines can gain a more complete picture of its operations from those profiles, than individual exceedance events. In **Trend Analysis**, airlines periodically aggregate and analyze exceedance events over time—for example, the number of unstabilized approaches at a particular airport per month, over the last 12 months. This type of analysis provides valuable information to the airline, especially in terms of whether the airline's performance is improving, holding steady, or deteriorating.

**Animation.** Exceedance events detected by software are selected and validated by specialists to confirm if they are unsafe operations, or if they are benign cases. Integrated 3-D flight animation is used to support the validation process. With libraries of cockpit panels, instruments, 3D aircraft models, airport runways and terrain information, flight animations can be created from the flight data, which help specialists to review what happened during the flight.



Figure 2.2 Flight Animation (Aerobytes Ltd., n.d.)

**Reporting.** Results from exceedance detection, statistical analysis, and trend analysis, are reported in various formats by GDRAS software. Exceedances are normally aggregated before reporting, yet individual exceedances can be examined as shown in Figure 2.3. Reports of trend analysis include aggregated exceedance information over time (Figure 2.4). Figure 2.5 shows an example of statistical analysis on airspeed deviation at 500 feet across fleet type.

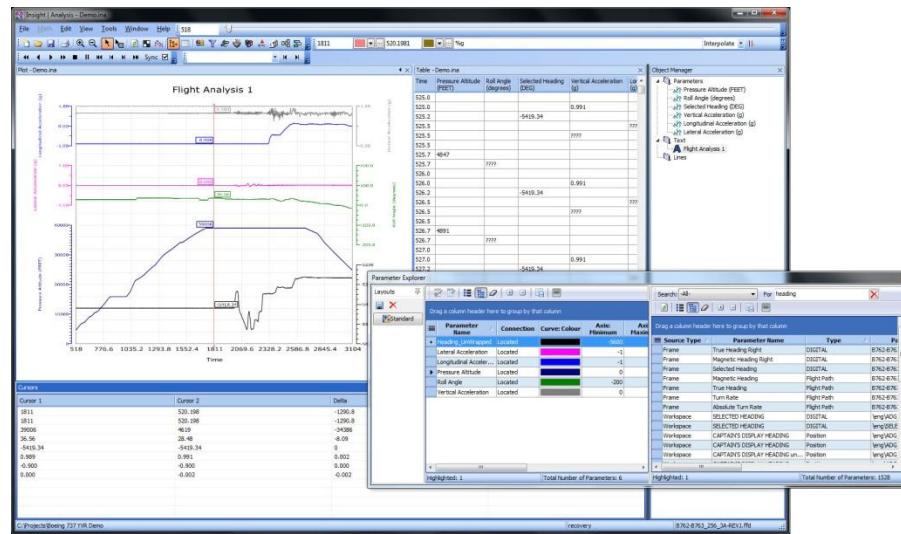


Figure 2.3 Flight Parameters for Exceedance Detection (CAE Flightscape, n.d.-a)



Figure 2.4 Report: Events per 100 Flights (CAE Flightscape, n.d.-b)

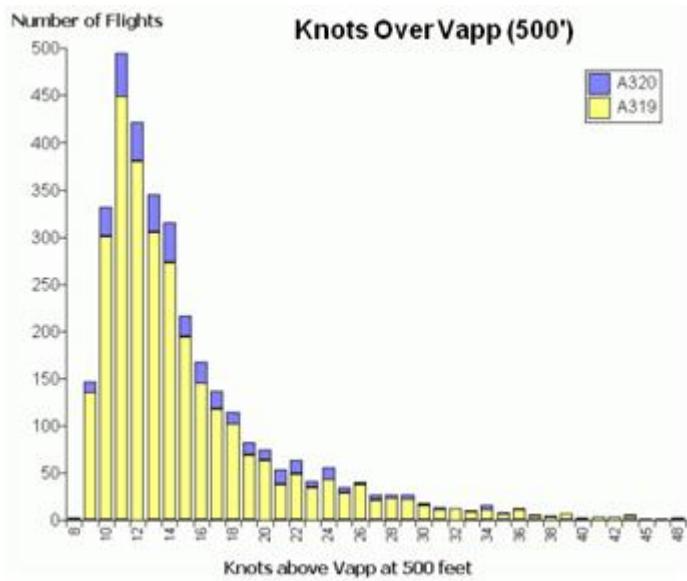


Figure 2.5 Statistical Reports of Fleet Activity (Sagem, n.d.)

In summary, current technology and equipment have enabled flight data monitoring, analysis, and anomaly detection. Flight data can be collected during flight, downloaded to ground, and analyzed to inform airlines about the safety and efficiency of their operations. Software tools currently available focus on event-based analysis, which requires events-to-watch to be defined in advance, with specific parameter criteria. Information on those pre-defined events can be evaluated through exceedance detection, statistical analysis, and trend analysis.

A limitation of the event-based data analysis approach is that it only detects known issues. One needs to specify what to look for in advance. However, accidents often happen because of unknown risks. The goal of this research is to develop a method to detect unknown issues without specifying what events to detect in advance.

## 2.2 Anomaly Detection

This thesis proposes a new approach that detects abnormal flights from routine airline operations using FDR data and asks domain experts to interpret the results and operational implications. Thus, anomaly detection algorithms will be developed to detect anomalies from FDR data.

Anomaly detection refers to the problem of detecting an observation (or patterns of observations) that is inconsistent with the majority members of the dataset. It is also referred to as novelty detection, anomaly detection, fault detection, deviation detection, or exception mining in different application domains. A significant number of anomaly detection techniques have been developed. While some of the techniques are generic and can be applied to different application problems, many of them are focused on solving particular types of problems in an application domain.

### 2.2.1 Anomaly Detection Categorization

Anomaly detection varies significantly depending on the requirements and constraints of the application. The different characteristics of anomaly detection applications bring the need for developing so many diverse techniques, as how to solve the problem largely depends on how the problem is formulated.

The anomaly detection problems are classified along three aspects as summarized in (Chandola, Banerjee, & Kumar, 2009): type of input data, type of supervision, and type of anomalies (Table 2.1.)

Table 2.1 Anomaly Detection Categorization

Problem Aspect	Categories
<b>Input data</b>	Binary, categorical, continuous; Univariate, multivariate; Point data, data with structures (time series, sequences, spatial data);
<b>Type of supervision</b>	Supervised, semi-supervised, unsupervised
<b>Anomalies</b>	context-based outliers, instantaneous anomalies, pattern-based outliers, correlation-based outliers

Input data are the data in which anomalies are detected. The input data can be univariate or multivariate depending on the number of variables. Each variable can be binary, categorical or continuous. Moreover, the input data can also be categorized based on whether the observations are independent or not. Independent observations are referred as point data (Chandola et al., 2009). Dependent observations have temporal structure, spatial structure, or other type of structures among them. Examples are time series, sequences and spatial data. Different types of input data require different techniques to extract relevant information from raw input data.

Depending on whether training data is needed or not, anomaly detection algorithms can be grouped into supervised techniques, semi-supervised techniques, and unsupervised techniques. Supervised anomaly detection techniques build a model based on training data sets which have both nominal and outlier classes. Semi-supervised techniques only need one labeled class in the dataset, such as nominal instances (or outlier instances). Most of the semi-supervised techniques construct a representative model for the nominal behavior based on the nominal instances and detect outliers if any test instances do not fit the model. Unsupervised anomaly detection techniques do not require labeled training data, but detect outliers assumes parametric distributions of difference classes of the data, or assumes that frequent occurrences are nominal while rare occurrences are outliers.

The exact notion of an anomaly is different for different application domains. Depending on the specific definition of anomalies, detection techniques are different. The most basic type of anomaly is a data sample with values that are inconsistent with the normal<sup>3</sup> samples in the entire dataset. For multivariate problems, correlation based anomalies are desired in some problems,

---

<sup>3</sup> “Normal” refers to the opposite of “abnormal”; it does not refer to having the property of a Gaussian distribution in this thesis.

since the relations among variables might be more important than individual variable values. In the case of input data with structures, it is often more meaningful to evaluate values in a specific context. A data sample is considered as an anomaly or not depending on when and where it occurs. Thus, the anomalies of interest might be sudden changes in time series, location specific deviations in spatial data, etc. Moreover, one might be interested in the patterns of temporal/spatial changes in the data, rather than values of individual data samples. Therefore, the pattern-based outliers are detected by examining trends over time/space, ordering of sequences, and frequency of occurrence.

### **2.2.2 Anomaly Detection in This Research**

The objective of the anomaly detection in this thesis is to detect abnormal flights within routine flight data without prior knowledge. The problem has the following characteristics:

The input data are multivariate time series, which are continuous, categorical or binary depending on the flight parameter.

Unsupervised techniques are required by the nature of the main objective of the thesis: detecting abnormal flights without prior knowledge.

The anomalies to detect are context-based and include both pattern-based anomalies and instantaneous anomalies. Context-based means that a data sample is considered as an anomaly or not depending on when it occurs. A modern aircraft is operated under many different system states. Whether a set of values is abnormal or not depends on which specific mode that aircraft is at that time. Within context-based anomalies, two types are of interest to detect: pattern-based anomalies which have abnormal patterns over a period of time, and instantaneous anomalies are instantaneous observations that are abnormal. Techniques used to detect these two types of anomalies are different.

Table 2.2 Characteristics of Anomaly Detection in This Research

<b>Problem Aspect</b>	<b>Characteristics</b>
<b>Input data</b>	Multivariate time series; Binary, categorical or continuous
<b>Type of supervision</b>	Unsupervised
<b>Anomalies</b>	Context-based: pattern-based anomalies and instantaneous anomalies

No existing anomaly detection technique can be directly applied to the anomaly detection problem in this thesis. The following literature review of anomaly detection methods are centered on approaches related to the anomaly detection problem in this thesis.

## 2.3 General Anomaly Detection Techniques

Many anomaly detection techniques have been developed to address anomaly detection problems in many application domains. Three main approaches have been taken: statistical approach, classification approach, and clustering approach. The categories are not mutually exclusive as some of the techniques adopt concepts from more than one basic approach. (Chandola et al., 2009; Hodge & Austin, 2004) provide the most recent and extensive review of the anomaly detection techniques in various domains. Typical techniques and algorithms of each approach are reviewed in this part.

### 2.3.1 Statistical Anomaly Detection Approaches

The earliest approaches used for anomaly detection were based on statistical models (Hodge & Austin, 2004). In these approaches, a statistical distribution of data is assumed or estimated during training phase. Whether a data instance is an outlier or not depends on how well it fits the distribution.

The most popular distribution assumed in this type of work is Gaussian model. The parameters of the Gaussian model are estimated in the training phase using techniques like Maximum Likelihood Estimates. Then statistical tests are used in the testing phase to determine if a given sample belongs to the distribution or not. Common outlier statistical tests include the Box-Plot rule (Laurikkala, Juhola, & Kentala, 2000), the Grubbs test (Anscombe & Guttman, 1960; Grubbs, 1969; Stefansky, 1972), and variants of the Grubb test (Gibbons, Bhaumik, & Aryal, 2009; Rosner, 1983).

Many techniques using a mixture of parametric models were developed for the situations in which a single statistical model is not sufficient to represent the data. If both normal and abnormal training data are available, separate parametric distributions are modeled for normal cases and abnormal cases using Expectation Maximization (EM) (Byers & Raftery, 1998; Eskin, 2000). The test data are evaluated based on which distribution they belong to. If only normal

training data are available, a mixture of models for only normal data is developed. The test data are considered as outliers if they do not belong to any normal models.

Another type of the statistical approaches uses non-parametric techniques that do not assume the knowledge of the data distribution. One stream of the non-parametric techniques was based on histogram analysis (Endler, 1998; Hofmeyr, Forrest, & Somayaji, 1998). Another stream was the kernel-based approaches, which approximate the density distribution using kernel functions (Desforges, Jacob, & Cooper, 1998).

The statistical approach was traditionally developed for univariate data, or multivariate data with known distributions. Most of the techniques are not directly applicable to complex unknown data because this approach is effective for analysis with independent variables. However, the statistical approach is the basis of many other techniques, which are built on statistical concepts. Many techniques convert the original multivariate problem into simple data representation (e.g. univariate) in order to use the basic statistical approach.

### **2.3.2 Classification-based Anomaly Detection Approaches**

The classification methods are used in supervised learning, which requires a training data set with examples of normal cases and (or) abnormal cases. In the training phase, boundaries between classes are created from learning the labeled data. Then in the testing phase, the trained classifier is used to assign a test data instance into one of the classes. Classification models used in this approach include neural networks, Bayesian networks, Support Vector Machines(Cortes & Vapnik, 1995), decision trees and regression models.

However, a dataset with both normal and abnormal class labels is not available for training in many anomaly detection problems. Variations and extensions of the traditional classification techniques are developed to perform semi-supervised learning. One-class classification (Tax & Duin, 1999) was developed to distinguish one class of objects from all other objects. Typically, data belonging to the normal class are available for training. Outliers are the data instances that fail to be classified into the normal class. Occasionally, only examples of abnormal cases are known. Anomalies are detected by testing data instance in the learnt anomaly dictionaries (Cabrera & Lewis, 2001).

The classification-based techniques are not directly suitable to address the anomaly detection problem in this research because the goal is to detect anomalies without prior knowledge of what is “normal” and what is “abnormal”.

### **2.3.3 Cluster-based Anomaly Detection Approaches**

Cluster analysis refers to techniques that identify groups of data points such that data points in the same group are similar to each other than to those in other groups. The groups of data points are called as clusters in these techniques.

Cluster-based anomaly detection techniques detect objects that do not belong to any cluster, or belong to very small clusters, assuming normal data are grouped in dense and large clusters. There are several different ways to perform the cluster analysis, such as partition-based, hierarchy-based, proximity-based, etc.

#### **Partition-based clustering**

K-means is the most basic and representative technique in this category. It partitions the observations into  $k$  clusters to minimize the intra-cluster difference and maximize the inter-cluster difference. It finds the “gravity” center of each cluster, named centroid. When it is used for anomaly detection, the clustering is usually performed on training data that only have normal classes. The centroids (or medoids) are considered as normal prototypes. The most distant point in a cluster determines the radius of that cluster. A test instance is compared with the  $k$  prototypes. The outlierness is measured by comparing the distance to the centroid with the cluster radius. Variations of k-means include k-medoids that chooses observations as centers, and fuzzy c-means that allows for fuzzy partition, rather than hard partition.

K-means has been used for novelty detection in online news and stories (Allan, Carbonell, & Doddington, 1998), and k-medoids has been used for fraud detection (Bolton & Hand, 2001). CLARANS is a popular k-medoids based algorithm, which can detect outliers as a by-product of the clustering process (Ng & Han, 1994).

The biggest limitation for the partition-based clustering algorithms is that the number of clusters needs to be specified in advance. If the number of clusters is not assigned appropriately, the clustering structure obtained can be ill-defined.

## Hierarchical clustering

A hierarchical clustering method seeks to build a hierarchy of clusters. Two types of hierarchical clustering methods are often distinguished: agglomerative and divisive, depending upon whether a bottom-up or top-down strategy is followed. In the agglomerative method, each object is in its own cluster at the bottom of the hierarchy, and pairs of clusters are merged based on similarity as one moves up the hierarchy, until all the objects are in a single cluster or until certain termination conditions are satisfied. The divisive approach works in the opposite way. In either approach, two measures are needed to decide how to merge (or split) the clusters: a measure of pairwise distance between observations, and a linkage criterion that specifies the dissimilarity between sets. The results of clustering are usually presented in a dendrogram. A hierarchy offers the clustering structure at various levels, so data can be reviewed for novelty at a fine-grained or less specific level. The basic hierarchical clustering method was not efficient in handling large datasets. BIRCH (Zhang, Ramakrishnan, & Livny, 1996) and CURE (Guha, Rastogi, & Shim, 1998) are two examples of the algorithms that employ hierarchical approach to clustering large datasets.

Hierarchy clustering is usually performed as a descriptive task before anomaly detection. Algorithm parameters for anomaly detection are better selected after understanding the data distribution. Hierarchy can also be directly used to identify outliers. For example, Baker et al. (1999) employs the hierarchical approach in news story monitoring.

## Proximity-based clustering

The proximity-based approaches find clusters by measuring how close a point is to its neighbors, assuming points in a cluster are close to each other and have similar proximity to its neighbors. The distance function needs to be defined to measure the proximity. DBSCAN is a common proximity-based clustering algorithm (Ester, Kriegel, Sander, & Xu, 1996). A cluster starts with  $k$  points within  $\epsilon$  distance neighborhood (the density criterion), and grows by finding the neighbors of the points already in the cluster which satisfy the density criterion. Outliers are the points that cannot form a cluster. The method does not require prior knowledge of the number of clusters in the data. Clusters are automatically formed until all data points have been processed. DBSCAN is used for one of the anomaly detection algorithms developed in this thesis.

Some techniques were developed with the only objective of detecting outliers. Outliers are identified in an optimized way, rather than as a by-product of clustering. The algorithm developed by (Breunig, Kriegel, Ng, & Sander, 2000) assigns an outlier score to every data point, referred as Local Outlier Factor (LOF), which is calculated by comparing the local density of an object to the local densities of its neighbors. Points that have a substantially lower density than their neighbors are considered outliers. LOF shares the similar concepts with DBSCAN and OPTICS in measuring density. LoOP (Kriegel, Kröger, & Zimek, 2009) was developed to overcome the problem of how to interpret the numeric outlier score in determining whether a data object indeed is an outlier. The LoOP technique provides an outlier score in the range of [0, 1] that is directly interpretable as a probability of a data object for being an outlier.

Along the proximity-based clustering, many techniques were developed to overcome the limitations of DBSCAN. DENCLUE (Hinneburg & Keim, 1998) was built on a firm mathematical basis and has good clustering properties in handing noise. The OPTICS algorithm (Ankerst, Breunig, Kriegel, & Sander, 1999) provides information of the intrinsic clustering structure that offers additional insights into the distribution of the data.

Besides the three clustering approaches introduced above, many other clustering techniques have been developed recently. The efforts were made to address problems with high dimensionality, large datasets, and time series. Some techniques use different measures for similarity, such as connectivity-based measures, angle-based measures, etc. Some techniques focused on reducing dimensions, such as subspace clustering and projected clustering. Since the research problem is a multivariate time series problem, the methods developed for time series are reviewed separately in Section 2.4.

The cluster-based approaches suit the requirements of this research problem the best among the basic anomaly detection approaches. One advantage is that cluster-based approaches do not have to be supervised. Moreover, they are capable of being used in an incremental mode. After learning the clusters initially, new flights data can be fed and tested for outliers, and the clusters can be updated with the new data.

Lastly, the three groups of anomaly detection approaches are not mutually exclusive. Some anomaly detection techniques combine different approaches together. For example, the Gaussian Mixture Model (GMM) is a combination of the statistical approach, classification approach and

clustering approach. A GMM is a parametric probability density function represented as a weighted sum of Gaussian component densities(Dempster, Laird, & Rubin, 1977; McLachlan & Basford, 1988; Reynolds, 2008). When it is used for anomaly detection, a mixture of Gaussian components for normal data is developed. Each component is a multivariate Gaussian that represents a type of normal data. The test data are considered as anomalies if they do not belong to any Gaussian component. Compared to K-means, GMM is able to give statistical inferences on clusters. Because of its flexibility and statistical features, GMMs have been commonly used to model complex multivariate data for a number of applications, most notably speech recognition(Reynolds, 2008). It is chosen to perform cluster analysis in one of the anomaly detection algorithms developed in this thesis.

## 2.4 Anomaly Detection Techniques for Time Series

Anomaly detection in time series has gained great interest due to an increasing need in many domains. In a wide range of fields, a huge amount of data is collected at specific time intervals and each sample is linked to previous and upcoming values; yet, it is challenging to utilize the collected data effectively. Most of the techniques described in Section 2.3 are dedicated to non-structured data and therefore are not adapted to exploit temporal relationships among observations. The special constraints and requirements brought by time series generate a number of specific anomaly detection techniques (Chandola et al., 2009). The clustering techniques for time series were reviewed in (Liao, 2005). A broader view on all kinds of data mining techniques for data streams was provided in (Gaber, Zaslavsky, & Krishnaswamy, 2005).

Common techniques for anomaly detection on time series data are reviewed in this part. Despite their differences, all of the techniques have two basic components: measuring the dissimilarity between time series, and identifying outliers based on the dissimilarity. The former component is the key for time series techniques. The latter component often applies one of the techniques described in Section 2.3 directly or with modifications. Thus, the techniques are grouped into two categories based on how the dissimilarity is measured: data-based and model-based as shown in Figure 2.6.

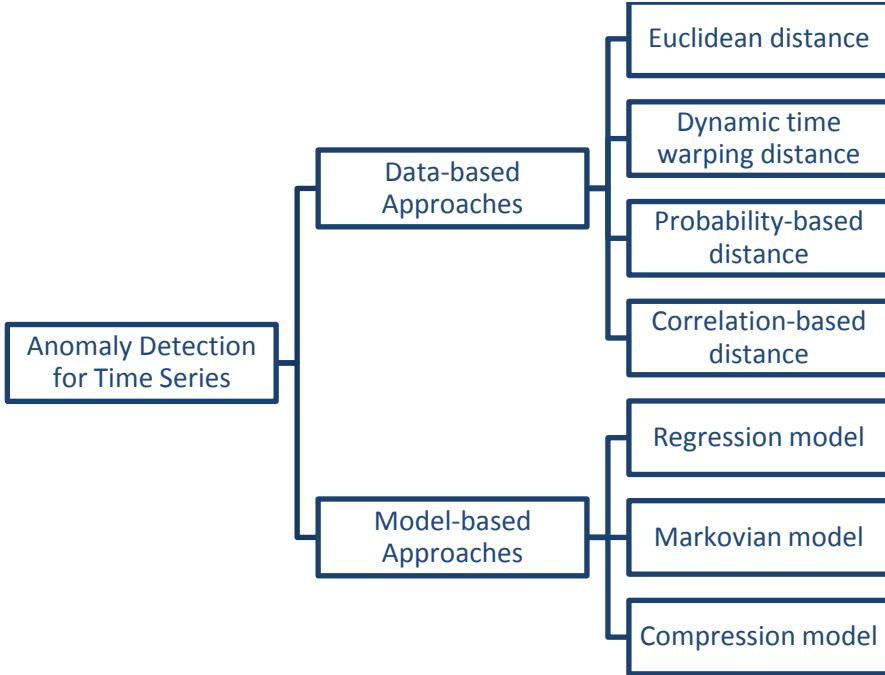


Figure 2.6 Categorization of Anomaly Detection Techniques for Time Series

## 2.4.1 Data-based Approaches

Data-based approaches measure the dissimilarity between two time series based on data observations directly, without fitting the data into a model. The dissimilarity is measured by a distance function, which differs by technique.

**Euclidean distance.** Approaches in this category measure the dissimilarity between two time series by aligning observations according to a temporal reference, and calculate the difference between every observation. They transform the sequential data into a vector of attributes and then calculate the Euclidean distance between vectors. The attribute value is determined by the value at each time. Then the problem of time series becomes a problem of static data with high dimensions. For example, a time series with a length of 50 samples can be converted to a vector of 50 dimensions. The temporal information is kept by increasing the number of dimensions. The similarity can be measured in the vector space using the standard Euclidean distance, the root mean square distance, Mikowski distance, or any other variations. This approach has been adopted in a wide range of time series analysis, such as identifying cyclone regimes in weather data (Blender, Fraedrich, & Lunkeit, 1997), monitoring the flight patterns in flight tracks data

(Gariel, Srivastava, & Feron, 2010), etc. Box Modeling was introduced as a transformation technique to incorporate un-equal length sequences (Chan & Mahoney, 2005). In general, Euclidean distance is very sensitive to any shift in the data. If two samples are not perfectly aligned, the distance can be huge. Also if the event involves several phases that can shrink or expand, Euclidean distance does not perform well, e.g. it is useless in voice recognition.

**Dynamic time warping distance.** Dynamic time warping (DTW) aligns two time series so that their difference is minimized. The DTW distance measures the dissimilarity between two series after an optimal alignment under certain constraints. Many algorithms use DTW as data pre-processing step before clustering or anomaly detection (Keogh, Lonardi, & Ratanamahatana, 2004). High-level patterns in the time series can be better compared by minimizing small shifts in time.

**Probability-based distance.** Some techniques extract probability information of each time sample from time series, rather than using raw data. Kumar & Woo (2002) assume each time series is made of samples drawn from Gaussian models, and then calculate the Chi-square statistics for every pair of time series as the similarity measure for hierarchical clustering. The distance function is scale-invariant so that clustering results are independent of the units of measuring data.

**Correlation-based distance.** The correlations between time series are used to measure dissimilarity in this approach. Various measures of correlation can be used. Golay et al. (1998) include the Pearson's correlation factor in a distance function. Möller-Levet, Klawonn, Cho, & Wolkenhauer (2003) proposed Short Time Series distance (STS) that measures the differences of slopes in two time series.

## 2.4.2 Model-based Approaches

Model-based approaches utilize the temporal structure in the time series. Time series models or sequence models are built based on the raw data. Then time series are compared based on the parameters of the models, or the residuals obtained from the models.

**Regression model.** Regression-based models are one of the earliest and widely used methods in modeling time series. Residuals or errors between the model prediction and the actual observation are used as the indicator of outliers. Rousseeuw & Leroy (1987) gave a comprehensive description on the use of robust regression to build models and detect outliers. Fox (1972) modeled

time series as an auto-regressive process (AR). Autoregressive moving average (ARMA) models were used in (Abraham & Box, 1979; Abraham, 1989). The autoregressive integrated moving average (ARIMA) model, a generalization of ARMA, was used in (Bianco, et al., 2001).

**Markovian model.** The markovian models are the most popular approach to model sequential behaviors in the time series that are not properly aligned in time. Time series are modeled using a Markov chain, Hidden Markov Model (HMM), or Finite State Machine (FSM). The outlierness of a sequence is determined by the likelihood of that sequence computed by a learnt model, or the difference between a prediction and an actual event. A great number of anomaly detection techniques has been developed using Markovian models, eg. Markov chain-based approaches (Ihler, Hutchins, & Smyth, 2006; Smyth, 1994). HMM-based techniques include (Boussemart, Las Fargeas, Cummings, & Roy, 2009; Ilgun, Kemmerer, & Porras, 1995; Sekar et al., 2002; Warrender, Forrest, & Pearlmutter, 1999). Most of data used in HMM approaches are symbolic sequences, so time series consisting of continuous values are converted to symbolic sequences before building Markovian models (Srivastava, 2005).

**Compression model.** Another branch of techniques extracts features from the time series using some compression methods, such as wavelet transformation, Fourier transformation, and Principal Component Analysis (PCA). The dissimilarity between time series is calculated based on extracted features. For example, (B. Amidan & Ferryman, 2000) summarized a time series by four coefficients (intercept, slope, quadratic, and error) using a quadratic equation model. (Vlachos, Lin, Keogh, & Gunopoulos, 2003) presented an approach to cluster time series at various resolutions using the Haar wavelet transform.

## 2.5 Anomaly Detection in Aviation

Anomaly detection in aviation systems has been focused on detecting defects on mechanical components in the past. Recently, more studies have been conducted to model and monitor complex systems. (B. Amidan & Ferryman, 2000; B. G. Amidan & Ferryman, 2005) were earliest efforts made to identify atypical flights using onboard recorded flight data. A software package called the “morning report” tool was developed. In their approach, each time series was summarized by four coefficients using quadratic equation model and then an “atypical score” for each flight was computed using the Mahalanobis distance. This approach was noteworthy but limited. First, it is a limited representation of a time series using only four coefficients. Important

features in the signal may not be captured by just four parameters. Second, using Mahalanobis distance as the only measure to detect outliers is not the best approach, as the distribution of flights in the feature space is complex and cannot be measured only by the distance to origin. Maille & Statler (2009) compared the “morning report” tool with the traditional FOQA analysis software on a set of digital flight recorded data. The study showed some potential value of using the “morning report” tool in finding newly emergent patterns. However, the performance and effectiveness of the “morning report” was not explicitly evaluated.

Some studies focused on detecting anomalies in discrete data. (S. Budalakoti, Srivastava, & Akella, 2006; Suratna Budalakoti, Srivastava, & Otey, 2008) developed an algorithm called sequenceMiner to discover anomalies in discrete parameter sequences recorded from flight data based on the Longest Common Subsequence (LCS) measure. Both clustering analysis and Bayesian model were used in this algorithm. (Srivastava, 2005) proposed a statistical framework that can work with both continuous data and discrete data based on HMMs. In the framework, the continuous data were pre-processed to symbolic data with a few representative states using clustering analysis, and then they were treated equally as the discrete data.

In another group of studies, normal behavior data were available and hence supervised or semi-supervised learning was the primary approach. (Iverson, 2004) developed the Inductive Monitoring System (IMS) software for real time monitoring of complex system. IMS used nominal data sets to extract general classes for typical system behaviors. During testing, real time operational data were compared with these classes to find any abnormal behaviors. Two application examples were presented, one for monitoring temperature sensors in the wings of a Space Shuttle Orbiter, and one for analyzing archived telemetry data collected from the ill-fated STS-107 Columbia Space Shuttle mission. (Schwabacher & Oza, 2007) compared four anomaly detection algorithms using data from two rocket propulsion testbeds. The four algorithms were Orca (Bay & Schwabacher, 2003), GritBot (RuleQuest Research software package), IMS (Iverson, 2004) and SVM (Tax & Duin, 1999). Except for Orca, training data were required for all four algorithms. They were designed for general complex systems, rather than explicitly for aerospace systems. Thus, temporal structures cannot be directly captured. (Das, Matthews, Srivastava, & Oza, 2010) introduced an approach based on kernel learning to incorporate the temporal structure. The method was a semi-supervised approach as it used one-class SVM for anomaly detection.

Some other studies on applying data mining techniques in aviation were generated due to the recent fast development in data mining techniques. (Larder & Summerhayes, 2004; Treder & Craine, 2005) were two examples that demonstrate the use of data mining techniques on various data collected at airlines, such as recorded flight data, safety reports, maintenance logs, etc. These two studies showed potential values of using data mining techniques in analyzing large amounts of data to find causalities, such as finding associations between incidents and flight information. However, anomaly detection problems in this research were not addressed in these studies.

# Chapter 3

## Cluster-based Anomaly Detection Algorithms

### 3.1 Challenges in Detecting Anomalies in FDR Data

This thesis develops a new approach to identify unknown issues from FDR data of routine airline flights. The new approach detects flights of interest without specifying what to look for in advance, and then relies on domain experts to review detected flights for operational implications. It is a step forward compared with existing methods, which rely on fixed flight parameters watch-list and predefined thresholds. Two challenges are identified in developing anomaly detection algorithms in this thesis:

The first challenge is detecting flights of interest from FDR data without specifying what to look for. No simple formulas can explain multivariate relationships among parameters as well as their temporal patterns. In general, flight parameters recorded in raw FDR data vary widely over time. An individual flight parameter is often associated with others depending on flight dynamics, specific procedures, and various environmental conditions. When one flight parameter changes, it will affect a number of others accordingly.

The second is the absence of prior knowledge on standard data patterns. Most existing anomaly detection methods assume one standard pattern, thus considering deviations from the standard as anomalies. However, multiple standard patterns exist in real-world operational data. Different phases of flight and variations in operational procedures, such as airport-specific procedures, air traffic control requirements, and company policies, may result in different data patterns. The assumption of one standard pattern is therefore not valid. The anomaly detection method in this thesis should be able to handle multiple standard patterns in the FDR data.

## 3.2 Concept of Cluster-based Anomaly Detection Algorithms

This section presents the concept of cluster-based anomaly detection algorithms and terminologies used in describing the algorithms: *cluster*, *outlier*, *anomalies*, and *abnormal flights*.

Two anomaly detection algorithms were developed to detect abnormal flights using the cluster-based concept. Cluster analysis is used to identify common patterns in FDR data. Multiple patterns exist in real-world operational data, however, the number of common patterns is finite because operations of commercial airline flights are highly standardized and a majority of flights share a few most common data patterns.

To facilitate cluster analysis, a prior step is to transform raw FDR data into high dimensional vectors, on which cluster analysis can be performed. The transformation can be performed in two ways: 1) convert the data of each flight for a specific phase into a vector 2) convert each data sample of the original FDR data into a vector. Both techniques were explored and developed into two detection algorithms: ClusterAD-Flight and ClusterAD-Data Sample. Details of each method will be presented in later sections in this chapter.

After data transformation, cluster analysis is performed on vectors in the high dimensional space. Groups of proximate vectors are identified as *clusters*, which represent common data patterns in the dataset; vectors that do not belong to any clusters are detected as *outliers*, which indicate uncommon data patterns.

In the last step, anomaly detection is performed based on cluster analysis result. In ClusterAD-Flight, outliers identified in cluster analysis are the *anomalies* to detect. In ClusterAD-Data Sample, both outliers and vectors that do not belong to appropriate clusters are the *anomalies* that we want to detect. Finally, the anomalies are summarized by flight. *Abnormal flights* are flights that have relatively more or severer anomalies.

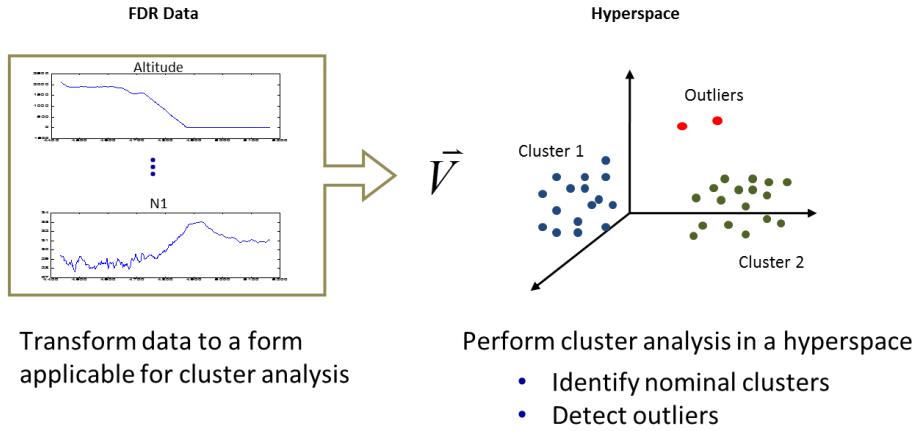


Figure 3.1 Concept of Cluster-based Anomaly Detection

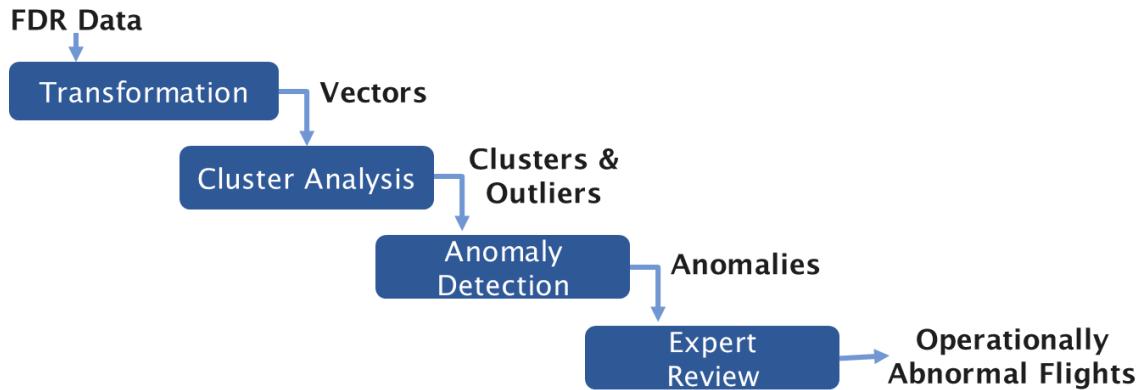


Figure 3.2 Framework of Cluster-based Anomaly Detection

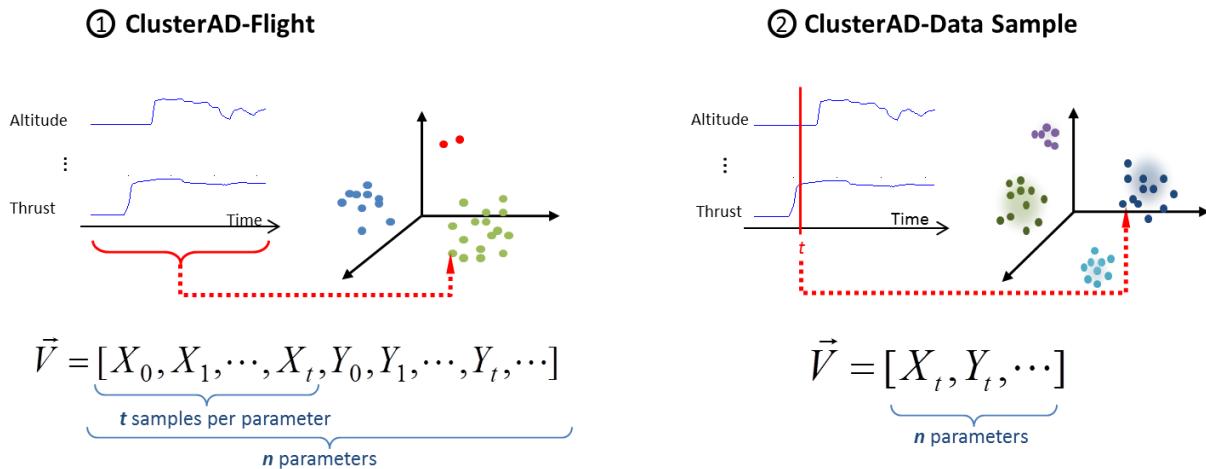


Figure 3.3 Two Different Ways of Data Transformation

### 3.3 Pattern-based Anomalies and Instantaneous Anomalies

Anomalies in FDR data can be categorized in two types: pattern-based anomalies and instantaneous anomalies. Figure 3.4 and Figure 3.5 shows examples of these two types of anomalies. In these two plots, anomaly signal is depicted in red. A normal profile is presented by blue areas – the center blue line shows the median of normal values, the dark blue area gives the range of 50% of normal values, and the light blue area depicts the range of 90% of normal values.

Pattern-based anomalies are data with abnormal patterns over a specific flight phase. As shown in Figure 3.4, pattern-based anomalies are observed in engine parameter “N1”, which measures fan speed and is representative of engine thrust. The profile of “N1” is different from the normal profile from 6nm before touchdown to 1nm before touchdown; each individual data sample is not significantly deviating from the normal value. In comparison, instantaneous anomalies are abnormal data that occurs instantaneously. Figure 3.5 gives an example of an instantaneous anomaly in “Angle of Attack”.

The objective of anomaly detection in this thesis is to detect both types of anomalies. Because distinct data transformation techniques are used in ClusterAD-Flight and ClusterAD-Data Sample, we expect the two methods to be sensitive to different types of anomalies. Thus, an evaluation on which types of anomalies are better detected by which method was performed in this research.

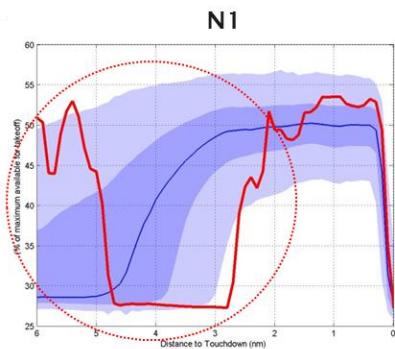


Figure 3.4 Pattern-based Anomaly Example

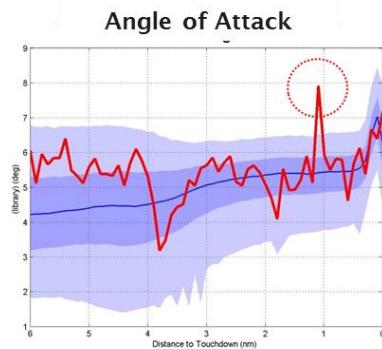


Figure 3.5 Instantaneous Anomaly Example

### 3.4 ClusterAD-Flight

ClusterAD-Flight converts data of an entire flight phase into a single point in a high-dimensional space, where data patterns are represented by vectors. It then uses cluster analysis to identify clusters and outliers in the high-dimensional space. Anomalies are detected from such outliers. ClusterAD-Flight consists of three key steps, as illustrated in Figure 3.6:

1. Data transformation: transforming time series into high-dimensional vectors
2. Dimension reduction: addressing problems of multicollinearity and high dimensionality
3. Cluster analysis: identifying clusters and outliers in high-dimensional space

The first step transforms multivariate time series data into high dimensional vectors. The transformation technique anchors time series by a specific event, such as touchdown, which reserves the temporal information and makes it comparable among different flights. Then, in the second step, techniques are developed to address problems of multicollinearity and high dimensionality. The dimensionality of vectors is reduced for computational viability while maintaining essential information. In the last step, cluster analysis is performed to detect outliers and clusters of normal flights in the feature space of reduced dimensions. Each step is described in detail in the following paragraphs.

ClusterAD-Flight is limited to flight phases that start or end with a specific event: takeoff or final approach. These two phases are critical phases in terms of safety because 53% of fatal accidents and 47% of onboard fatalities happened during those two phases for worldwide commercial jet fleet from 2002 to 2011 (Boeing Commercial Airplanes, 2012).

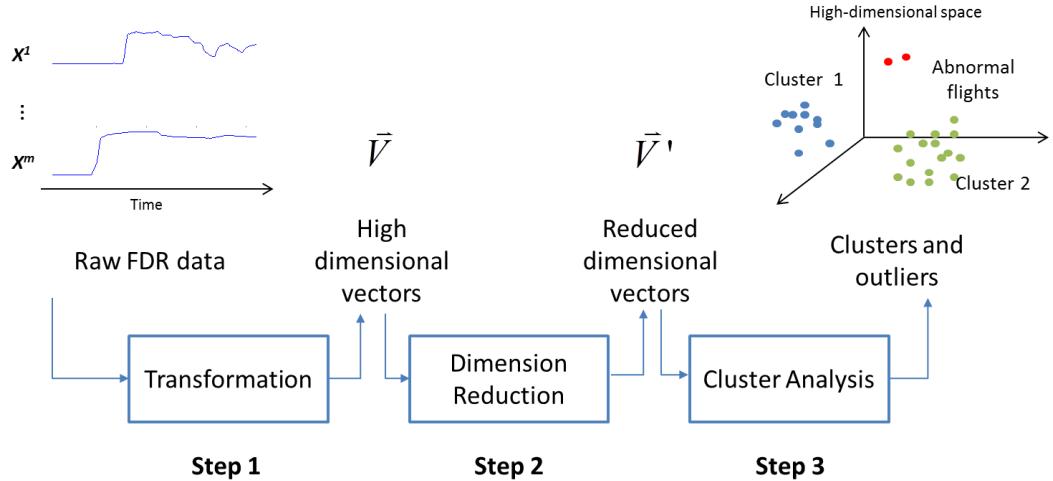


Figure 3.6. Cluster-based Detection Algorithm: ClusterAD-Flight

### 3.4.1 Data Transformation

In order to map raw data into comparable vectors in the high dimensional space, time series data from different flights are anchored by a specific event to make temporal patterns comparable. Then, every flight parameter is sampled at fixed intervals by time, distance or other reference from the reference event. All sampled values are arranged to form a vector for each flight:

$$[x_{t_1}^1, x_{t_2}^1 \dots, x_{t_n}^1, \dots, x_{t_j}^i, \dots, x_{t_1}^m, x_{t_2}^m, \dots, x_{t_n}^m]$$

where  $x_{t_j}^i$  is the value of the  $i^{th}$  flight parameter at sample time  $t_j$ ;  $m$  is the number of flight parameters;  $n$  is the number of samples for every flight parameter. The total dimensionality of every vector is  $m * n$ . Each dimension represents the value of a flight parameter at a particular time. The similarity between flights can be measured by the Euclidian distance between the vectors.

Raw FDR data are anchored from a specific event and sampled at fixed intervals by time, distance, or other reference. For the takeoff phase, the time of takeoff power application is used as the reference time and a number of samples are obtained at fixed time intervals, as shown in Figure 3.7. For the approach phase, the time series are first transformed into a “distance-series” and then a number of samples are obtained backtracking from the touchdown point (Figure 3.8).

Distance is used as the reference rather than time in the approach phase as procedures during approach are often specified based on distance or height above ground.

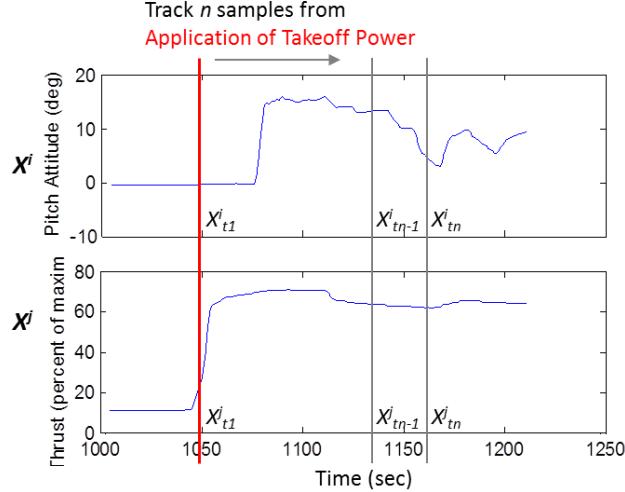


Figure 3.7. Sampling Time Series in Takeoff Phase

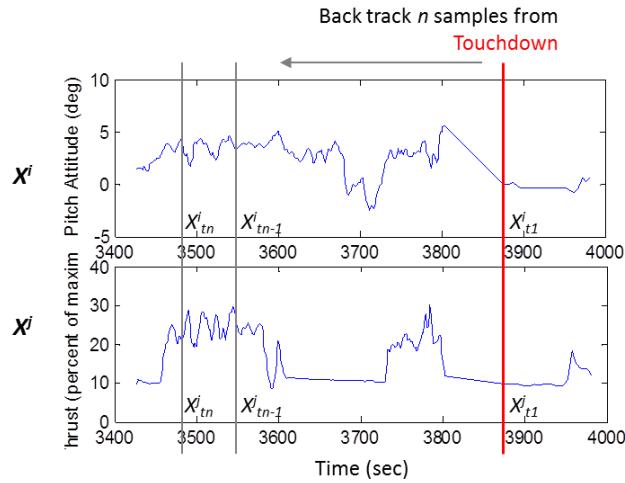


Figure 3.8. Sampling Time Series in Approach Phase

### 3.4.2 Dimension Reduction

Because of the temporal aspect, the vectors formed in the first step will normally have thousands of dimensions. For example if 100 parameters are evaluated over 100 time steps, this will result in an 10,000 dimension analysis space. In comparison, the number of daily flights at a large airline is on the order of 1000 flights. This implies the typical daily dataset will have more

dimensions than data points. It is difficult to identify data clouds in such sparse distribution. Therefore, Principal Component Analysis (PCA) was used to reduce the number of dimensions before performing cluster analysis. PCA is a common procedure to transform data into an orthogonal coordinate system based on the variance in the data (Hotelling, 1933). The coordinates in the new system are referred as components. The largest variance by any projection of the data comes to lie on the first component, the second largest variance on the 2nd, and so on. As a consequence, a number of last components could be dropped to reduce the hyperspace dimension without losing significant information. In this study, the first  $K$  components that capture 90% of the variance in the data are kept.

$$\sum_{i=1}^K \lambda_i \Bigg/ \sum_{i=1}^N \lambda_i > 90\%$$

where  $\lambda_i$  is the variance explained by principal component  $i$ .  $N$  is the total number of principal components, which equals to the original number of dimensions.  $K$  is the number of principal components kept. The magnitude of dimensional reduction will vary with the dataset but can be significant. As an example, in the initial testing of ClusterAD-Flight discussed later in this chapter (Section 3.5), the dimensions were typically reduced from 6188 to 77 for the takeoff data and from 6279 to 95 for the landing data using this criterion.

The use of PCA is unnecessary for large datasets that are dense enough to apply cluster analysis. However, anomaly detection could be biased by correlations among parameters in the absence of PCA. For example, if a majority of parameters in a FDR dataset are engine related, anomalies would be dominated by engine problems in the absence of PCA. Correlations between parameters are common in FDR datasets. As an example, linear correlations among parameters in the dataset used in the initial testing of ClusterAD-Flight are shown in Figure 3.9.

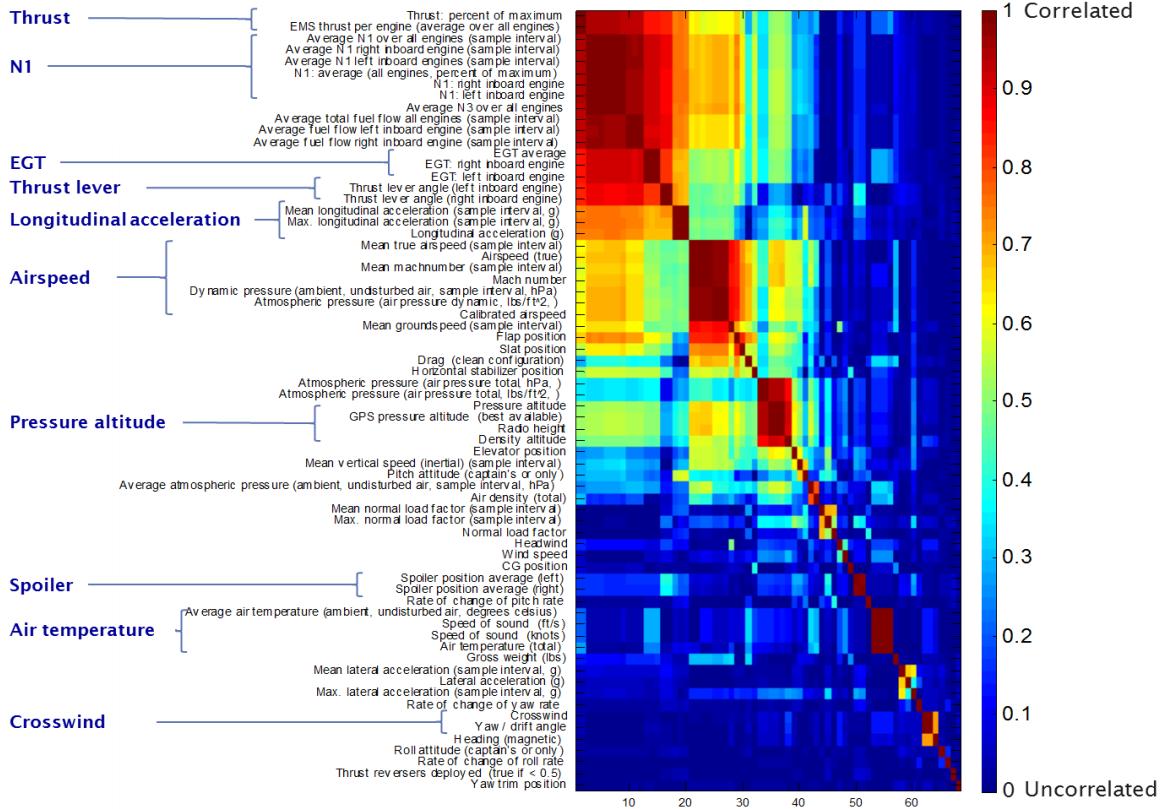


Figure 3.9 Correlation Matrix of Original Flight Parameters (Dataset: 365 B777 flights)

The thesis proposes a solution to weaken the effect of correlated parameters by first identifying sets of correlated parameters and then combining each set of correlated parameters into two measures: the average and the maximum differences of all parameters in the set. The former captures the general trend, while the latter examines abnormal patterns. Pearson correlation coefficients were used to identify sets of correlated parameters. It is widely used as a measure of the strength of linear dependence between two variables. It was developed by Karl Pearson from a related idea introduced by Francis Galton in the 1880s (Pearson, 1896; Rodgers & Nicewander, 1988; Stigler, 1989). After correlated parameters are identified, they are modified into new variables by set, which have much weaker linear dependence between each other. Figure 3.10 shows the linear correlations among modified parameters after de-correlation was performed.

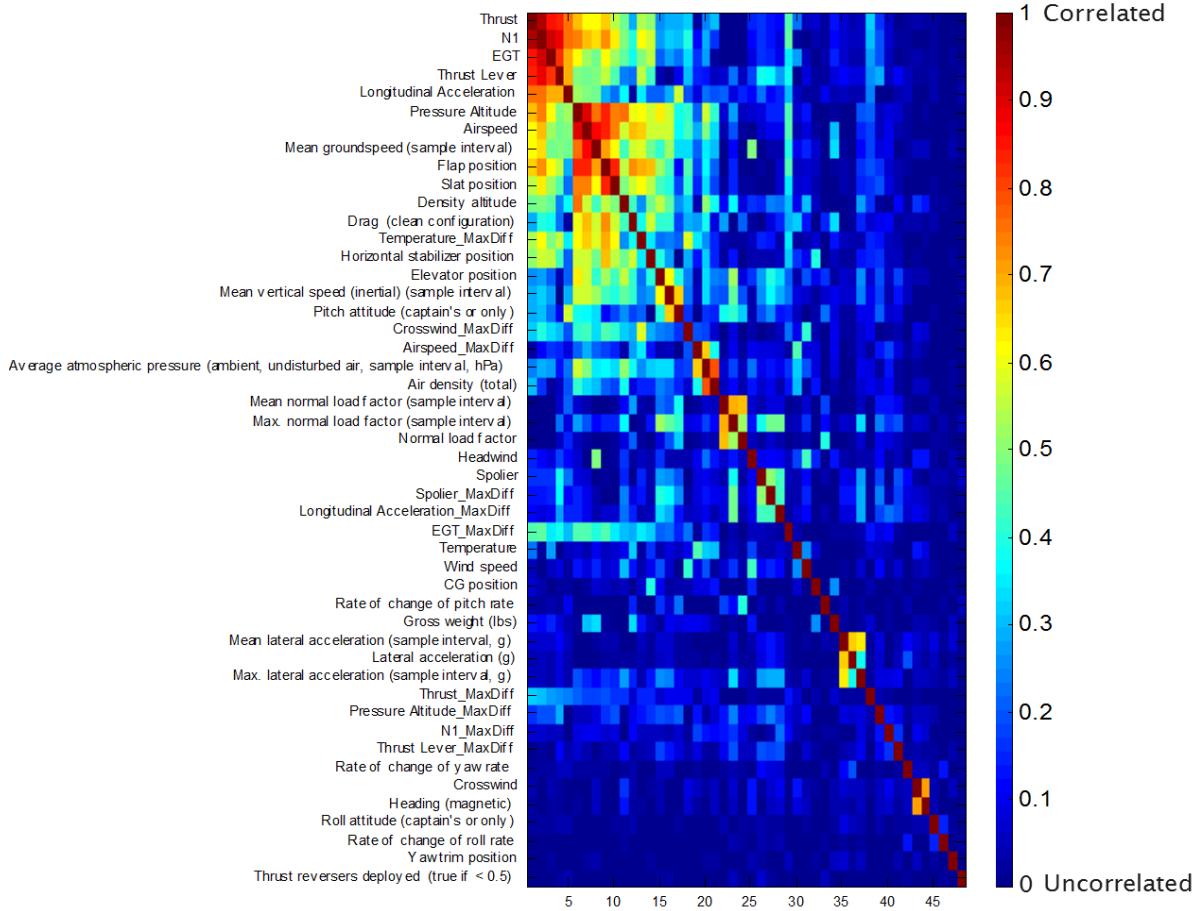


Figure 3.10 Correlation Matrix of Modified Parameters after De-correlation (Dataset: 365 B777 flights)

### 3.4.3 Cluster Analysis

The clustering analysis aims to identify clusters of data in the feature space. Then, outliers can be detected based on the identified clusters in the feature space. The Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm (Ester et al., 1996) was chosen to perform the cluster analysis because: 1) it can automatically determine the number of clusters; 2) it can handle data with noise/outliers; 3) it can detect outliers while identifying clusters.

DBSCAN is a density-based clustering algorithm. It progressively finds clusters based on a density criterion. A cluster forms if at least  $MinPts$  points are within  $\varepsilon$  radius of a circle. The cluster grows by finding the neighbors of the cluster, which also satisfy the same density criterion until no other point can be added into the existing cluster. At this point, it starts to search for a

new cluster. Outliers are the points that do not belong to any cluster. Other than the two parameters  $MinPts$  and  $\varepsilon$  to set the density criterion, no other parameters are required in DBSCAN.

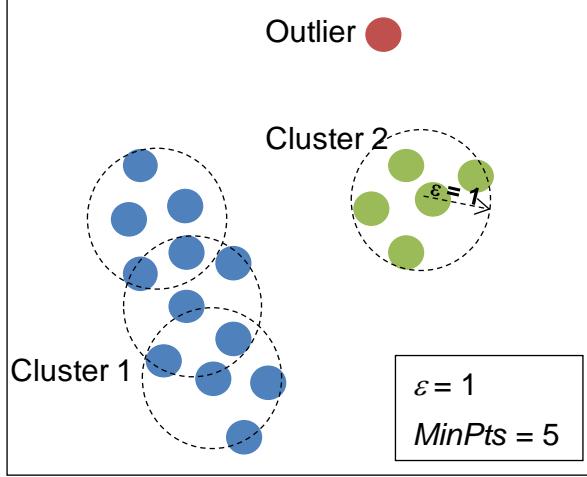


Figure 3.11. Example of DBSCAN Clustering Process

The selection of the two parameters is based on sensitivity analysis. For a fixed value of  $MinPts$ , DBSCAN is performed multiple times using a series of  $\varepsilon$  values ranging from the minimum pairwise distance to the maximum pairwise distance in the data.  $MinPts$  is chosen to be the minimum number of similar flights that can be considered as a separate nominal group. While the number of outliers is sensitive to the value of  $\varepsilon$ , the value of  $\varepsilon$  was set to match user's preferences: finding the top  $x\%$  outliers.

After the cluster analysis, outliers and clusters are identified in the space of reduced dimensions. Outliers represent the abnormal flights to be detected; clusters capture different types of normal flights in a dataset.

### 3.5 Initial Testing of ClusterAD-Flight

An initial testing of ClusterAD-Flight was performed using a representative DFDR dataset obtained from an international airline. The dataset consisted of 2881 flights including 7 aircraft types with 13 model variants, e.g. B777, A319, and A320. To obtain relatively homogeneous data, the dataset was filtered by model variant. Among the 13 model variants, the set of B777 was the

largest, which had 365 flights occurring over one month with various origins and destinations. This section presents results of testing the method on this dataset.

Outlier detection was conducted separately for the approach phase and the takeoff phase. Three sets of DBSCAN parameters were tested to identify the top 1%, 3% and 5% outliers identified by the method. All the identified abnormal flights were further analyzed to determine the flights were abnormal and if so to characterize the abnormal behaviors. In addition, when more than one cluster was present the differences in the nominal cluster data patterns were investigated to understand the cause of the different nominal behaviors.

### 3.5.1 Dataset and Data Preparation

The FDR dataset used for the initial testing of ClusterAD-Flight contained 365 B777 flights with various origins and destinations from an airline. Every flight included 69 flight parameters including engine parameters, aircraft position, speeds, accelerations, attitudes, control surface positions, winds, and environmental pressures and temperatures. Radio height was only available during approach phase.

To allow flights at different airports to be compared, the position related flight parameters were first converted to values relative to the airport. For instance, the original recorded altitude values (e.g. pressure altitude, density altitude) were transformed to relative altitudes (e.g. height above takeoff, height above touchdown).

For the transformation from time-series to vectors, observations were obtained at 1-sec intervals from takeoff power up to 90 seconds after takeoff for the takeoff phase. For the approach phase the same number of observations was obtained from 6 nm before touchdown to touchdown. After performing the PCA, the number of dimensions was reduced from 6188 (68 flight parameters \* 91 samples) to 77 for the takeoff phase and from 6279 (69 flight parameters \* 91 samples) to 95 for the approach phase.

The sensitivity to cluster selection criteria ( $\epsilon$  and MinPts) is shown for the Approach and Takeoff data in Figure 3.12 and Figure 3.13. It was observed that the selection was insensitive to MinPts (between 3 and 15) but that fewer flights are identified as outliers when  $\epsilon$  increases. Therefore MinPts was set at a value of 5 and the value of  $\epsilon$  was selected to find the top 1%, 3% and 5% outliers.

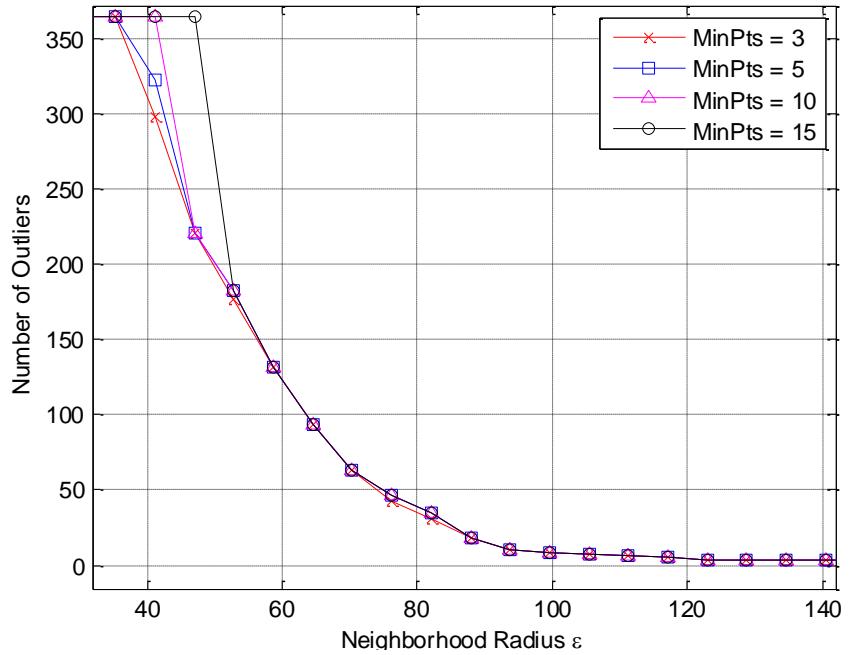


Figure 3.12. Sensitivity to  $\varepsilon$  and  $MinPts$  (Approach)

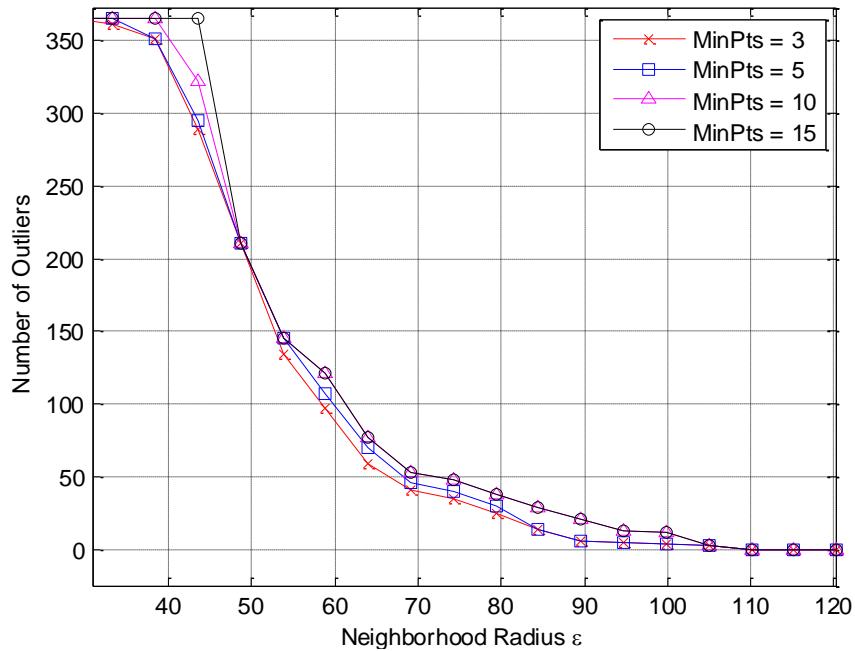


Figure 3.13. Sensitivity to  $\varepsilon$  and  $MinPts$  (Takeoff)

### 3.5.2 Results Overview

Three sets of abnormal flights were identified using different parameter settings to match the top 1%, 3% and 5% outlier criterion. The results are summarized in Table 3.1. Further

examination confirmed that abnormal flights found using a smaller outlier criterion were always included in the results obtained using a larger outlier criterion, as shown in Table 3.2.

Table 3.1. Number of Abnormal Flights Identified (Dataset: 365 B777 flights)

Find Top x% Abnormal Flights		DBSCAN Setting ( <i>MinPts</i> = 5)	Number of Abnormal Flights
Approach Phase	1%	$\varepsilon = 122.5$	3
	3%	$\varepsilon = 93.9$	10
	5%	$\varepsilon = 89.7$	16
Takeoff Phase	1%	$\varepsilon = 100.0$	4
	3%	$\varepsilon = 85.8$	9
	5%	$\varepsilon = 83.4$	22

Table 3.2 Abnormal Flights Identified Using Different Detection Thresholds in Takeoff Phase

Flight ID	Find top x% abnormal flights		
	1%	3%	5%
370715	x	x	x
380219	x	x	x
371045	x	x	x
371046	x	x	x
377862		x	x
385702		x	x
378692		x	x
386369		x	x
370723		x	x
380217			x
383285			x
384110			x
385160			x
379636			x
369755			x
385444			x
370019			x
368486			x
368487			x
372209			x
373921			x
369204			x

### **3.5.3 Abnormal Behaviors in Flights Detected**

All of the top 5% abnormal flights were further analyzed to determine if they exhibited abnormal behaviors by comparing their flight parameters with distribution of flight parameters from all the flights.

Two examples are presented in details in this section to show how the abnormal behaviors were identified. The most distinctive flight parameters for each example are presented in graphs that use the same format. The abnormal flights are shown by red lines. The patterns of most flights are depicted by blue bands. The dark blue bands indicate the 25<sup>th</sup> to the 75<sup>th</sup> percentile of all flights' data; the light blue bands encompass the 5<sup>th</sup> to the 95<sup>th</sup> percentile. Respectively, the dark blue region contains 50% of the data, while the light blue region covers 90%. The details on how these plots are generated are presented in Section 4.2.1 in Chapter 4.

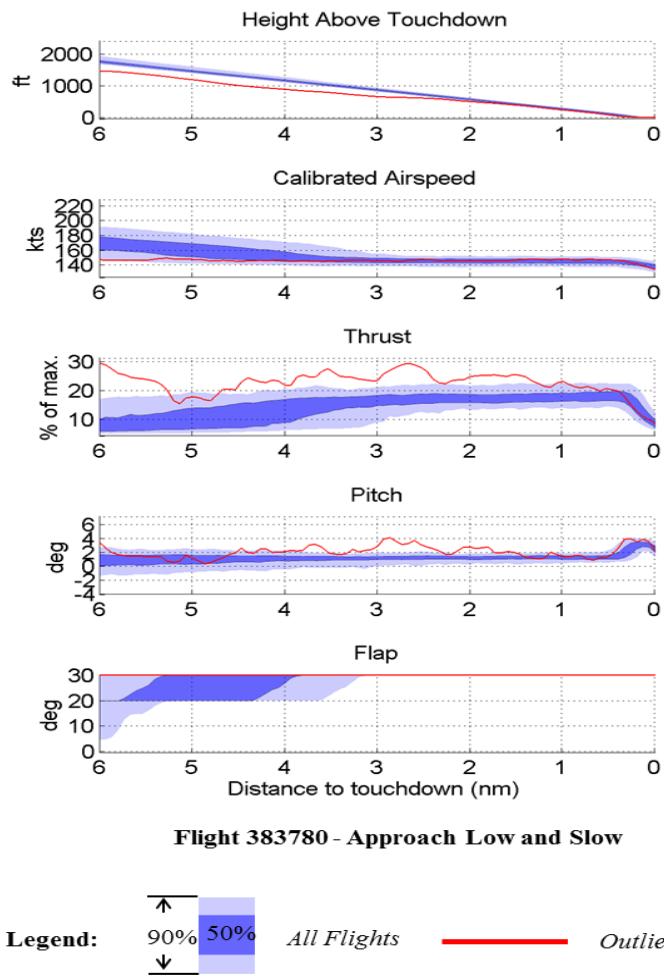
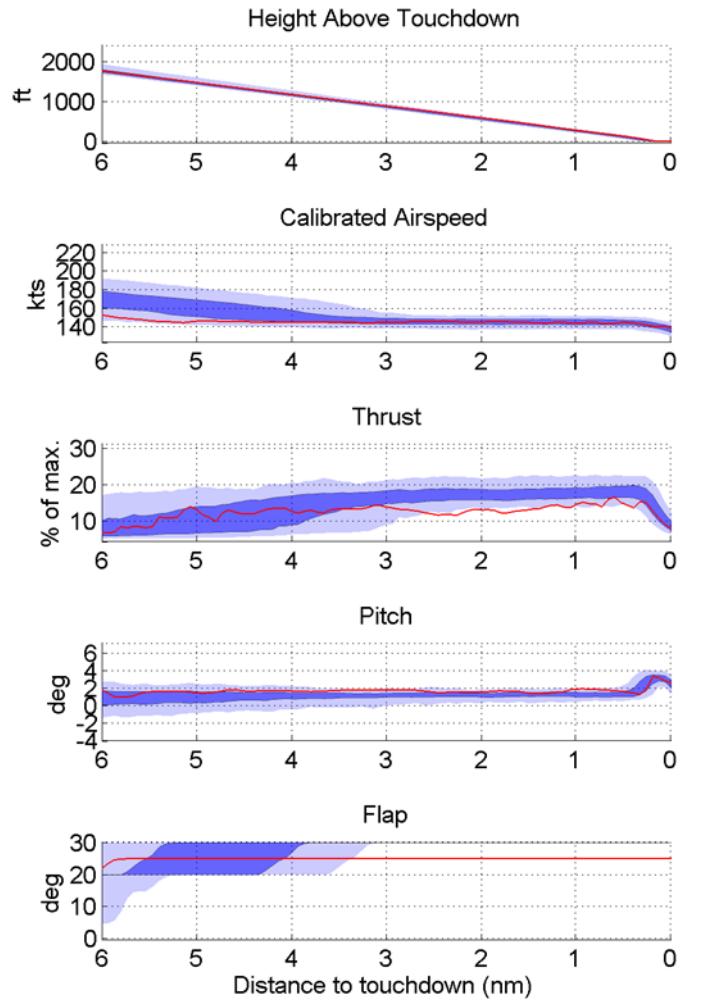


Figure 3.14. Example Abnormal Flights in Approach Phase: Approach Low and Slow

Flight 383780 is a low and slow approach (Figure 3.14). The vertical profile is always below the common glide slope until 2 nm before touchdown and the calibrated airspeed is lower than most other flights until 3 nm before touchdown. Moreover, the flap is set to the landing configuration, 30, from at least 6 nm before touchdown. Therefore, this flight has to use a much higher thrust than most others until touchdown. It is also noted that a higher than normal pitch attitude is used to catch the glide slope between 3 nm and 2 nm before touchdown.



**Flight 383285 - Unusual Flap**



Figure 3.15. Example Abnormal Flights in Approach Phase: Unusual Flap Setting

Flight 383285 uses Flap 25 all the way up from 6 nm before touchdown until landing, while most other flights are using Flap 30 as the landing configuration, as shown in Figure 3.15. So less thrust is needed for the final part of the approach than most flights. Meanwhile, major indicators of the approach performance, the altitude, the airspeed and the pitch, are within the 90% normal range.

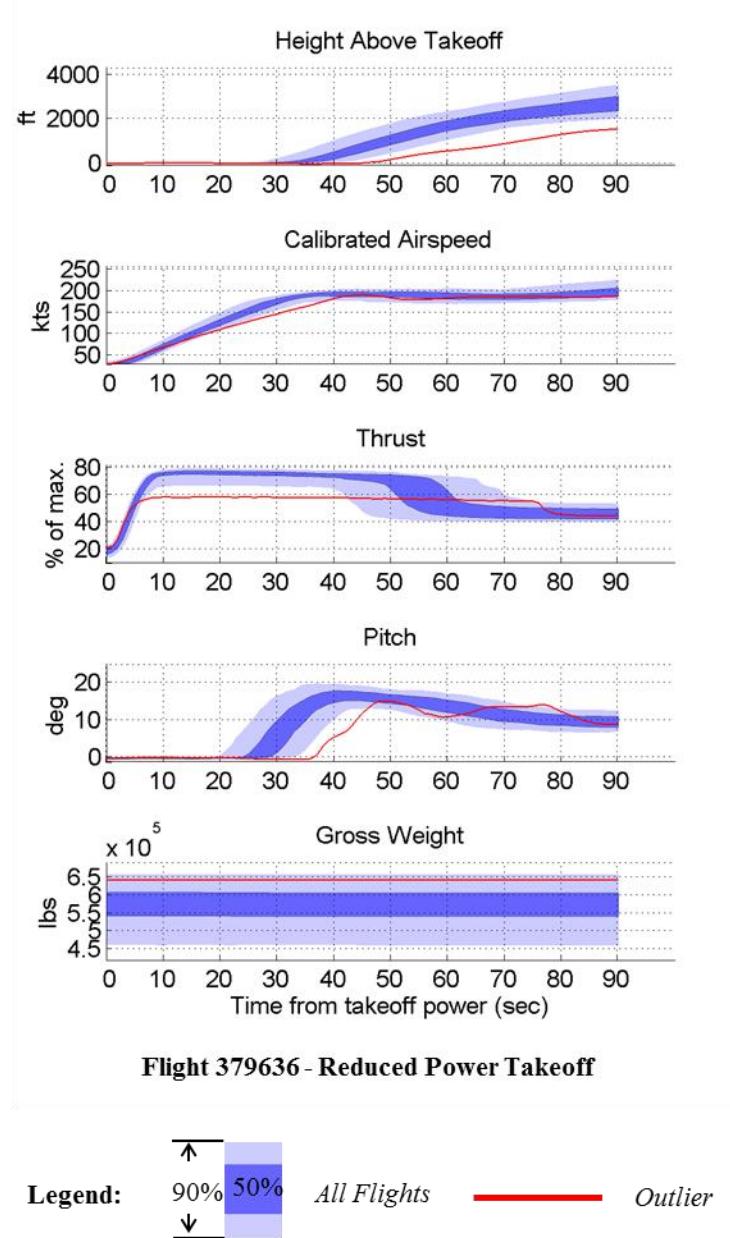
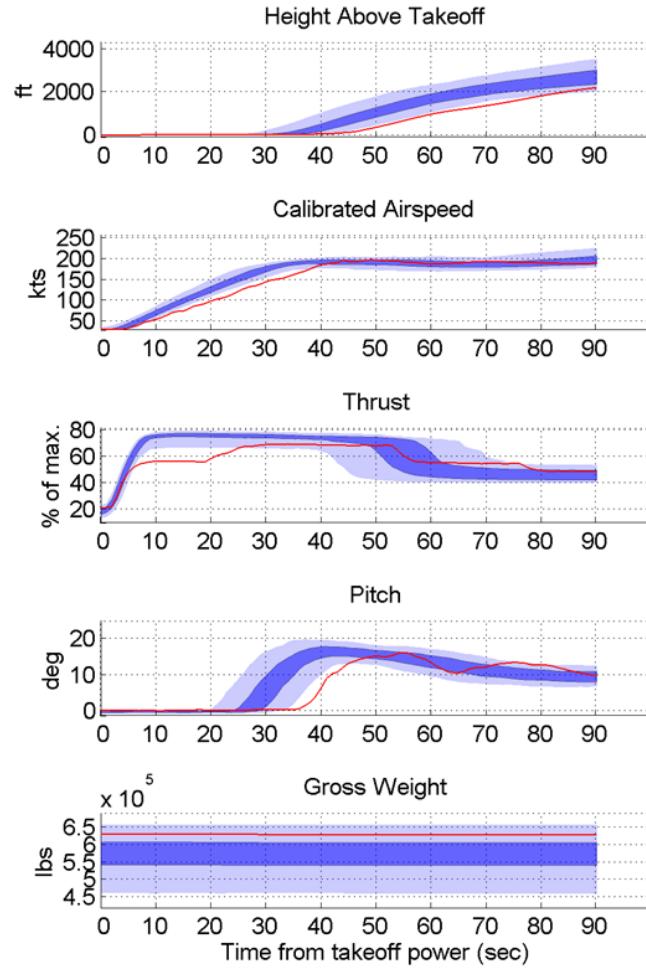


Figure 3.16. Example Abnormal Flights in Takeoff Phase: Reduced Power Takeoff

Example takeoff anomalies are shown in Figure 3.16 and Figure 3.17. Flight 379636 shown in Figure 3.16 used a much lower takeoff power than most other flights although it was at a relatively heavy weight. The degraded takeoff performance is apparent as the aircraft accelerates slowly and the rotation is not made until the airspeed reaches the required level. In addition, at 80 seconds after applying takeoff, the pitch reaches 15 degree that is similar to the angle during

initial rotation. As the aircraft is relatively underpowered, the climb rate is much lower than other flights as well.



**Flight 372209 - Takeoff Power Change**



Figure 3.17. Example Abnormal Flights in Takeoff Phase: Changed Power Takeoff

Flight 372209 (Figure 3.17) displays behavior similar to Flight 379636 for the first 20 seconds after applying takeoff power, as shown in Figure 3.16 right column. However, the power setting is changed back to normal level before rotation happens. As a result, the takeoff performance is better than Flight 379636. However, the climb rate and the acceleration are still lower than most other flights.

The same analysis was performed for all abnormal flights detected. All identified flights exhibited some identifiable degree of anomaly. For the approach phase, the most frequent abnormal behaviors were high energy approaches and low energy approaches (Table 3.3). Some flights were found to have unusual operations, such as abnormally high pitch, unusual flap settings, and lining up with localizer relatively late. In addition, environmental anomalies, such as strong crosswind and high atmospheric temperature, were found in some of the abnormal flights.

Table 3.3. Abnormal Behaviors in Approach Phase of the Top 5% Abnormal Flights (Dataset: 365 B777 flights)

Flight ID: Abnormal behaviors
<b>High energy approaches</b>
371040: Fast
373547: Fast
377860: Fast
378688: Fast, unstable airspeed
377844: High, line up late
377288: Initially fast, then normal
379685: Initially fast, then slow
<b>Low energy approaches</b>
383780: Low, slow
375698: Low, high power
383270: Low, unusual yaw trim
<b>Other Unusual operations</b>
383285: Unusual flap setting
384110: Unusual flap setting
371044: Abnormal high pitch
371045: Line up late
<b>Environmental anomalies</b>
372235: High atmosphere temperature
379665: Strong crosswind

The abnormal behaviors in takeoff are summarized in Table 3.4. The most frequent abnormal behaviors were high and low power takeoffs which often include other notable factors. Also observed were: excessive reduction of power after takeoff, double rotation, and high pitch attitude during takeoff. It should be noted that not all abnormal flights identified indicate safety concerns. Some flights were identified as abnormal but were benign cases, such as the takeoff in strong wind and the flight that turned soon after takeoff.

Table 3.4. Abnormal Behaviors in Takeoff Phase of the Top 5% Abnormal Flights (Dataset: 365 B777 flights)

Flight ID: Abnormal behaviors
<b>High power takeoffs</b>

377862: Early rotation, high & fast climb out 380217: Early rotation, early turn 380219: Early rotation, crosswind 370715: Light, accelerate fast, climb fast 383285: Light, early rotation, early turn 384110: Light, climb out high, early turn 385160: High climb out, high pitch rotation
<b>Low power takeoffs</b>
368486: Reduced power, low & slow climb out 368487: Reduced power, low & slow climb out 369755: Reduced power, low & slow climb out 370019: Reduced power, low & slow climb out, extended period of high pitch 371045: Reduced power, low climb out 371046: Reduced power 379636: Reduced power, low climb out 385444: Reduced power, low & slow climb out
<b>Abnormal power settings</b>
369204: Excessive power reduction after takeoff 372209: Start with reduced takeoff power then switch to normal takeoff power, low & slow climb out 378692: Extended period of takeoff power
<b>Other Unusual operations</b>
373921: Double rotation 385702: High pitch rotation, climb out high 386369: Early turn after takeoff
<b>Environmental anomalies</b>
370723: Rise of spoiler, strong wind

### 3.5.4 Nominal Data Patterns from Clusters

The cluster analysis method can also be used to recognize different nominal patterns in the data which can be identified by different clusters. Each cluster represents a type of nominal data pattern. Typical operational behaviors can be characterized by retrieving flights from these clusters.

In this dataset, a single dominant cluster was found in the approach phase; while in the takeoff phase, a large cluster and two small clusters were identified. The result shows that most takeoffs shared a common data pattern and two small groups of takeoffs involved other patterns in this dataset. Table 3.5 summarizes the cluster structure identified using different density criterion. Cluster 2 was labeled as a separate cluster by all three outlier criteria, which indicates

that flights in Cluster 2 were distinctive from most flights. Flights in Cluster 3 were identified to belong to a separate cluster only when the 3% outlier criterion is used. The flights in Cluster 3 were merged into Cluster 1 using the 1% outlier criterion and were classified as outliers using the 5% outlier criterion. Cluster 3 can be viewed as a sub cluster at the border of Cluster 1.

The differences between the clusters can be seen in Figure 9. Flights belong to Cluster 2 were takeoffs at OR Tambo International Airport (ICAO: FAJS), near the city of Johannesburg, South Africa. Due to the high altitude (5558 ft MSL), the takeoff performance is degraded compared to most other flights, as shown in green in Figure 3.18. Flights belonging to Cluster 3 were reduced power or de-rated takeoffs. They are shown in orange in Figure 3.18. They show reduced power settings with subsequently late rotations and lower climb rates.

Table 3.5. Number of Flights in Clusters by Outlier Criterion in Takeoff Phase

	Outlier Criterion		
	1%	3%	5%
<b>Cluster 1</b>	353	341	335
<b>Cluster 2</b>	8	8	8
<b>Cluster 3</b>	--	7	--
<b>Outliers</b>	4	9	22

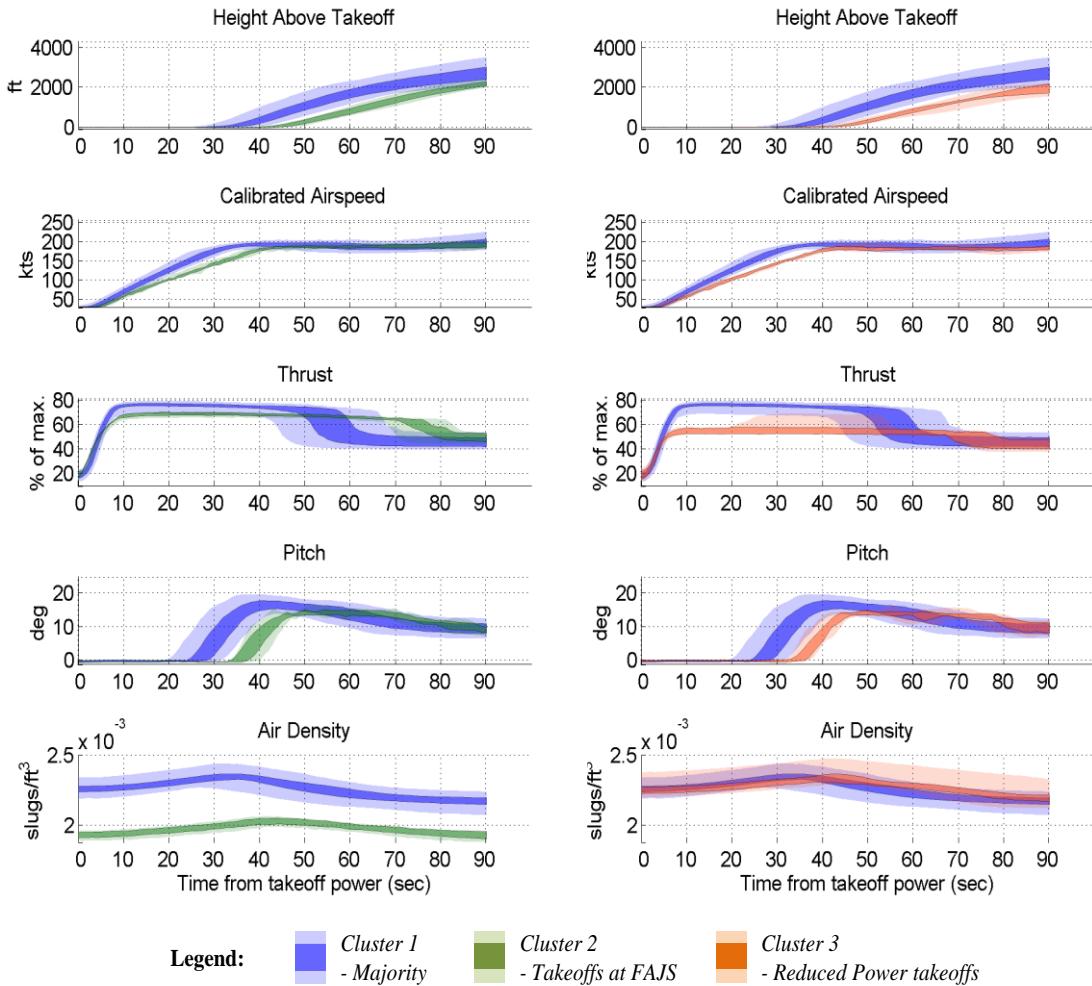


Figure 3.18. Patterns of Flights in Cluster 2 (Left) and Cluster 3 (Right) in Takeoff Phase

In summary, ClusterAD-Flight was applied to a representative Digital FDR dataset of 365 B777 flights. Abnormal flights were detected. The examination on these flights showed uncommon operations, e.g. high energy approaches, unusual pitch excursions, abnormal flap settings, high wind conditions, etc. In addition, multiple clusters representing nominal conditions were also detected. Three distinct takeoff clusters were identified in this data set: one represented a majority of the takeoff cases, one associated with a specific high altitude airport, one correlated with reduced power takeoffs.

## 3.6 ClusterAD-Data Sample

ClusterAD-Data Sample converts instantaneous data samples of FDR data into vectors for cluster analysis. Clusters identified in this space represent nominal modes of aircraft operations. An operational mode is a relatively stable state of aircraft, and it often takes place during a certain flight phase and can be described by a set of flight parameters. Commercial passenger flights follow highly standardized procedures, and these procedures result in a finite number of operational modes that an airplane could be in. Examples of an operational mode include level flight at a certain cruising speed, descent with a particular descent rate, final flare, touchdown, etc.

Operational modes determine values of flight parameters. Therefore, instantaneous data samples can be evaluated based on operational modes identified in cluster analysis. Whether a data sample is nominal or not is evaluated by two factors: 1) whether it belongs to a mode, 2) whether the mode is an appropriate one, e.g. flare mode is not appropriate at the beginning of an approach. In order to calculate these two factors, a parameterized model of modes and mode distribution are required.

Therefore, the following three steps are performed in ClusterAD-Data Sample:

1. Identification of nominal modes
2. Characterization of mode distribution
3. Anomaly detection

In the first step, nominal modes of an aircraft from FDR data are identified using cluster analysis. In the second step, the distribution of nominal modes is summarized by counting number of observations. Every nominal mode has a range that it is most likely to happen, e.g. the “final flare” mode is more likely to happen at the end of the approach phase. In the last step, anomalies are detected based on the nominal modes and their temporal distribution. Each step is described in detail in the following sections.

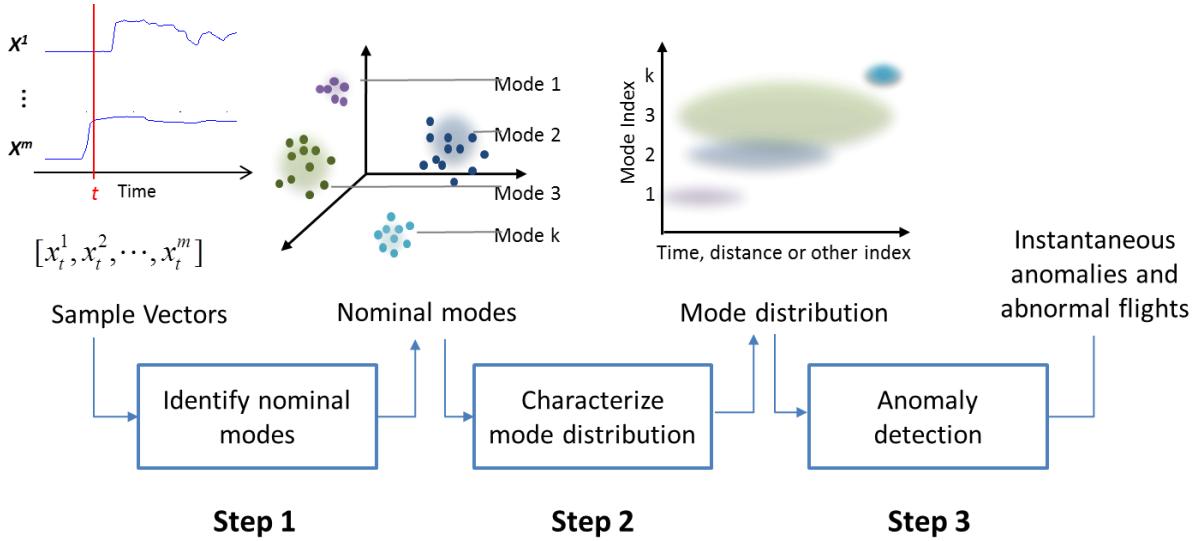


Figure 3.19 Cluster-base Detection Algorithm: ClusterAD-Data Sample

### 3.6.1 Identification of Nominal Modes

In the first step, instantaneous data samples are converted into vectors in a hyperspace, and then clusters of proximate vectors are identified in the hyperspace. Each cluster represents a frequently observed operational mode in the dataset, which is referred as a nominal mode. These clusters will be used to assess the abnormality of data samples.

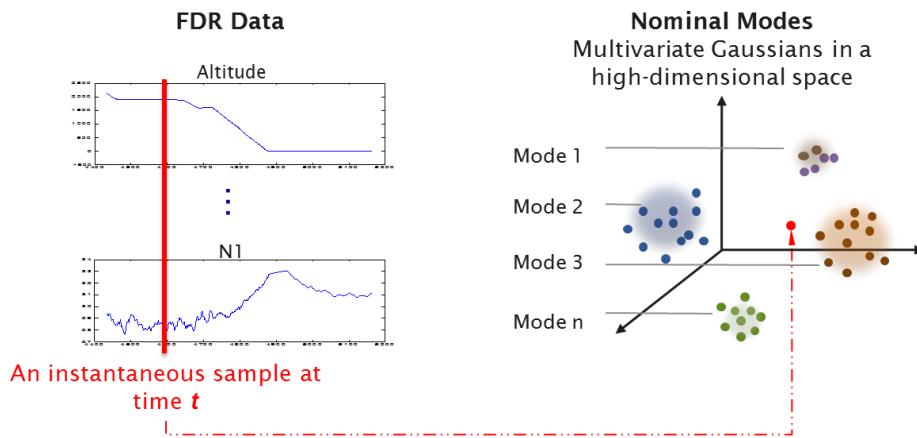


Figure 3.20 Transforming Instantaneous Data Samples to Hyperspace

In detail, ClusterAD-Data Sample maps original FDR data into a hyperspace sample by sample. Each sample of FDR data from flight  $f$  at time  $t$ , distance  $t$ , or other reference index  $t$ , is represented by a vector in the hyperspace, as in the form,

$$\mathbf{x}_t^f = [x_t^1, x_t^2, \dots, x_t^m], \quad (3.1)$$

where  $x_t^m$  is the value of the  $m^{th}$  flight parameter at reference index  $t$ . All samples during a phase of flight for all flights are transformed into the hyperspace. The samples of relatively stable states are naturally clustered together, because the values of flight parameters are determined by the states.

Gaussian Mixture Model (GMM) is then used to identify these clusters. The identified clusters, namely nominal modes, are defined by multivariate Gaussian distributions. A GMM is a parametric probability density function represented as a weighted sum of Gaussian component densities(Dempster et al., 1977; McLachlan & Basford, 1988; Reynolds, 2008). GMMs are commonly used as a parametric model of the probability distribution of continuous measurements or features. Compared to K-means, GMM is able to give statistical inferences on clusters.

A GMM with  $K$  components is given by the equation,

$$p(\mathbf{x}|\lambda) = \sum_{i=1}^K w_i g(\mathbf{x}|\mu_i, \Sigma_i) \quad (3.2)$$

where  $\mathbf{x}$  is a D-dimensional continuous-valued data vector,  $w_i, i=1, \dots, K$ , are the mixture weights, and  $g(\mathbf{x}|\mu_i, \Sigma_i), i=1, \dots, K$ , are the component Gaussian densities. Each component density is a D-variate Gaussian function of the form,

$$g(\mathbf{x}|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2} |\Sigma_i|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \mu_i)' \Sigma_i^{-1} (\mathbf{x} - \mu_i) \right\} \quad (3.3)$$

with mean vector  $\mu_i$  and covariance matrix  $\Sigma_i$ . The mixture weights satisfy the constraint that

$$\sum_{i=1}^K w_i = 1.$$

The GMM components represent nominal modes observed in the dataset. The nominal modes exhibit features of typical operational states of aircraft during a phase of flight. Typical operational states include “ILS approach”, “Flare”, “Touchdown”, “Thrust reverser deployment”,

etc. For example, Figure 3.21 shows a GMM component identified as the operational mode “ILS approach”. All parameter values are normalized in this figure. A few discrete parameters related to ILS approach are labeled by their operational meanings.

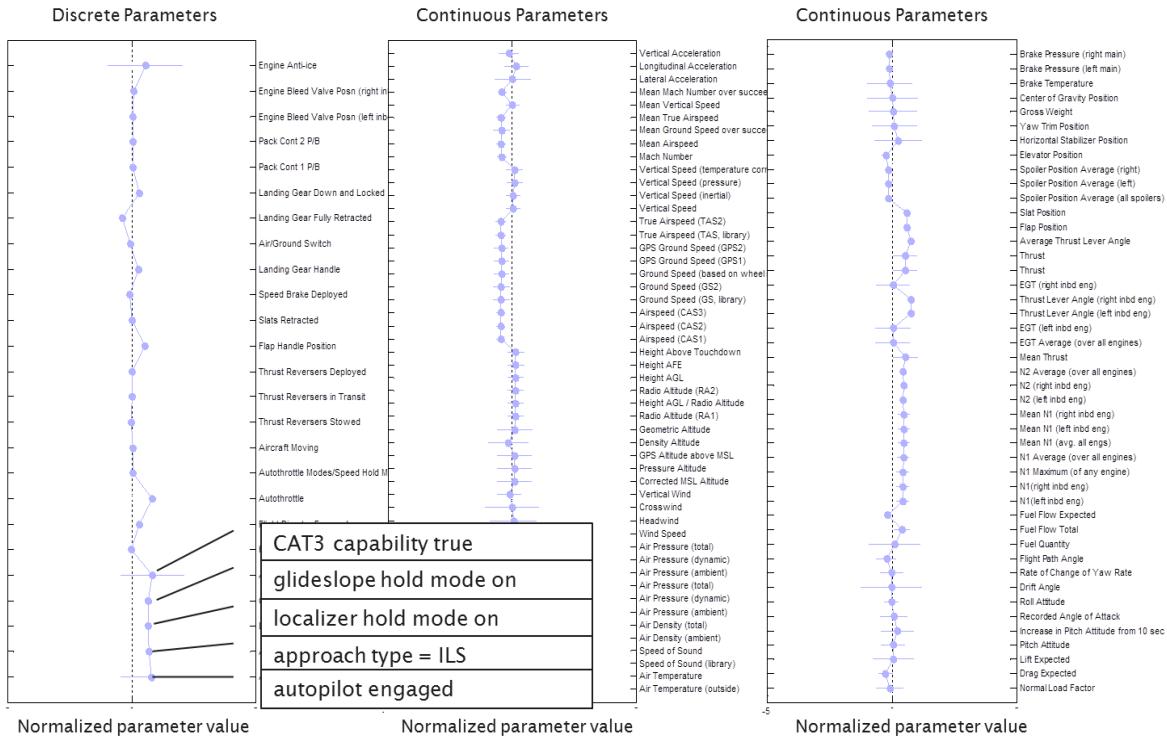


Figure 3.21 An Example of Nominal Mode: ILS Approach

A GMM configuration needs to be specified to build a GMM model, including full or diagonal covariance matrices, shared parameters or not among the Gaussian components and the number of components. These are often determined by the amount of data available for estimating the GMM parameters and how the GMM is used in a particular application.

In this particular application, we choose to use diagonal covariance matrices, independent parameters among Gaussian components, and estimate the number of mixture components (K) by sensitive analysis.

The covariance matrices are restricted to be diagonal in order to reduce computational complexity. Because the component Gaussians are acting together to model the overall vector density, full covariance matrices are not necessary even if flight parameters are not statistically independent. The linear combination of diagonal covariance basis Gaussians is capable of modeling

the correlations between vector elements. The effect of using a set of full covariance matrix Gaussians can be equally obtained by using a larger set of diagonal covariance Gaussians.

The parameters of each Gaussian are not shared among components, because we assume operational modes are independent from others.

The number of mixture components ( $K$ ) is estimated by sensitive analysis. A series of  $K$  values are tested and the optimal one is chosen based on Bayesian Information Criterion (BIC) (Schwarz, 1978). BIC is a measure of the relative goodness of fit of a statistical model. It has been widely used for model identification in time series and linear regression(Abraham & Box, 1979). BIC rewards model accuracy and penalizes model complexity. With the increase of  $K$ , BIC gets lower because of improved goodness of fit when  $K$  is relatively small; with further increase of  $K$ , BIC starts to increase as the penalty of overfitting rises. The value of  $K$  is chosen that gives the minimum BIC.

After a GMM configuration is selected, the parameters of a GMM ( $\lambda = \{w_i, \mu_i, \Sigma_i\}, i = 1, \dots, K$ ) are obtained using the expectation-maximization (EM) algorithm(Dempster et al., 1977). The basic idea is, beginning with an initial model ( $\lambda = \{w_i, \mu_i, \Sigma_i\}, i = 1, \dots, K$ ) to estimate a new model which has a larger likelihood given the training data. The new model then becomes the initial model for the next iteration and the process is repeated until some convergence threshold is reached.

After a GMM is obtained, the probability of a data sample belonging to a nominal mode can be calculated,

$$p(\mathbf{x} \text{ is from nominal mode } q) = g(\mathbf{x} | \mu_q, \Sigma_q) \quad (3.4)$$

$\{\mu_q, \Sigma_q\}$  are the estimated parameters for component  $q$ .

### 3.6.2 Characterization of Mode Distribution

Training a GMM on FDR data provides a parameterized model of nominal modes. However, mode distribution across flight phases is not included in the model. To assess the abnormal level of a data sample, we also need to know which nominal modes are appropriate at a particular time, distance, or other reference index during a flight phase.

In order to know which modes are more likely to be appropriate as a function of time or distance, the number of observations of every nominal mode is counted across different time, distance, or other reference during a phase of flight. Figure 3.22 illustrates the distribution of nominal modes across distance to touchdown during approach phase from a test dataset. The number of observations of a mode is indicated by dot size and color density. This example shows that Mode 33 is the most appropriate mode at the time of touchdown in this dataset.

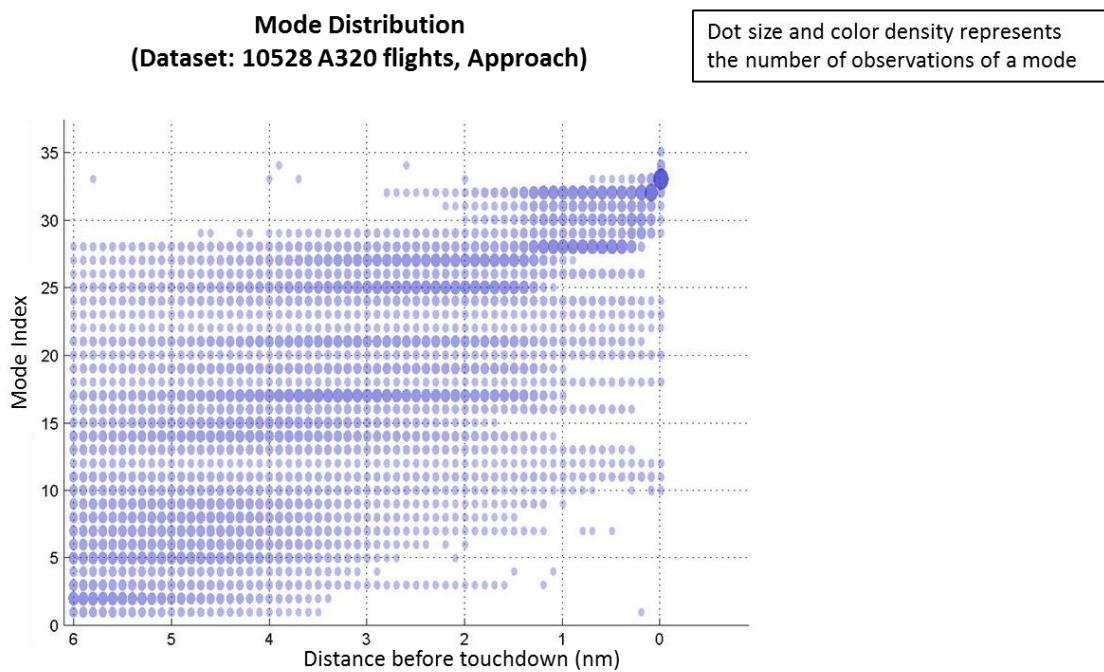


Figure 3.22 Temporal Distribution of Nominal Modes from a Dataset

Based on the mode distribution, which modes are appropriate at a particular reference ( $t$ ) during a flight phase can be calculated. The measure is the probability of a mode at  $t$ ,  $p(\text{nominal mode } q \text{ at } t)$ , which is computed using the following formulas. Given a GMM obtained, the posterior probability of a data sample belonging to a particular component  $q$  is given by the equation,

$$\Pr(q|\mathbf{x}_t^f, \lambda) = \frac{w_q g(\mathbf{x}_t^f | \mu_q, \Sigma_q)}{\sum_{i=1}^K w_i g(\mathbf{x}_t^f | \mu_i, \Sigma_i)} \quad (3.5)$$

where  $\mathbf{x}_t^f$  is a data sample from flight  $f$  at time  $t$ , distance  $t$ , or other reference  $t$ . The probability of a nominal mode  $q$  being proper given all the identified nominal modes, is estimated by aggregating the posterior probabilities of all samples at  $t$ , as given by this formula,

$$p(\text{nominal mode } q \text{ at } t) = \frac{\sum_{f=1}^n \Pr(q|\mathbf{x}_t^f, \lambda)}{\sum_{i=1}^K \sum_{f=1}^n \Pr(i|\mathbf{x}_t^f, \lambda)} \quad (3.6)$$

where  $n$  is the total number of flights in the dataset,  $K$  is the total number of nominal modes.

### 3.6.3 Detecting Anomalies Based on Nominal Modes

In the last step, a probability of being nominal for every data sample is computed, and abnormal flights are detected based on individual samples. The probability of a data sample being nominal is determined by 1) which nominal mode it is likely to belong to, 2) whether that nominal mode should be observed. Mathematically, it is a sum of the probability of a sample belonging to a mode, weighted by the probability of the mode being proper, over all modes, as in the form,

$$\begin{aligned} p(\mathbf{x}_t^f \text{ is nominal}) \\ = \sum_{i=1}^K p(\mathbf{x}_t^f \text{ is from mode } i) \cdot p(\text{mode } i \text{ is proper}) \end{aligned} \quad (3.7)$$

where

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

$$p(\text{mode } q \text{ is proper}) = \frac{\sum_{f=1}^n \Pr(q|\mathbf{x}_t^f, \lambda)}{\sum_{i=1}^K \sum_{f=1}^n \Pr(i|\mathbf{x}_t^f, \lambda)},$$

and  $K$  is the number of nominal modes identified in the dataset. Both  $p(\mathbf{x}_t^f \text{ is from mode } i)$  and  $p(\text{mode } i \text{ is proper})$  are available through calculations in previous steps, as described in

Section 3.6.1 and Section 3.6.2. As a result, a probability profile for every flight can be constructed. Figure 3.23 shows a probability profile of a flight during approach. The x-axis is the distance to touchdown. The y-axis is the logarithm of  $p(\mathbf{x}_t^f \text{ is nominal})$ , - a higher value indicates the data sample being relatively normal comparing to other samples in the dataset; a lower value means relatively abnormal. The probability is displayed in a logarithmic scale because the original data covers a large range of values.

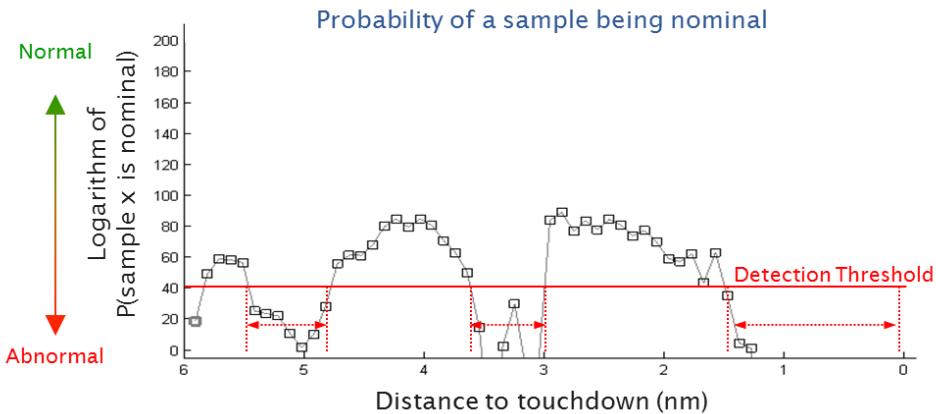


Figure 3.23 Probability Profile of a Flight during Approach

Anomalies are detected by identifying any samples with  $p(\mathbf{x}_t^f \text{ is nominal})$  that is lower than a threshold. A threshold value is set based on the distribution of all samples'  $p(\mathbf{x}_t^f \text{ is nominal})$ . An example is given in Figure 3.24. It shows a distribution of probabilities for all data samples in a dataset with 10528 A430 flights during approach phase. Without assuming the distribution to be Gaussian or other common statistical distribution, a threshold value can be set based on a percentile value, e.g. if threshold is chosen to be the 1<sup>st</sup> percentile value, the top 1% anomalies will be detected.

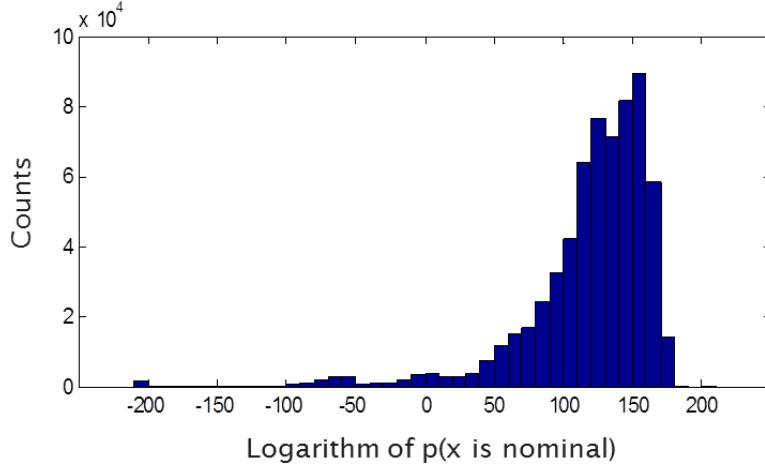


Figure 3.24 Distribution of  $p(\mathbf{x}_t^f \text{ is nominal})$  for All Data Samples

The evaluation of flights is based on the sum of  $p(\mathbf{x}_t^f \text{ is nominal})$  values over all samples for a flight during a flight phase. The sum, which is referred as nominal score, indicates the overall level of a flight being nominal. Abnormal flights are identified as the flights with the lowest nominal scores.

## 3.7 Initial Testing of ClusterAD-Data Sample

This section presents an initial testing of ClusterAD-Data Sample on a large set of data from an airline. This set of FDR data contains 10528 A320 flights. The large dataset is required because the number of observations needs to be much larger than the number of dimensions in the data; clusters might be ill-defined if the density of data is low in the hyperspace.

This section demonstrates the steps in ClusterAD-Data Sample and how algorithm parameters were selected given a new set of data. Examples of anomalies detected are presented in this section.

### 3.7.1 Dataset and Data Preparation

In order to test ClusterAD-Data Sample, a set of FDR data which contains 10528 A320 flights with same engine configurations was used. These flights are operations from a commercial airline in recent years. There are 36 airports involved as either the origin or the destination of these flights.

Each flight's recording has 142 flight parameters (113 continuous ones and 29 discrete ones). To allow flights at different airports to be compared, position related flight parameters were first converted to values relative to the airport. For instance, original recorded altitude values (e.g. pressure altitude, density altitude) were transformed to relative altitudes (e.g. height above takeoff, height above touchdown).

The analysis focused on the approach phase, which was defined as from 6nm before touchdown to touchdown.

### 3.7.2 Optimal Number of Nominal Modes

To construct a Gaussian Mixture Model in ClusterAD-Data Sample, the number of mixture components ( $K$ ) is the only parameter need to be specified; all the others are trained from the data. The mixture components represent nominal modes. A sensitivity analysis of  $K$  was performed in order to select the optimal  $K$ . Different GMMs with a series of  $K$  were trained on the data. Figure 3.25 shows the changes of BIC with  $K$  increases. When  $K$  is smaller than 35, the model fits the data better with the increase of  $K$ ; then BIC starts to increase with further growth of  $K$  because it penalizes the complexity of a model – a tradeoff between goodness of fit and overfitting. Based on a BIC curve shown in Figure 3.25, the optimal  $K$  value is 35.

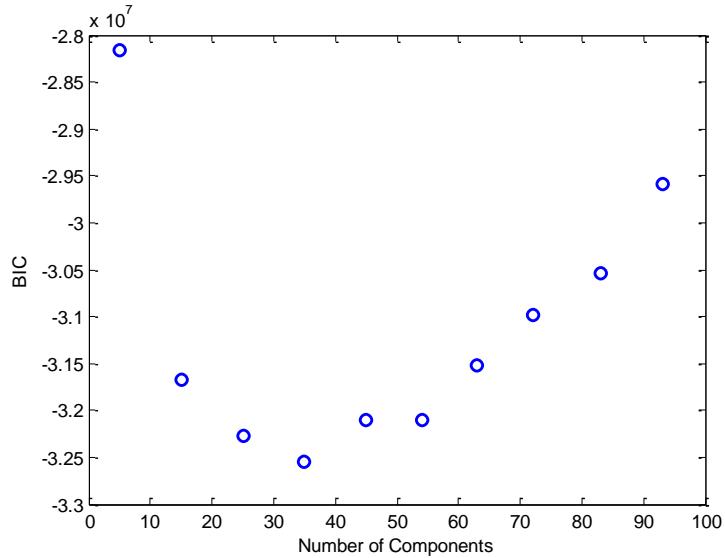


Figure 3.25 Sensitivity Analysis of Number of Components Based on BIC

### 3.7.3 Nominal Modes

A GMM with  $K = 35$  was trained on the data, which identified 35 nominal modes. Each nominal mode was defined by a Gaussian mixture component, which specifies mean values and standard deviations of all flight parameters. Based on the mean values and standard deviations, each Gaussian mixture component was further examined to identify features of typical operational modes during approach phase. A nominal mode identified as “ILS approach” is shown in Figure 3.26. The mean values and standard deviations of flight parameters in this mode are normalized and displayed in this figure. The discrete parameter values indicate this mode is an ILS approach: CAT3 capability = true, Glideslope hold mode = on, Localizer hold mode = on, Approach type = ILS and Autopilot = engaged. In addition, the continuous parameters have values that are normal during an ILS approach, which confirms that this mode is an ILS approach.

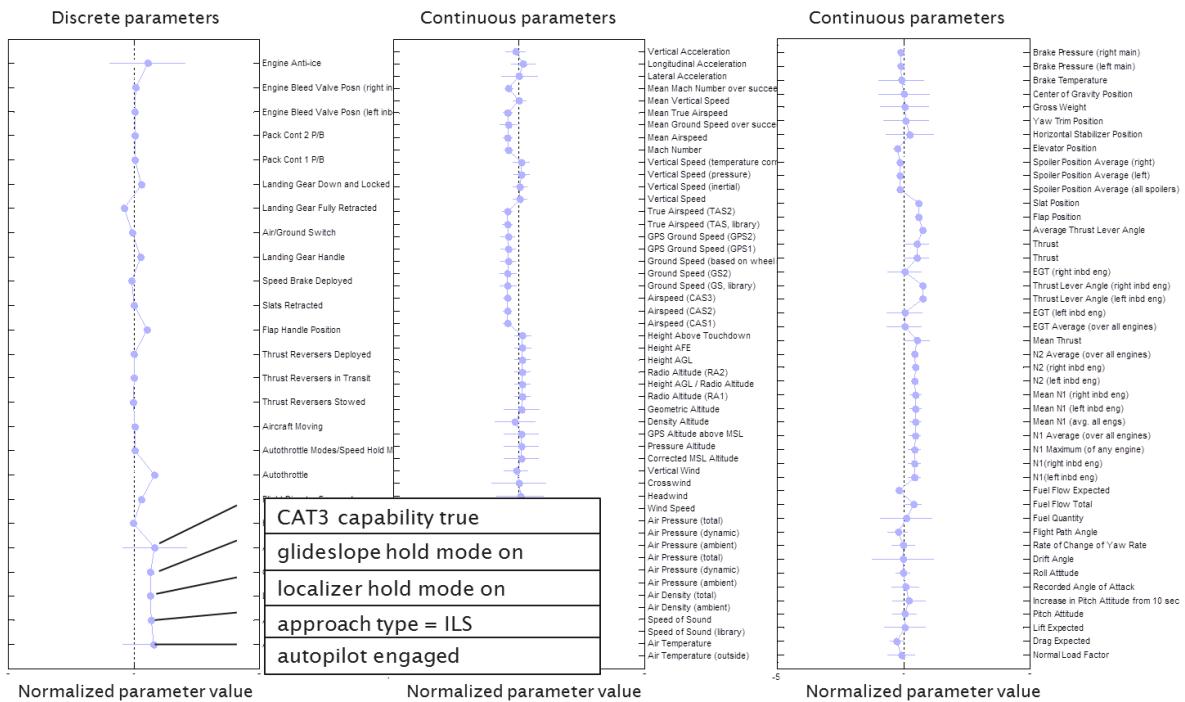


Figure 3.26 Nominal Mode Example: “ILS Approach”

A nominal mode identified as “Thrust reverser deployment” is shown in Figure 3.27. In this mode, discrete parameters show that Air/Ground indicator = ground, Landing gear = down,

Thrust reverser deployed = true, Thrust reverser in transit = true, Thrust reverser stowed = false. Also, the vertical speed related parameters indicate that vertical speed is 0 in this mode.

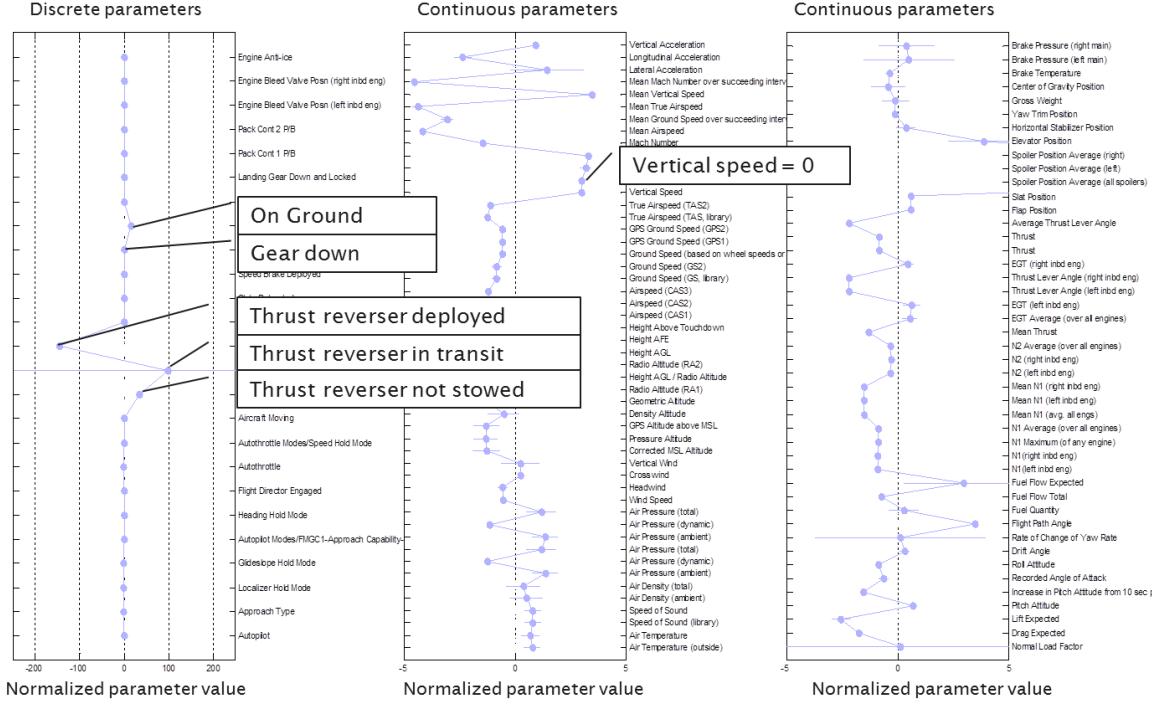


Figure 3.27 Nominal Mode Example: “Thrust Reverser Deployment”

A number of typical operational modes were identified by examining original flight parameter values. These operational modes include “Descent with autopilot on, flap 2, gear up”, “Descent with autopilot on, flap 2, gear down”, “ILS approach”, “Visual approach without flight director”, “Visual approach with flight director”, “Flare with auto-throttle on”, “Flare with auto-throttle off”, “Touchdown”, “Thrust reverser in transit”, “Thrust reverser deployment”, as shown in Figure 3.28.

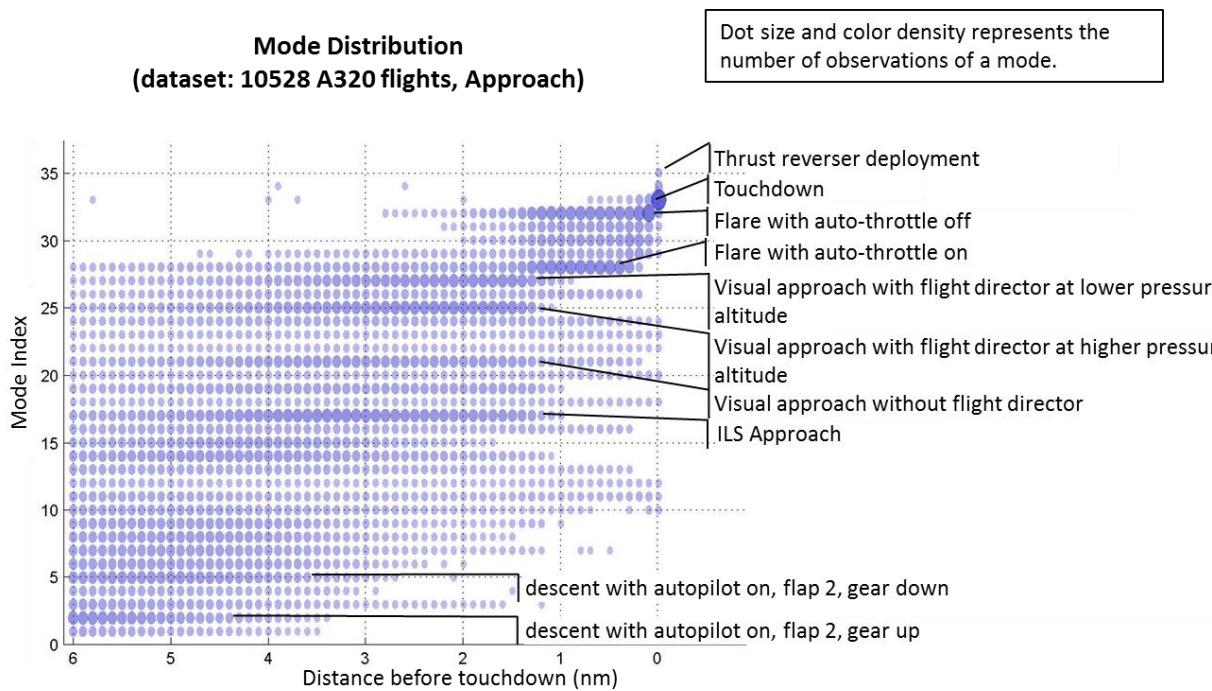


Figure 3.28 Distribution of Nominal Modes during Approach Phase

The distribution of these nominal modes during approach phase is shown in Figure 3.28. Nominal modes representing typical operational states are labeled. Many flights started the approach phase with Mode 2: “Descent with flap 2 and gear up” and Mode 5: “Descent with flap 2 and gear down”. Between 4 nm and 2 nm before touchdown, the most frequently observed mode is Mode 17: “ILS approach.” After the “ILS approach” mode, typical modes include Mode 21: “Visual approach without flight director”, Mode 25: “Visual approach with flight director at lower pressure altitude” and Mode 27: “Visual approach with flight director at higher pressure altitude”. At the end of the approach phase, Mode 33: “Touchdown” has significantly more observations than other modes, as indicated by the large dark blue dot at Mode Index = 33 and Distance to touchdown = 0nm.

### 3.7.4 Anomaly Detection

Whether a sample is abnormal or not is measured by the probability of a sample being nominal,  $p(\mathbf{x}_i^f \text{ is nominal})$  - a higher value indicates a sample is more nominal. The probability is

calculated based on the trained GMM and the distribution of nominal modes, using Equation (3.7).

Abnormal samples, namely instantaneous anomalies, were detected by a detection threshold set based on the distribution of  $p(\mathbf{x}_t^f \text{ is nominal})$  across all data samples in this dataset as shown in Figure 3.29. The distribution has a long left tail, which indicates that detecting low-value outliers is relatively insensitive to the value of detection threshold (Figure 3.30).

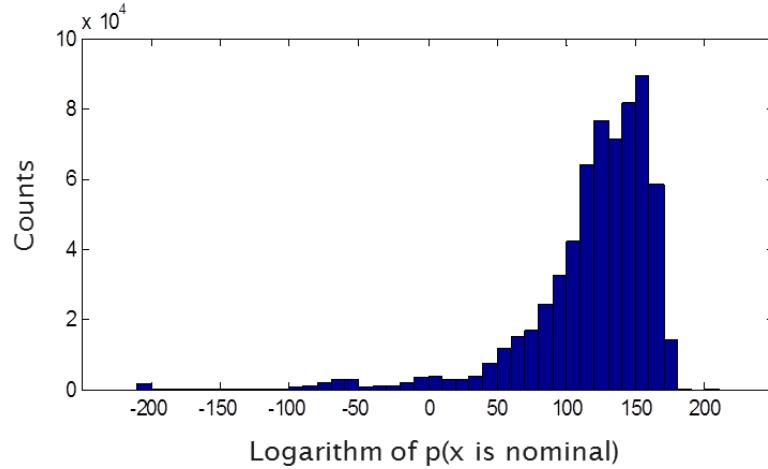


Figure 3.29 Distribution of  $p(\mathbf{x}_t^f \text{ is nominal})$  for All Data Samples

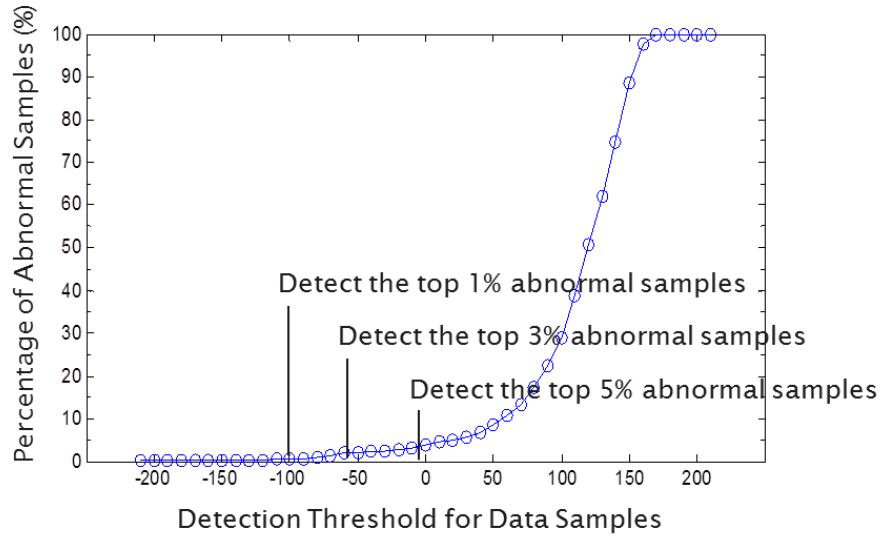


Figure 3.30 Sensitivity Analysis of Detection Threshold for Instantaneous Anomalies

Abnormal flights were detected based on an evaluation on a flight level, calculated by the sum of  $p(\mathbf{x}_t^f \text{ is nominal})$  over all data samples in a flight during a focused flight phase. The distribution of  $p(\text{Flight } f \text{ is normal})$  was heavily left-skewed as shown in Figure 3.31. In order to detection the top 0.5%, 1%, and 3% abnormal flights, the detection threshold could be set as -35000, -5000, and 3000 accordingly.

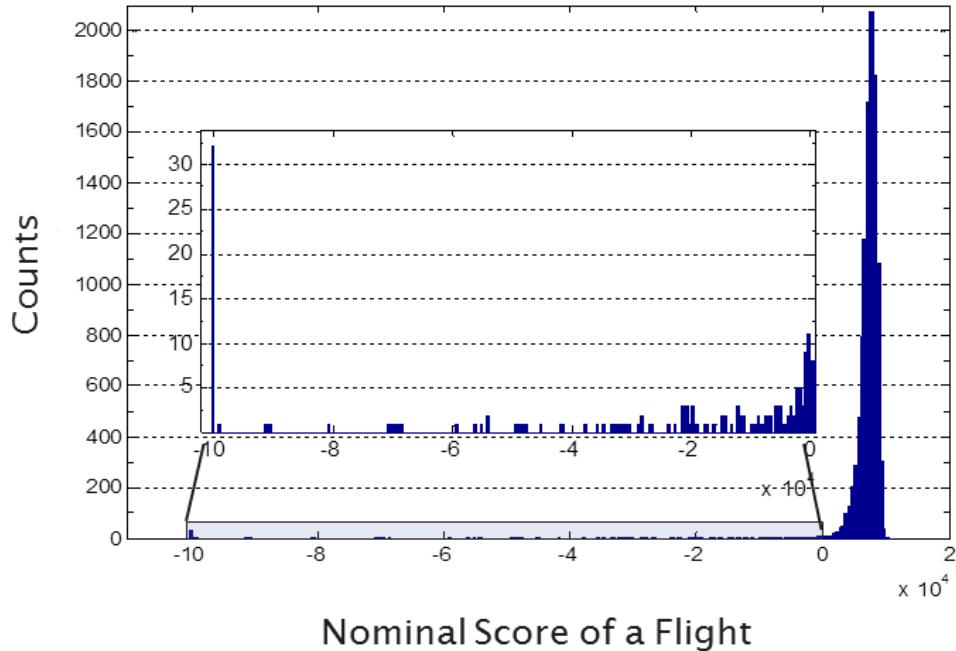


Figure 3.31 Distribution of Flight Nominal Score

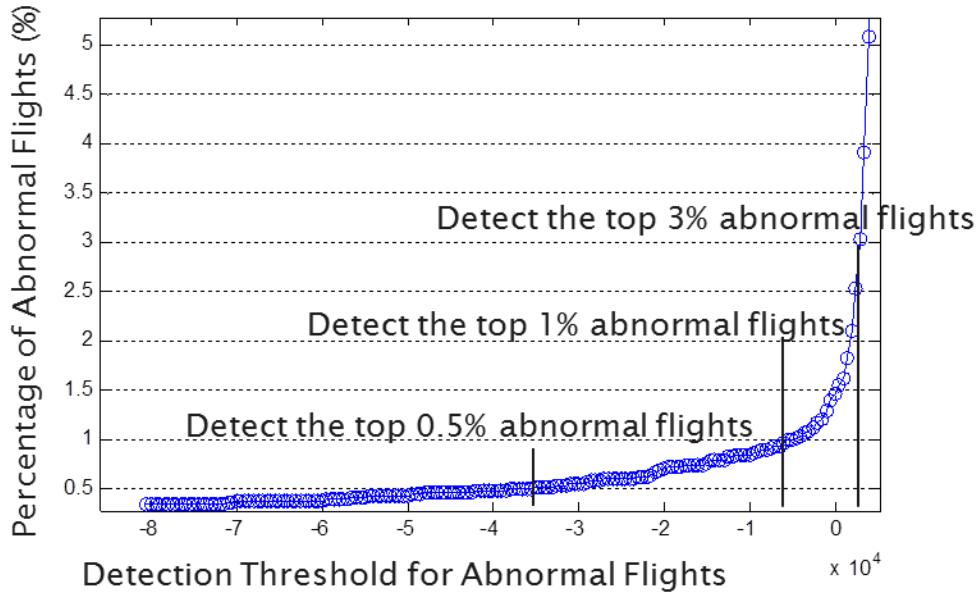


Figure 3.32 Sensitivity Analysis of Detection Threshold for Abnormal Flights

A list of abnormal flights was detected. An evaluation with domain experts was performed to evaluate the operational significance of flights detected. Details will be presented in Chapter 5. This section gives an example of an abnormal flight detected at detection threshold = 1%. The probabilities of data samples of this flight are shown by a red line in FX. The blue region represents the most centered 90% values of all data samples. The flight was abnormal for a short period of time between 3nm before touchdown and 2nm before touchdown. An inspection of the original flight parameters found that parameters related to left engine had abnormal values during that period.

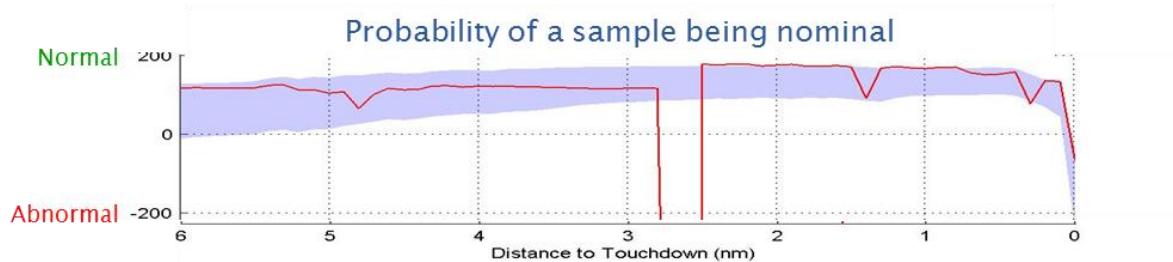


Figure 3.33 Probability Profile of an Abnormal Flight

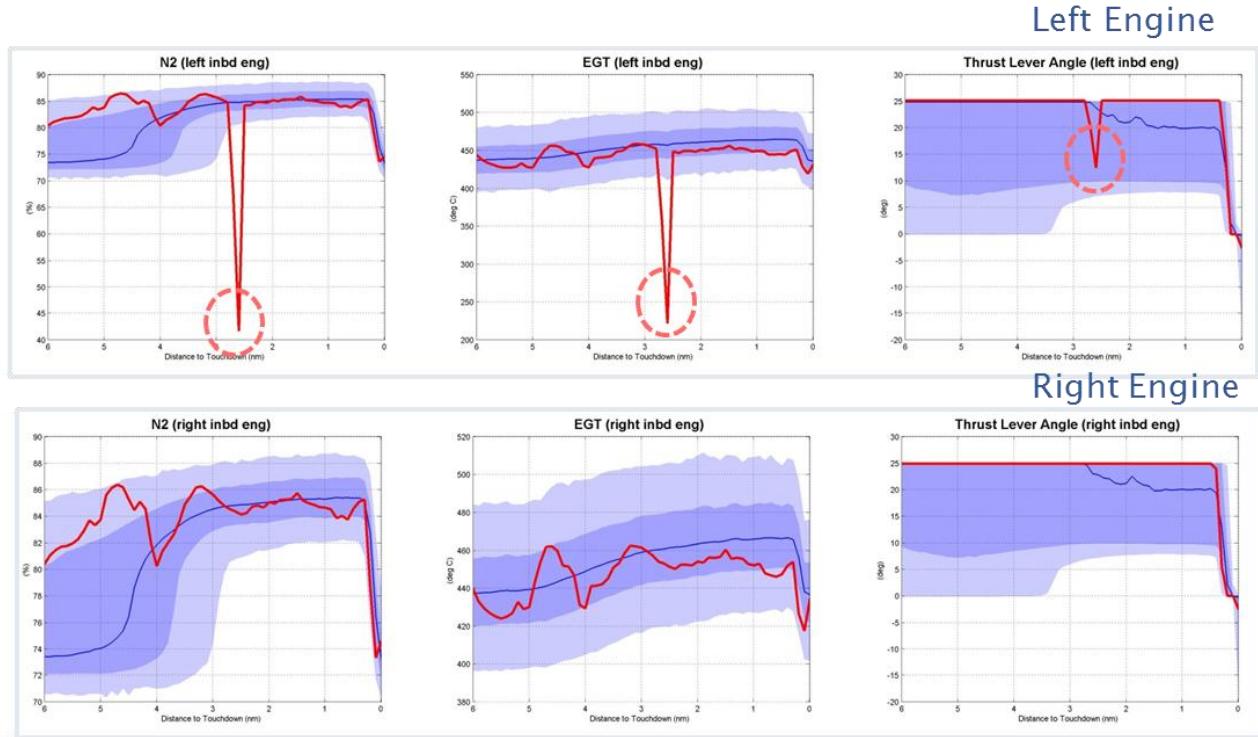


Figure 3.34 Engine Parameters of the Abnormal Flight

## 3.8 Summary

Two anomaly detection algorithms were developed to identify abnormal flights without specifying parameter limits. There are two different techniques to convert raw FDR data into a form applicable for cluster analysis. Both techniques were explored. ClusterAD-Flight converts data of a flight for a specific phase into a hyperspace for cluster analysis. ClusterAD-Data Sample converts data samples into a hyperspace for cluster analysis. This chapter presented the methodologies of two algorithms and demonstrated both algorithms on testing datasets.

# Chapter 4

## Expert Review and Data Representation

This chapter presents a review process of examining abnormal flights detected by anomaly detection algorithms and discusses practical challenges in implementing the review process. In order to address the practical challenges, data visualization tools are developed and tested to facilitate expert review. An experiment with domain experts was conducted to test data visualization tools developed in the thesis. The results showed that the data visualization tools were effective in presenting the information of abnormal flights and locating abnormality across flight parameters and time.

### 4.1 Expert Review Process and Practical Challenges

Domain experts' review is essential to obtain operationally meaningful results from the abnormal flights detected. ClusterAD-Flight and ClusterAD-Data Sample provide a way to identify abnormal flights with atypical data patterns. However, these flights need to be reviewed by domain experts, in order to determine their operational significance, whether they are operationally abnormal or not and whether they are indicating emerging risks or not.

The flow of anomaly detection and expert review process is illustrated in Figure 4.1. Reviewing abnormal flights requires many sources of information available to domain experts, such as data of the flight to be reviewed, normal data patterns, operation standards, weather information, airport information, and local procedures. Domain experts then rate the level of operational significance and identify signs of safety hazards, if any, based on the information presented and their operational experience and knowledge.

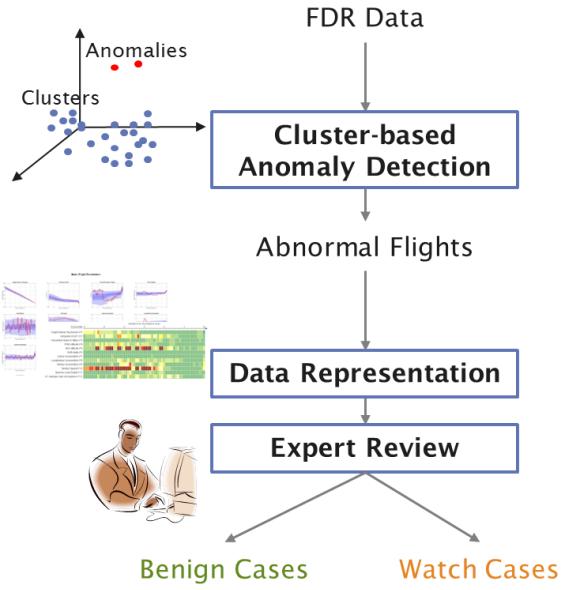


Figure 4.1 Flowchart of Anomaly Detection and Expert Review

Since the review process is performed by flight safety analysts, practical factors need to be considered in the design of the expert review process. It matters how the review process is performed, how the information is presented, and how the results are organized, as it is a subjective evaluation task. Practical challenges are found during initial tests of expert review process.

The workload constraints are the most critical among the practical challenges. The review process is labor intensive, examining tens to hundreds flight parameters for each flight. Figure 4.2 is a mosaic of some flight parameters of a flight, which gives a sense of how large the information is for each flight to be reviewed by a safety analyst. At an early stage of this study, at least 30 minutes were needed on average for a domain expert to review a flight. Each flight had 51 flight parameter plots as shown in Figure 4.2. Without data representation tools to support the review process, it is extremely time consuming to go through each flight parameter and to look for what is abnormal about the flight and indications of safety hazards.

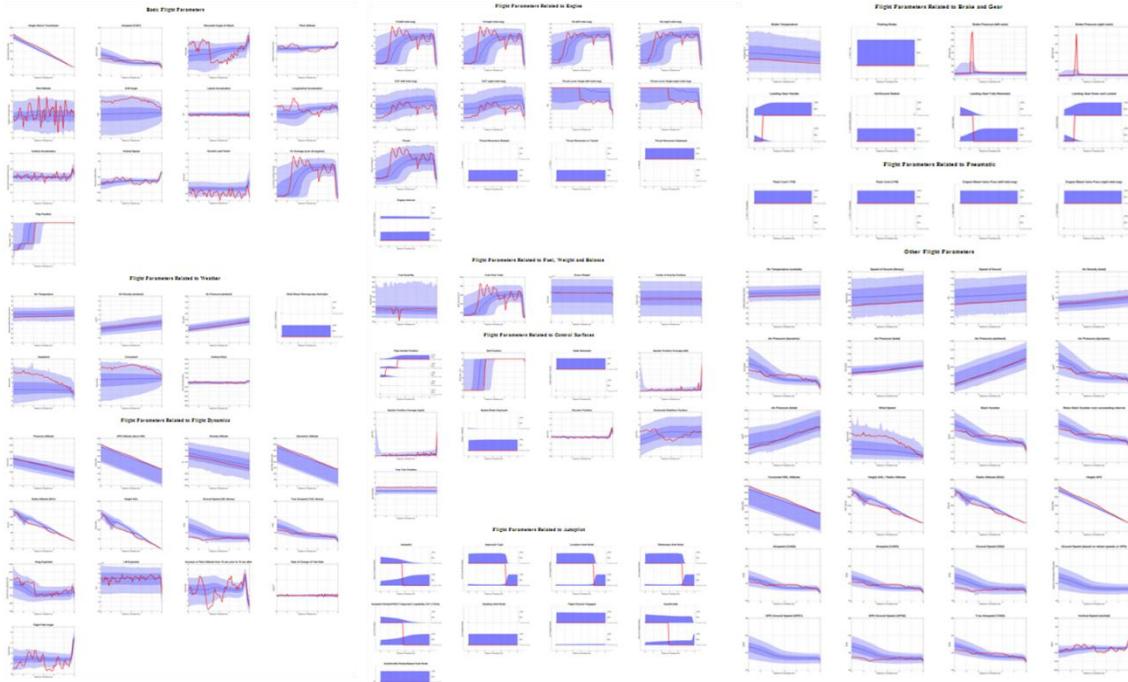


Figure 4.2 Part of Flight Parameters to Review for a Flight

Given the practical challenges, data visualization tools were developed to support the review process with the purpose of maximizing the chance of identifying operational issues and keeping the workload minimal. These visualization tools were designed to present relevant information effectively, make the abnormality directly identifiable, and reduce the time needed to review a flight. Section 4.2 describes the data visualization tools.

## 4.2 Data Visualization Tools for Expert Review

Flights detected by anomaly detection algorithms are referred to domain experts to look for operational significance. A large amount of information about these flights are provided to domain experts, in order to understand what happened during the flight, whether it was operationally abnormal or not, and if it was abnormal, whether the abnormality indicated any kind of emerging risks. Two types of data visualization tools were developed to support the review process. Flight parameter plots were developed to present FDR data of individual flight parameters. Flight abnormality visualization was developed to help analysts quickly identify sources of anomalies across flight parameters and time in a flight phase.

#### 4.2.1 Flight Parameter Plot

An essential part of the review process is to present raw FDR data in graphics. Plots that show raw FDR data of a flight have been used for a long time. The format becomes standardized across software tools: lines to depict time-series of flight parameters over time, as shown in Figure 4.3. However, these plots only show the information of the flight to be reviewed. No reference information is available in the plots, such as normal data patterns, operational standards, weather information, airport information, and local procedures. To determine whether the flight is abnormal or not, analysts need to either use their own knowledge or look up the reference information from other sources. Some efforts have been made to incorporate reference information in time-series plots, for example, a tool developed by Amidan and Ferryman in 2005 gives a performance envelope in the background (Figure 4.4), which is a contour plot that consists of superimposed gray to black boxes displaying the number of flights that shared that value at that time(B. G. Amidan & Ferryman, 2005).

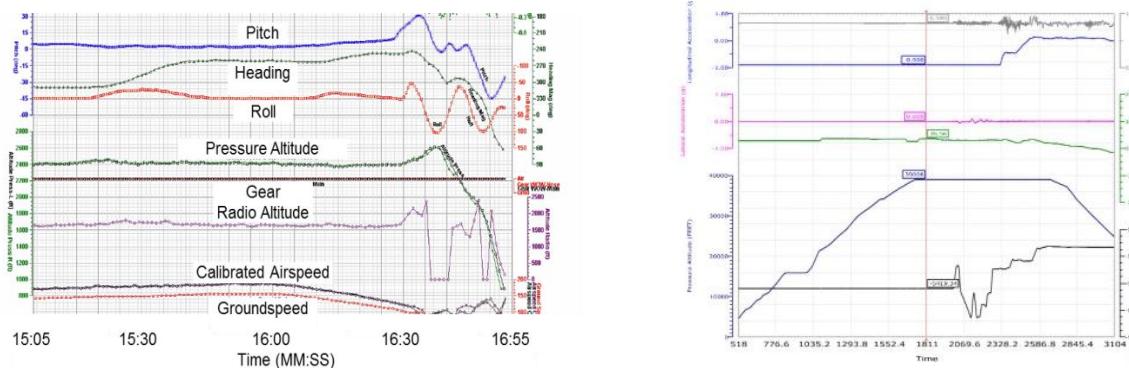


Figure 4.3 Examples of Traditional FDR Time-series Plots

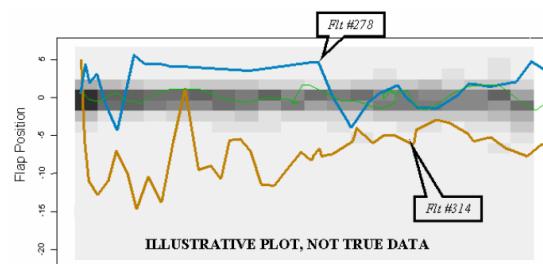


Figure 4.4 Performance Envelope Plot of Flap Position with Three Flight Traces (B. G. Amidan & Ferryman, 2005)

A new design of flight parameter plot is developed in this thesis. Each plot shows the information of a flight parameter with two elements: 1) information of the flight to review; 2) reference information of normal flights. The information of the flight to review is shown in a standard format: a time-series plot of the flight parameter over a period of time and is highlighted in red. The reference information of normal flights is overlaid on the same plot of the flight to review. It shows the normal range of the flight parameter over time, and it is always color-coded in blue. Making the reference information directly available supports domain experts to visually spot abnormalities. Using the same concept, the detailed design slightly differs for continuous flight parameters and discrete parameters to best present information based on data type.

### Flight Parameter Plot for Continuous Parameters

An example of the plot for continuous flight parameters is shown in Figure 4.5. In this plot, the airspeed value of the flight to review is indicated in red. The reference information of normal flights is given by the blue bands. In the lighter blue band, the lower limit is the 5<sup>th</sup> percentile value of all flights, for that flight parameter at that time; the upper limit is the 95<sup>th</sup> percentile value. The darker blue band is plotted in a similar way – the lower limit is the 25<sup>th</sup> percentile value, and the upper limit is the 75<sup>th</sup> percentile value. Therefore, the lighter blue band indicates the center distribution of 90% of the data, and the darker blue band depicts the most centered 50% of the data. The blue bands show the changes in value and variation of a flight parameter over time. By reading the red line reference to blue bands makes the comparison between the flight to review and normal flights straightforward. For example, the airspeed is higher than most flights from 6nm to 3 nm before touchdown in Figure 4.5.

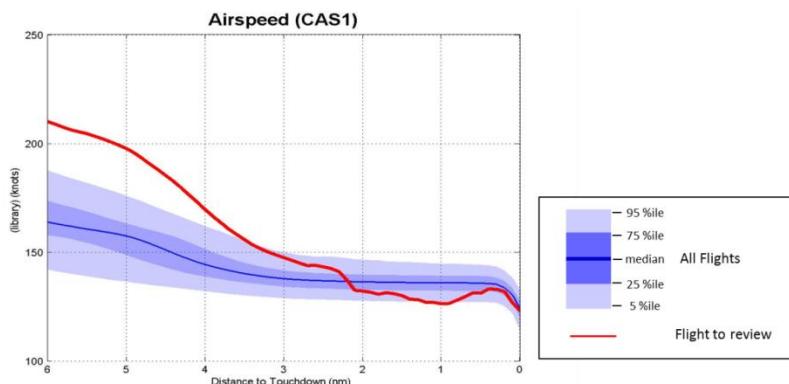


Figure 4.5 Example of Flight Parameter Plot for Continuous Parameters

In this design, the reference information is given by percentiles of all flights, instead of standard deviations. The reason is that percentiles are more robust than standard deviations when applied to real-world data. It is a non-parametric measure, thus, does not require the data fit a normal distribution. In addition, it is not sensitive to outliers with extreme values.

## Computational Techniques to Generate Flight Parameter Plots

Plotting flight parameter plots requires that percentile statistics are generated based on raw FDR data. Traditional computing techniques require all data to be loaded into memory before calculating the statistics. The memory capacity limits the size of data sets that can be processed. Thus, when massive amounts of data need to be processed, one-pass algorithms are preferred, because they scan through observations without storing all of them in the memory. Only necessary statistics are recorded, so the storage requirements are low and fixed regardless of the number of observations. Methods to construct one-pass algorithm are straightforward for parametric statistics, e.g. mean and standard deviation. However, for non-parametric statistics, e.g. median and percentiles, one-pass algorithms use heuristics and other techniques to approximate those statistical measures.

In this study, a one-pass algorithm:  $P^2$  algorithm for dynamic calculation of quantiles and histograms without storing observations (Jain & Chilamtac, 1985) is applied to estimate percentiles of a data set. Instead of storing a complete distribution, it stores only five markers and updates the markers as more observations come in. The five markers are the minimum, the maximum, and the current estimates of  $(p/2)$ -,  $p$ -, and  $(1+p)/2$ -quantiles, to get the  $p$ -quantile of  $n$  observations, as shown in Figure 4.6.

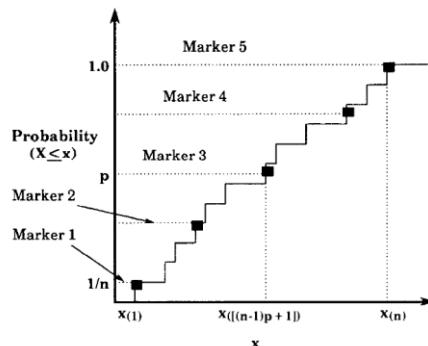


Figure 4.6 The Five Markers in  $P^2$  Algorithm to Estimate a  $p$ -quantile

## Flight Parameter Plot for Discrete Parameters

In the plots of discrete flight parameter, the format is changed to best suit features of discrete variables. If the same plotting format for continuous parameters is used for discrete parameters, the blue bands are not informative, as shown in Figure 4.7, which shows when the autopilot is disconnected. Because the value of a binary flight parameter is either 0 or 1, no matter what percentile to choose, the blue bands either cover across 0 to 1, or shrink to a line on 0 or 1. It cannot effectively present information of when and how many flights switch from one level to another of the discrete parameter. Therefore, the discrete parameters are plotted in a way as shown in Figure 4.8. The red line indicates the parameter value of the flight to review, same as before. However, the reference information is presented differently. Since only a finite number of values exist for discrete parameters, all possible values of the discrete parameter are depicted. Blue areas give the distribution of all flights at different values of the discrete parameter – the height of blue area on a value level indicates the percentage of flights whose parameter value equals to that level at that time.

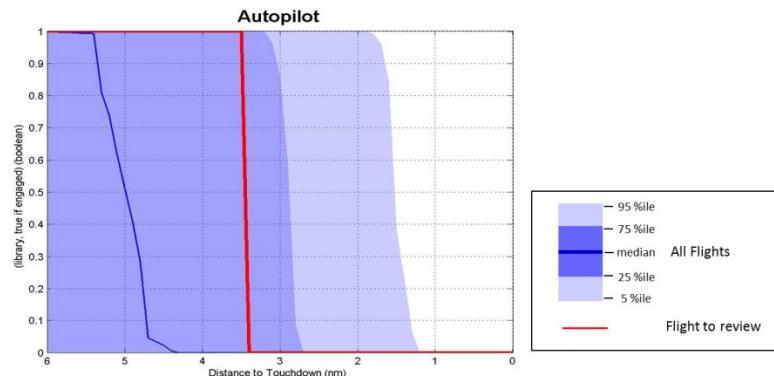


Figure 4.7 Discrete Flight Parameter Plotted in the Format of Continuous Parameter Plot

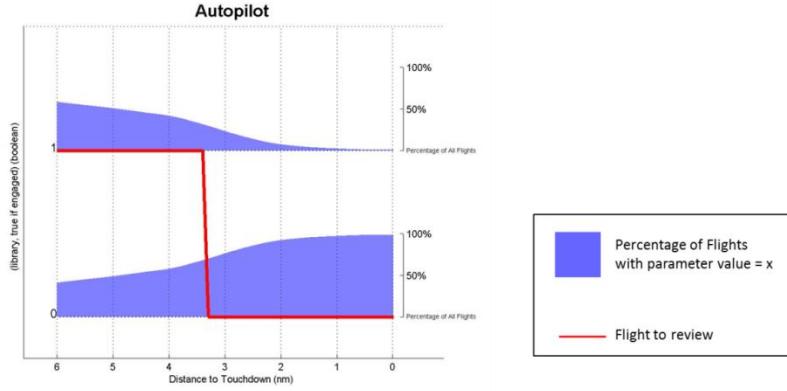


Figure 4.8 Example of Flight Parameter Plot for Discrete Parameters

### Organization of Flight Parameter Plots

Depending on data sets, the total number of flight parameter plots varies from tens to hundreds. It is necessary to organize these parameter plots in an effective way to help analysts to browse through and locate plots of interest. In this visualization tool, users have the option to configure their preferred layout. As an example, Figure 4.9 is a prototype configuration after consulting with domain experts. The plots are arranged into 10 groups: 1) Basic flights parameters, including altitude, airspeed, aircraft attitudes, accelerations, thrust, and flap settings. 2) Flight parameters related to weather. 3) Flight parameters related to flight dynamics. 4) Flight parameters related to engine. 5) Flight parameters related to fuel, weight and balance. 6) Flight parameters related to control surfaces. 7) Flight parameters related to autopilot modes. 8) Flight parameters related to brakes and gear. 9) Flight parameters related to pneumatic. 10) Other flight parameters in the data set.

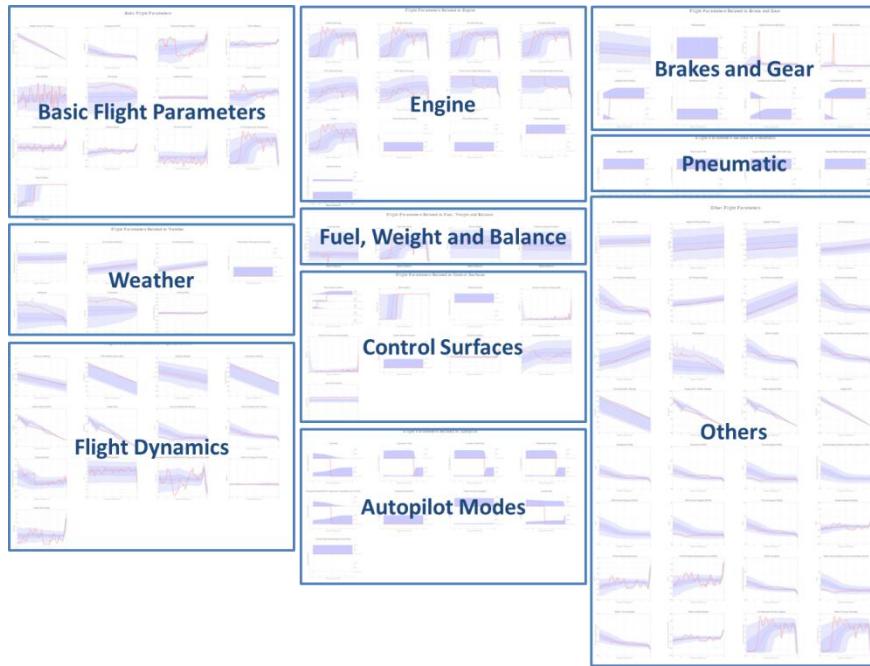


Figure 4.9 A Configuration to Organize Flight Parameter Plots

### 4.2.2 Flight Abnormality Visualization

#### Why is it needed?

The Flight Parameter Plots provide a way to inspect individual flight parameters; however, it is very time-consuming to go through pages of plots to locate what is abnormal about a flight. One option to reduce the time is to present only a few parameters. But which parameters shall be presented? The chance of identifying issues of interest is limited to the presented parameters. The conventional approach of pre-defining a list of important parameters by domain experts is limited, as abnormal behaviors might be hidden in other “not-so-important” flight parameters for different cases.

A new visualization tool is proposed to address this issue, reducing the time needed to review all available flight parameters without pre-defining a limited list of parameters. The data visualization tool, namely Flight Abnormality Visualization, gives an overview of the abnormality level of all available flight parameters across time, distance or other reference in the flight phase of

interest. It enables domain experts to quickly locate which flight parameters and when exhibit abnormal behaviors.

## Method

Flight Abnormality Visualization aims to provide an overview of the abnormality level of a flight across flight parameters and time. Figure 4.10 illustrates the Flight Abnormality Visualization of a flight. The color of each block indicates the level of abnormality for a flight parameter at a particular time. Green means normal, red means abnormal and yellow means in between. All available flight parameters are listed along the y-axis. Temporal reference is provided along the x-axis using the distance to touchdown to be consistent with other plots for the approach phase.

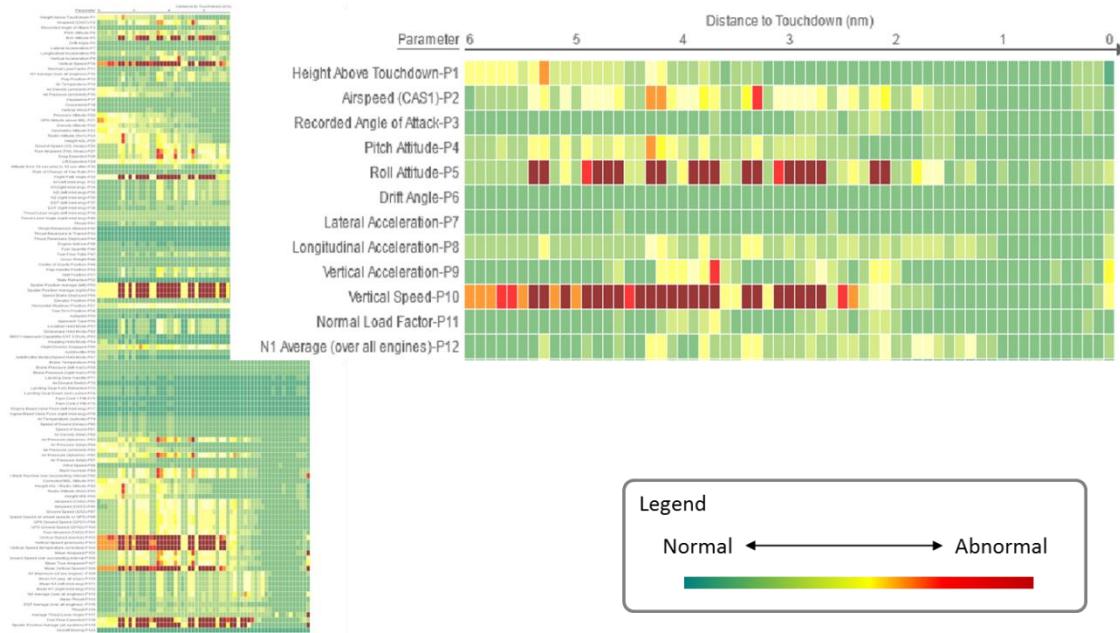


Figure 4.10 Flight Abnormality Visualization of Flight 1264410

The flight shown in Figure 4.10 was a high energy approach, which exhibits abnormal behaviors in a number of flight parameters in the Flight Abnormality Visualization Tool. For example, in the set of basic flight parameters, the most abnormal ones are Roll Attitude and Vertical Speed, which contain a relatively large area of red blocks. Further inspection of Roll Attitude and Vertical Speed confirms the abnormality, as shown in Figure 4.11.

Another feature of the Flight Abnormality Visualization is that each block is directly linked to a flight parameter plot - a user can click on a block of interest to open the plot of corresponding

flight parameter. It gives the user the ability to quickly inspect the original FDR data of individual flight parameters.

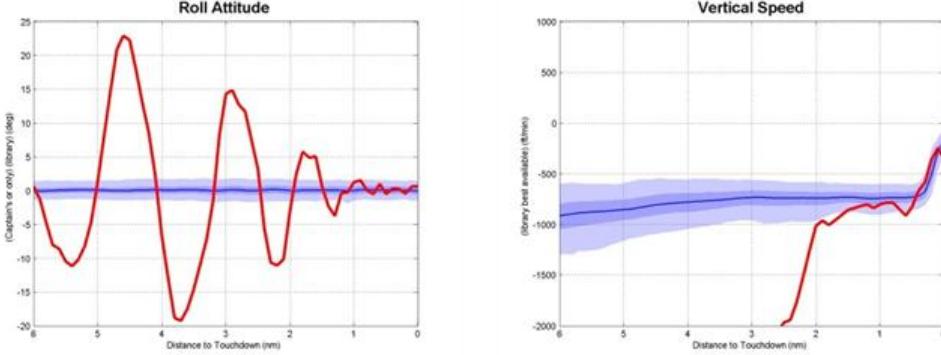


Figure 4.11 Roll Attitude and Vertical Speed of Flight 1264410

An index of abnormality is calculated to determine the color of a block in Flight Abnormality Visualization based on the model trained by ClusterAD-Data Sample algorithm. It measures the probability of a flight parameter value being normal at a particular time, given the distribution of that flight parameter at different times during a flight phase. It is calculated using the probability density function formulated as,

$$p(x_t^p | \lambda^p) = \sum_{i=1}^K w_i g(x_t^p | \mu_i^p, \delta_i^p) \quad (4.1)$$

where  $x_t^p$  is flight parameter  $p$ 's value at time  $t$ , and  $\lambda^p = \{w_i, \mu_i^p, \delta_i^p\}, i=1, \dots, K$  are GMM parameters related to parameter  $p$ . They are extracted from the GMM model trained in ClusterAD-Data Sample.

The index calculated using Equation (4.1) gives a relative sense of the abnormality level comparing to all flight parameters' samples over the phase of flight. Therefore, the color of each block is coded in a relative scale as well. The distribution of the index is mapped into a color scale that varies smoothly from green through yellow to red. A higher value of the index means normal and is colored in green, while a lower value of the index indicates abnormal and is colored in red. The mapping is based on percentiles: index values larger than 50<sup>th</sup> percentile are colored in green, index values smaller than 5<sup>th</sup> percentile are colored in red, and all index values in-between are linearly mapped into a color between green and red.

# **Chapter 5 Evaluation Studies**

## **5.1 Overview of Evaluation Studies**

### **5.1.1 Evaluation Challenges**

The evaluation of the cluster-based anomaly detection approach is challenging due to a number of factors. First, it is difficult to obtain FDR datasets. Because of labor agreements between airlines and pilot unions, the FDR data of routine airline flights are confidential and highly sensitive information in the U.S. and a number of other countries. Second, there is no validation information available against which detection results can be compared. Any abnormal flights detected by the anomaly detection approach are all from airlines' routine flights without incidents, let alone accidents. The safety implications of these abnormal flights are difficult to be measured as it is beyond the capacity of existing methods. Third, it is difficult to compare cluster-based anomaly detection with the current method Exceedance Detection because Exceedance Detection is conducted via proprietary software.

### **5.1.2 Overview of Evaluation Studies**

Ideally, an evaluation of the cluster-based anomaly detection approach would be conducted on a FDR dataset, in which detection results would be compared against a validation standard or the current method Exceedance Detection. Given the challenges described in previous chapter, an ideal evaluation was not possible; instead, a number of evaluation studies were performed using available resources to gain some insights of the cluster-based anomaly detection approach. Two initial tests and three evaluation studies were conducted during different phases of this research, which are summarized in Table 5.1.

Table 5.1 Overview of Evaluation Studies

Testing/Evaluation Study	To Evaluate			Evaluation Method
	ClusterAD-Flight	ClusterAD-Data Sample	Data Visualization Tools	
Initial Testing of ClusterAD-Flight	x			N/A, Preliminary testing
Initial Testing of ClusterAD-Data Sample		x		N/A, Preliminary testing
Evaluation Study I: Comparison of ClusterAD-Flight, MKAD and Exceedance Detection*	x			Comparison with MKAD and Exceedance Detection; Domain expert review
Evaluation Study II: Comparison of ClusterAD-Flight, ClusterAD-Data Sample, MKAD and Exceedance Detection*	x	x		Comparison with MKAD and Exceedance Detection
Evaluation Study III: Evaluation of ClusterAD algorithms and Data Visualization Tools with Domain Experts	x	x	x	Comparison between ClusterAD-Flight and ClusterAD-Data Sample; Domain expert review; Questionnaires
*Collaboration with NASA, restricted data access				

### Initial Testing of ClusterAD-Flight

In order to test if the method works, an initial testing of ClusterAD-Flight was performed using a preliminary FDR dataset obtained from an international airline. The initial testing is presented in Chapter 3 Section 3.5, after the method of ClusterAD-Flight is described.

### Initial Testing of ClusterAD-Data Sample

Similarly, an initial testing of ClusterAD-Data Sample was performed in order to check if the method works. The initial testing is presented in Chapter 3 Section 3.7.

### Evaluation Study I: Comparison of ClusterAD-Flight, MKAD and Exceedance Detection

Since no validation information available against which anomaly detection results can be compared, ClusterAD-Flight was assessed via a comparison with Multiple Kernel Anomaly Detection (MKAD), an anomaly detection algorithm developed at NASA, and Exceedance Detection, baseline method currently in use. The three algorithms were applied on a same set of flight data. The abnormal flights detected by different algorithms were compared to assess the commonalities and differences across ClusterAD-Flight, MKAD and Exceedance Detection.

This comparative study was conducted in collaboration with NASA, who provided a large set of FDR data and its associated exceedance information. However, access to this dataset was restricted. Flights detected could only be reviewed by domain experts within NASA. Special procedures and protocols need to be followed before releasing any information associated with original flight parameters. This study is presented in Section 5.2 of this chapter.

#### **Evaluation Study II: Comparison of ClusterAD-Flight, ClusterAD-Data Sample, MKAD and Exceedance Detection**

In order to examine its performance in detecting exceedances, ClusterAD-Data Sample was applied to the dataset used in Evaluation Study I, which was the only available dataset that had exceedance information. In this study, ClusterAD-Data Sample was compared with ClusterAD-Flight and MKAD on the ability of detecting exceedances which are considered as known issues. Access to the dataset and results were also restricted by the collaboration agreement with NASA. Only aggregated statistical results were available in this study, e.g. number of exceedances, number of flights detected by ClusterAD-Data Sample, and number of commonly detected flights. Detailed information of specific flights was not available. This study is presented in Section 5.3 of this chapter.

#### **Evaluation Study III: Evaluation of ClusterAD algorithms and Data Visualization Tools with Domain Experts**

An overall evaluation study of detection algorithms (ClusterAD-Flight and ClusterAD-Data Sample) and data visualization tools was performed in addition to Evaluation Study I and II. An experiment was designed to demonstrate the cluster-based anomaly detection process and to test the algorithms and data visualization tools. In the experiment, domain experts were asked to review a number of flights detected by ClusterAD algorithms. A number of data visualization tools were used to facilitate the review process. The experiment was designed to obtain domain experts feedback on the data visualization tools and to compare operational characteristics of flights detected by ClusterAD-Flight and ClusterAD-Data Sample. A comparison with the current approach Exceedance Detection was absent due to a lack of exceedance information. This study is presented in Section 5.4 in this chapter.

### 5.1.3 Overview of FDR Datasets

In this thesis, three sets of FDR data were obtained during different phases of this research. A summary of the three datasets is presented in this section.

#### **Dataset I: 2881 flights of 13 aircraft model variants**

It consisted of 2881 flights including 7 aircraft types with 13 model variants, e.g. B777, A319, and A320. The largest set of a particular model included 365 B777 flights. The data were de-identified so all flights were anonymous, and 69 flight parameters were available for each flight. No exceedance information was available for this dataset.

Dataset I was exercised as a preliminary dataset for the initial testing of ClusterAD-Flight.

#### **Dataset II: 25519 A320 flights (via collaboration with NASA)**

Dataset II contained 25519 A320 flights landed at an anonymous airport. Each flight consisted of 367 discrete and continuous parameters sampled at 1 Hz with the average flight length between 2 and 3 hours. Exceedance information were available through a separate analysis by a company specialized in current FOQA analysis working with NASA. However, access to this dataset was restricted. It was available through a collaboration agreement with NASA. Flights detected could only be reviewed by domain experts within NASA. Special procedures and protocols need to be followed before releasing any information associated with original flight parameters.

Dataset II was employed in Evaluation Study I: Comparison of ClusterAD-Flight, MKAD and Exceedance Detection and Evaluation Study II: Comparison of ClusterAD-Flight, ClusterAD-Data Sample, MKAD in Detecting Exceedances, because these two studies were conducted in collaboration with NASA.

#### **Dataset III: 10528 A320 flights**

Dataset III consisted of 10528 A320 flights that originated from or arrived at 36 airports. Each flight has 142 flight parameters sampled at 0.5 Hz or 0.1 Hz depending on the altitude during two flight phases: 1) from takeoff to 10000 ft AGL, 2) from 10000 ft AGL to touchdown. No exceedance information was available for this dataset and data entries were de-identified. The dataset was obtained from an oversea airline with the support of the FAA.

Dataset III was utilized for the initial testing of ClusterAD-Data Sample and Evaluation Study III: Evaluation of ClusterAD algorithms and Data Visualization Tools with Domain Experts. Compared with Dataset II, Dataset III was more convenient to use because it was

obtained and available at MIT. Compared with Dataset I, Dataset III had a larger data size, which should improve cluster analysis performance and make it closer to real-world operations.

Table 5.2 summarizes the three datasets and the associated evaluation studies.

Table 5.2 Overview of FDR Datasets

FDR Dataset	Access	Exceedance Information	Evaluation Studies
I 2881 flights of 13 aircraft model variants (largest subset: 365 B777 flights)	Full	Not available	<ul style="list-style-type: none"> <li>Initial Testing of ClusterAD-Flight</li> </ul>
II 25519 A320 flights	Limited, via collaboration with NASA	Available	<ul style="list-style-type: none"> <li>Evaluation Study I: Comparison of ClusterAD-Flight, MKAD and Exceedance Detection</li> <li>Evaluation Study II: Comparison of ClusterAD-Flight, ClusterAD-Data Sample, MKAD and Exceedance Detection</li> </ul>
III 10528 A320 Flights	Full	Not available	<ul style="list-style-type: none"> <li>Initial Testing of ClusterAD-Data Sample</li> <li>Evaluation Study III: Evaluation of ClusterAD algorithms and Data Visualization Tools with Domain Experts</li> </ul>

## 5.2 Evaluation Study I: Comparison of ClusterAD-Flight, MKAD and Exceedance Detection

### 5.2.1 Background: Evaluation Challenges in Anomaly Detection

A number of efforts have been made to develop algorithms to detect anomalies in sensory data from a complex system. These algorithms build a detection method directly from data collected, rather than based on prior knowledge of the system. Many of the algorithms face the challenge of validating new discoveries from real-world data, e.g. all abnormal flights detected are safe, and the degree of hazards are evaluated differently by different experts.

In order to gain some insights of the performance of different algorithms, we conducted a cross-comparison study of three anomaly detection algorithms: ClusterAD-Flight, Multiple Kernel

Anomaly Detection (MKAD), and Exceedance Detection. The Exceedance Detection method is used as a baseline. It is an analysis tool widely used in airlines' FOQA programs. It detects exceedance events when certain flight parameters exceed pre-specified thresholds. Only known safety concerns are examined by this method. All three methods were independently tested on a same set of flight data from an airline's normal operations.

### 5.2.2 Comparison Design and Setup

The objective of this study was to compare three anomaly detection algorithms, ClusterAD-Flight, MKAD, and Exceedance Detection. The three algorithms were applied on a same set of flight data, Dataset II. The abnormal flights detected by different algorithms were compared to assess the commonalities and differences across ClusterAD-Flight, MKAD and Exceedance Detection.

## Algorithms

This study focused on comparing ClusterAD-Flight with MKAD algorithm, using the traditional method, Exceedance Detection, as a baseline. ClusterAD-Flight and MKAD detect abnormal flights based on a model learned from the flight data, while the traditional method, Exceedance Detection, detects abnormal flights based on domain knowledge, standard operating procedures (SOPs), or other prior domain knowledge.

**ClusterAD-Flight.** It is developed in this thesis. It detects abnormal flights using cluster analysis. It is designed to analyze flight phases that have standard procedures and clear time anchors, such as take-off and final approach. Details of the algorithm can be found in previous sections. ClusterAD-Flight algorithm can handle situations when multiple standard operations exist in the data; even the number of standard operations is unknown. In addition, ClusterAD-Flight tends to work well with continuous flight parameters. However, it is not sensitive to the sequence of the discrete parameters (eg. sequences of switches in the cockpit); the discrete flight parameters are processed in the same way as the continuous ones, only state differences are observable in the algorithm.

**MKAD.** Multiple Kernel Anomaly Detection (MKAD) is an anomaly detection algorithm developed at NASA (Das et al., 2010). It is based on one-class Support Vector Machine (SVM). It

combines information of various data types simultaneously to identify anomalies by the “multiple kernels” approach, e.g. discrete, continuous, text, and network data. A separate kernel is built for each data type. The kernels of different data types are combined into a general kernel. The general kernel maps the original data into a high-dimensional feature space where the data are linearly separable. One-class SVM assumes all normal flights are alike and center on the origin in the feature space. It constructs an optimal hyper plane in the feature space to separate the abnormal flights from the normal ones.

The kernels used in MKAD account both discrete parameters by their sequences, while continuous parameters by their converted sequences, using the normalized Longest Common Subsequence (nLCS) based kernel. Thus, MKAD is able to detect abnormal sequences of switches of cockpit operations.

**Exceedance Detection.** Exceedance detection is the traditional flight data analysis method that is widely used in the airline industry. It consists of checking if particular flight parameters exceed the predefined limits under certain conditions. The list of flight parameters to watch and the limits of those parameters need to be specified by safety specialists in advance. The watch list is always chosen to coincide with the airline’s standard operating procedures, such as the pitch at takeoff, the speed at takeoff climb, the time of flap retraction, etc. Therefore, this approach requires a pre-defined watch list of key parameters under certain operational conditions and, in addition, precisely defined thresholds of the key parameters. Known safety issues can be accurately examined by Exceedance Detection; however, the unknown emerging risks remain latent.

In this study, we leveraged on the results from a standard Exceedance Detection currently used by an airline. The standard Exceedance detection detects three levels of exceedance events. Level 1 indicates minor deviations from the performance target; Level 2 indicates moderate deviations, while Level 3 indicates the severest deviation from the target value.

## Algorithm Settings

All three algorithms can be set at different sensitivity levels for detection. The sensitivity level of ClusterAD-Flight and MKAD is set by the “detection threshold” parameter. It determines how many flights in the dataset will be identified as abnormal. For example, by specifying a detection

threshold equals to 1%, the algorithm detects the top 1% most abnormal flights from the dataset. We compared ClusterAD-Flight and MKAD using a series of detection thresholds to test algorithm performance on different sensitivity levels (detection threshold = 1%, 3%, 5%, 10%). The comparison between ClusterAD-Flight and MKAD was always made on the same detection threshold to make the results comparable.

Exceedance Detection detects “exceedance events” when aircraft performance fails to meet target value ranges during specific maneuvers. The industry normally uses three levels to detect exceedance events. Level 1 indicates minor deviation from the performance target, Level 2 indicates moderate deviation, and Level 3 indicates the severest ones. Because Level 3 exceedance events are the issues of most concerns to airlines, the comparisons with Exceedance Detection were focused on Level 3 in this study.

The algorithm settings are shown in Figure 5.1. Each cube is a scenario with the corresponding algorithm settings tested in the study.

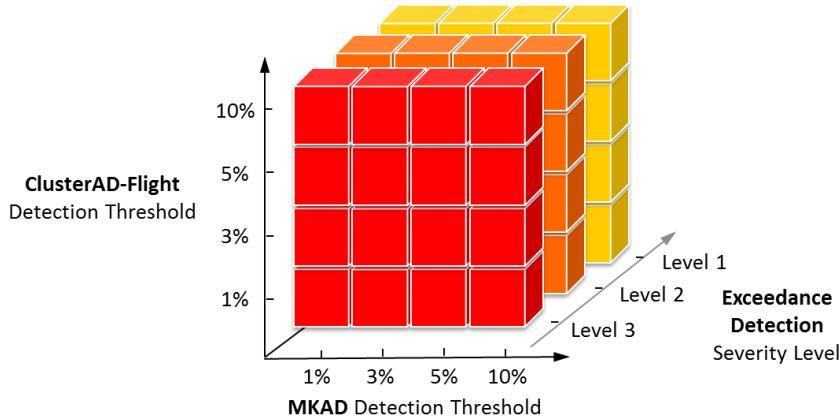


Figure 5.1 Testing Scenarios with Detection Sensitivity Levels

## Dataset

The FDR data used in this study was Dataset II, which was from a commercial passenger airline, including 25519 A320 flights landing at a standard European airport. The analysis focused on the approach phase (from 6nm before touchdown to touchdown).

Each flight consisted of 367 discrete and continuous parameters sampled at 1 Hz with the average flight length between 2 and 3 hours. However, we used a subset of the flight parameters (see Table 5.3) based on domain expert's feedback in order to focus on detecting abnormalities in

crew operations. All flight parameters were analyzed by their original engineering values, except for the flap parameter. It is categorical in nature and has a finite number of values during the approach phase: 10, 15, 20, 40 degree. It was decomposed into 4 binary state variables (shown in Table 5.4) based on information from domain experts, because of data pre-processing requirements in MKAD.

Exceedance information were available through a separate analysis by a company specialized in current FOQA analysis working with NASA. However, access to this dataset was restricted. It was available through a collaboration agreement with NASA.

Table 5.3 List of Flight Parameters Used in the Cross-Comparison Study

Data Type	Flight Parameters
Discrete	Autopilot and all Autopilot related modes, Auto-throttle, Flight Director, Glide Slope, Stall Indicator, Flap Positions (derived parameter), Ground Proximity Warning System, Altitude Mode, Flare Mode, Flight Path Angle Mode.
Continuous	Altitude, Target Air Speed, Computed Air Speed, Engine-related Measures, Pitch Angle, Roll Angle, Rudder Position, Angle of Attack, Aileron Position, Stabilizer Position, Aircraft Gross Weight, Latitude, Longitude and Normal Accelerations, Derived parameters like Above Stall Speed, Vertical Speed.

Table 5.4 Mapping Flap Position into Binary State Variable

Flap Position Original Value (in degree)	Binary States' Values			
	Flap0	Flap1	Flap2	FlapFull
10	1	0	0	0
15	0	1	0	0
20	0	0	1	0
40	0	0	0	1

### 5.2.3 Results Overview

The total number of abnormal flights detected by each method is summarized in Table 5.5 and Table 5.6. As expected, it increases with less restricted detection sensitivity for all three methods. The “detection threshold” controls the number of abnormal flights in ClusterAD-Flight and MKAD - more flights were considered as abnormal when a higher detection threshold was used, as shown in Table 5.5. In Exceedance Detection, the severity level is the main factor impacting the number of flights being detected. As shown in Table 5.6, almost all flights (18888 out of 25519) were found to have at least one Level 1 exceedance event, while only less than 3% flights had at least one Level 3 exceedance event.

Table 5.5 Number of Abnormal Flights Detected by Data-Driven Algorithms

	Detection Threshold			
	1%	3%	5%	10%
ClusterAD-Flight	277	753	1271	2539
MKAD	203	704	1206	2483

Table 5.6 Number of Abnormal Flights Detected by Exceedance Detection

	Level 1	Level 2	Level 3
Exceedance Detection	729	3581	18888

The results of ClusterAD-Flight and MKAD are expected to be different because of their different detection strengths: ClusterAD-Flight tends to work well with continuous flight parameters, while MKAD is able to incorporate the sequence of discrete flight parameters. This is confirmed by the results from comparing flights detected by the two algorithms. Each algorithm detected a set of flights at a detection threshold. A portion of the flights was commonly detected by both algorithms. The number of common flights detected by both varied from 33 (Detection Threshold = 1%), 147 (Detection Threshold = 3%), 355 (Detection Threshold = 5%), to 955 (Detection Threshold = 10%). The agreement between the two methods increased when the detection criteria became more relaxed, as shown in Figure 5.2.

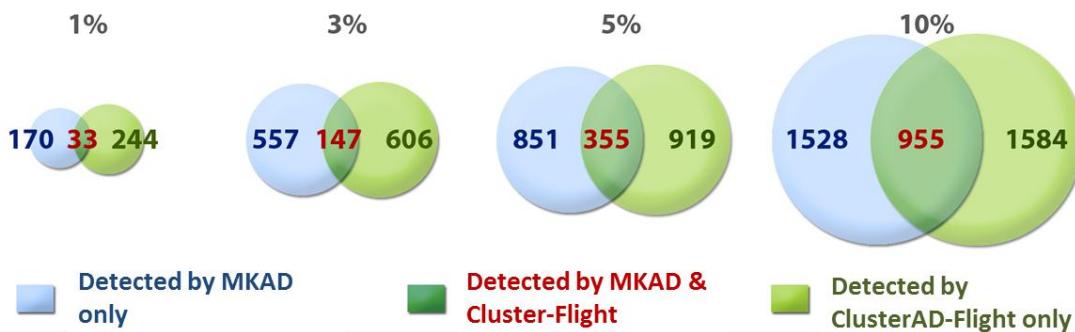


Figure 5.2 Comparison between ClusterAD-Flight and MKAD

A large number of auto-landing flights were detected as abnormal flights in MKAD. Among a total of 568 auto-landings in the dataset, 103 flights (18% of all auto-landings) were considered as abnormal by MKAD at a 10% detection threshold, while only 20 of them were considered abnormal by ClusterAD-Flight, as shown in Table 5.7. It can be explained by 1) MKAD is

sensitive to abnormal sequences in discrete flight parameters; 2) Auto-landings have different activities in discrete flight parameters comparing to ILS approach which is the majority of the approach types in the dataset.

Table 5.7 Landing Type and Abnormal Flights Detected by Data-Driven Algorithms  
(detection threshold = 10%)

	Landing Type				
	ILS Approach	Visual Landing	Auto-landing	Non-precision Landing	Others
All Flights	21960	2987	568	2	2
Flights detected by ClusterAD-Flight	1141	1160	20	2	0
Flights Detected by MKAD	961	1416	103	2	1

### 5.2.4 Comparison between ClusterAD-Flight and MKAD

In order to further understand the strength and the limitations of ClusterAD-Flight and MKAD, we reviewed the detected flights with domain experts in detail and cross-checked with the Exceedance detection results, to look for operational significance. We encountered practical issues during the reviewing process, which is discussed in detail in Chapter 5. We selected three groups from all the flights detected by any method:

- Flights detected by ClusterAD-Flight on all detection thresholds, but not detected by MKAD on any detection threshold
- Flights detected by MKAD on all detection thresholds, but not detected by ClusterAD-Flight on any detection threshold
- Flights detected by both ClusterAD-Flight and MKAD on all detection thresholds

In this section, we present several representative examples in each group. Two types of graphs are used to show the information of a flight: (1) Speed and flap setting during final approach; (2) Time-series plots of most distinctive flight parameters. Regarding to the third type, the same format is used to show all the distinctive flight parameters. The detected flight is shown by black lines. The patterns of most flights are depicted by blue bands. The dark blue bands indicate the 25th to the 75th percentile of all flights data; the light blue bands encompass the 5th to the 95th

percentile. Respectively, the dark blue region contains 50% of the data, while the light blue region covers 90%. The wind plots and latitude & longitude plots in Type 2 graph do not have blue bands because those parameters (wind, latitude, longitude) were not included in the parameter list for anomaly detection algorithms.

## **Operational characteristics of flights detected by ClusterAD-Flight, but not detected by MKAD**

ClusterAD-Flight algorithm detects abnormal flight by grouping normal flights into clusters. The time series of flight parameters are anchored by a specific time (eg. the application of takeoff power during take-off, or the touchdown in final approach). It requires the part of the data being analyzed have specific time anchors to make the comparison.

ClusterAD-Flight tends to perform well with continuous flight parameters and is influenced by the magnitude of deviations. In this section, we present two examples of flights detected by ClusterAD-Flight, yet not by MKAD. Both flights had significant deviations in continuous flight parameters. ClusterAD-Flight can be considered as a variation of Exceedance detection as it works in a similar way in considering flight parameter deviations; however, ClusterAD-Flight automatically inspects all available flight parameters and makes the comparisons based on nominal values summarized from the data itself, rather than pre-specified. In addition, ClusterAD-Flight can handle situations when multiple standard operations exist in the data and the number of standard operations is unknown.

### *Very High Airspeed*

This flight was detected by ClusterAD-Flight at all detection thresholds, yet not detected by MKAD on any detection threshold. It was a very high airspeed ILS approach (Figure 5.3 and Figure 5.4). The airspeed profile was always much higher than the normal airspeed and also than the target airspeed until less than 2 nm before touchdown. As the airspeed was too high, the engine was set to idle until 3 nm before touchdown. Moreover, many flight parameters, e.g. the pitch, the target airspeed, the stabilizer position, the vertical speed, etc., had an abrupt change around 3.5 nm before touchdown. It is not clear what caused this change and why there was a significant drop in pitch even though the airspeed was too high.

This flight was also detected in Exceedance Detection. The detected events were “Speed High in Approach (at 1000ft)” (Level 3), “Speed High in Approach (at 500ft)” (Level 3), “Flaps Late Setting at Landing” (Level 2), and “Deviation below Glideslope (1000ft - 300ft)” (Level 2). In addition, five Level 1 events were also found in this flight.

This example shows that ClusterAD-Flight can detect approaches with excessive airspeed, which is one type of rushed and unstable approach. The rushed and unstable approaches are one of the contributory factors in (Controlled Flight Into Terrain) and other approach-and-landing accidents, because they can result in insufficient time for the flight crew to correctly plan, prepare, and execute a safe approach.

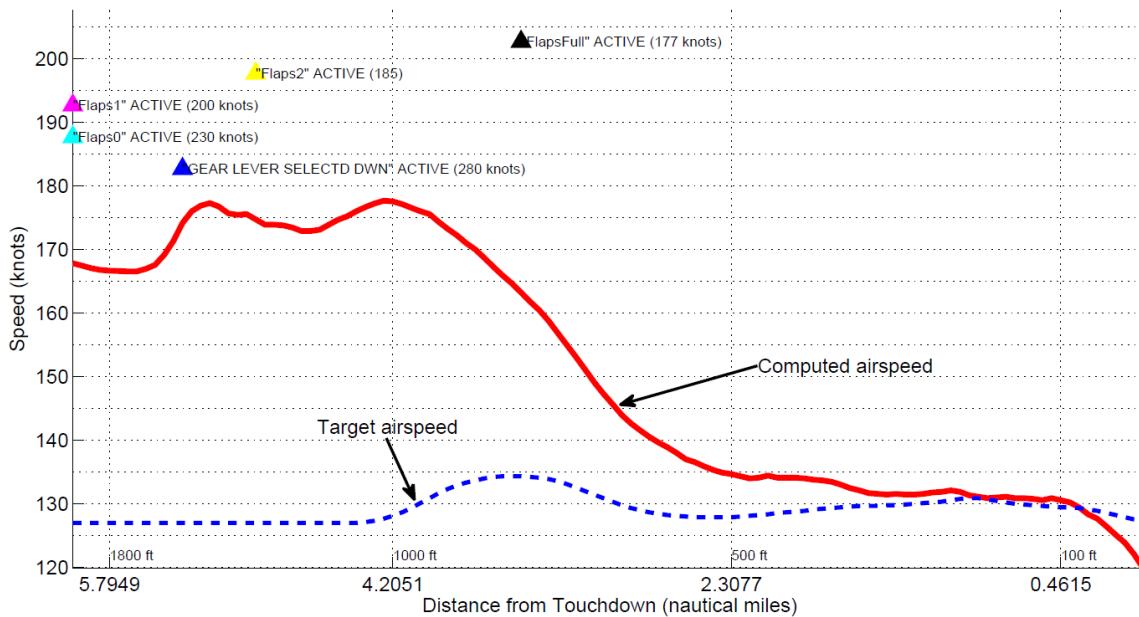


Figure 5.3 Very High Airspeed – Airspeeds, Flaps and Gear

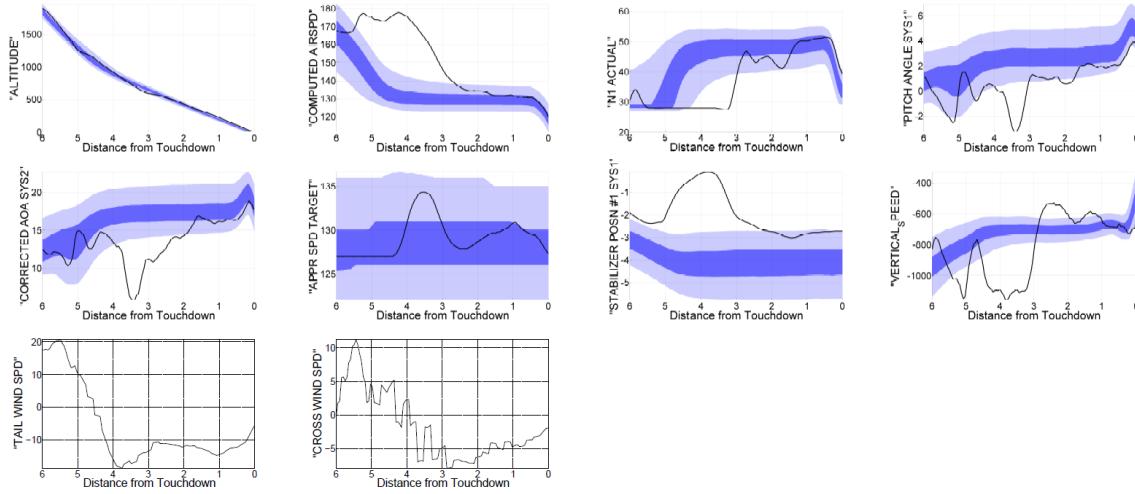


Figure 5.4 Very High Airspeed

### Landing Runway Change

Another type of flight detected by ClusterAD-Flight not by MKAD was flights with landing runway changes in final approach. The flight shown in Figure 5.5 is a representative example of this type. The flight was originally lined up for the right runway. Then, it turned and landed on the left runway. Although the ground reference position information (e.g. latitude and longitude) was not included in the anomaly detection analysis by ClusterAD-Flight and MKAD, ClusterAD-Flight was able to capture the abnormal behaviors in other flight parameters caused by the change of runway turn and identified the flight as abnormal on all detection thresholds.

None of these abnormal behaviors was considered severe by the standard in Exceedance Detection. No Level 3 or Level 2 events were detected in the flight for this flight phase. Only four Level 1 events were detected. This type of abnormal flight could be operationally benign, because the change of landing runway happens due to many reasons, e.g. ATC assignment to accommodate traffic flows, ILS instrument limitations, etc. However, to identify this type of abnormal operations and then to track the trend can help to understand the operations better, such as whether it happens at a particular airport, during a specific time of the day, or under certain weather conditions. Moreover, further analysis may bring insights on whether there is a correlation between the approaches with runway change and the unstable approaches.

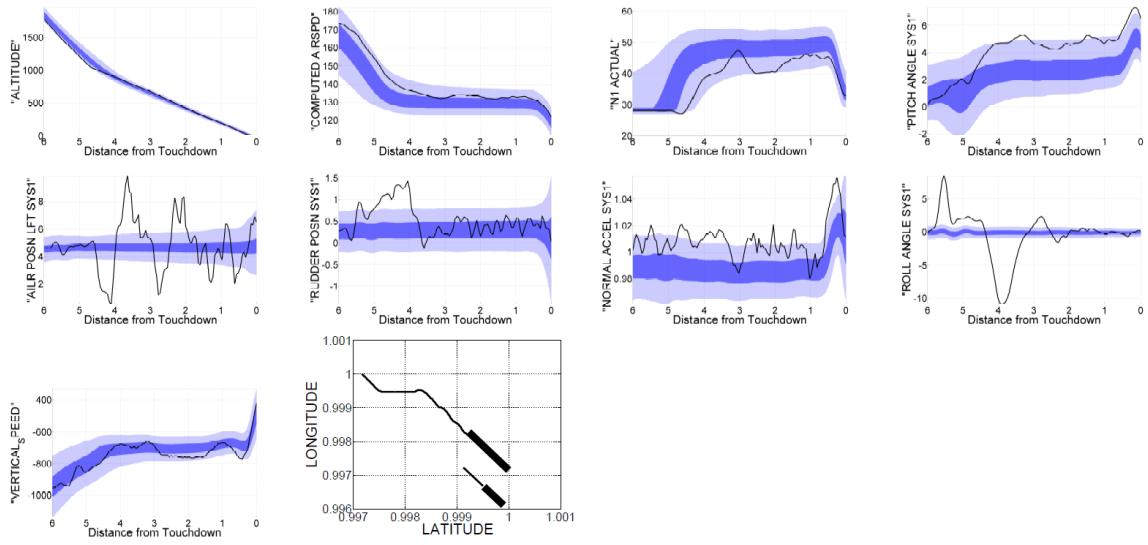


Figure 5.5 Landing Runway Change

## **Operational characteristics of flights detected by MKAD, but not detected by ClusterAD-Flight**

MKAD models discrete sequences and continuous sequences for a given process using a normalized Longest Common Subsequence (nLCS) based kernel, which captures sequential features. Therefore it is able to detect some events that ClusterAD-Flight is not designed to detect.

### *Unusual Auto Landing Configuration*

In this section we describe two flights identified by MKAD mostly due to some atypical patterns in the switching sequences generated from the discrete parameters. In both cases the flights used auto-land systems which have been designed to control the aircraft automatically during approach and landing. In this data set we had a small fraction (around 2% of the 25519 flights) of auto-landing examples and the deviations from normal switching behaviors for auto-landing made them statistically significant as compared to rest of the flights. Auto-landing is mostly used in poor visibility conditions and/or bad weather where the crew can see the runway lights just few seconds before landing. In many such occasions with poor visibility conditions, visual landing may not be possible or may be considered unsafe and therefore auto-landing is preferred. The presence of automatic guidance systems with human in-the-loop makes the auto-land an extremely accurate and safe maneuver. However there are strict requirements which are imposed by the authorities on airborne elements and ground environment, as well as special crew qualification for auto-landing.

The first flight engaged auto-landing without full flap setting. Under normal circumstances the auto-land is executed with both autopilots engaged and with flaps configured full. The use here of the flap setting prior to full introduced some differences from usual auto-land patterns. Out of all auto-landing examples, more than 90% of flights performed this operation with full flap settings.

While legal from an operational stand point this was still reported by the algorithm due to some statistically significant activities (or signatures) in parameters like autopilot and autopilot modes and flight directors. The exceedance based method indicated “Pitch High at Touchdown”, “Short Flare Time” and “Short Flare Distance” with considerable severities (Level 2 and Level 3).

The second example was another atypical auto-land configuration. In this flight the flaps were configured full. The weather was reported as foggy with 0.1 mi visibility. It is common to use only

one autopilot for an approach which does not require an auto-land. The autopilot is then disengaged when the runway is in sight or at least by the minimum charted descent altitude. Procedures specify that an approach which requires an auto-land, however, must be started with two autopilots. If one autopilot then fails, a “fail-operational state” exists and the automatic landing can be completed. This flight departed from normal operational requirements by utilizing only one autopilot for the entire approach and auto-land. This scenario could be of interest because under bad weather conditions (like extremely low visibility conditions) with further degradation of the system the auto-landing capability may be lost at an extremely inopportune time. The algorithm was able to find it due to the uncharacteristic settings of autopilots for landing aircraft. For example, out of all auto-landing examples, only 2 flights with different tail numbers performed this kind of operation. The exceedance based method indicated “Speed Low at Touch down” and “Flaps Questionable Setting at Landing” with considerable severities (Level 2 and Level 3).

## **Operational characteristics of flights detected by both ClusterAD-Flight and MKAD**

Common flights detected by both ClusterAD-Flight and MKAD are those that have significantly different data patterns from other flights. A number of flight parameters of these flights are distinctively different compared to the nominal values of most flights. Some of the common flights indicate interesting operational implications, while some of them are benign as they are abnormal due to the low occurrence of the operation. Four examples of the common flights are presented in detail: High energy approach, High-airspeed with low-power approach, Recycling flight director, and Influence of wind.

### ***High Energy Approach***

The high energy approach was a flight that is detected by both ClusterAD-Flight and MKAD on all detection thresholds. There are two basic conditions which may result in a high energy approach - the aircraft may be too high or too fast or both. This flight has been categorized as a high energy approach with unusually high air speed when compared to a set of reference flights landing at that airport.

The flight review with domain expert suggested that this flight might indicate an energy state awareness problem. The speed profile and power profile of this flight could be precursors to runway excursion for shorter runways. The high speed was not due to the wind. Ideally, in these cases it should be a go around. The landing operation was performed in a cloudy weather condition with average visibility of 8.2 miles and with almost no wind. "Flap 0", "Flap 1" and "Flap 2" along with the landing gear was deployed before 1800 ft (or 6 nm from touch down). During this process, a gradual turn was initiated to align with the runway in preparation for landing. This flight intercepted the glide slope (see altitude plot in Figure 5.6) from below and was slower than most other flights at the beginning of the approach. During this period the pitch was high. Immediately after this the pilot spooled up the engines for some time to increase the speed. This was followed by lowering the pitch with engine idle which further accelerated the aircraft. The target airspeed (140 knots) was higher than most others (126-130 knots). The power was high until 3 nm before touchdown, which resulted in a high and unstable airspeed and a significant decrease in N1 for the rest of the approach. In addition, the pitch angle profile and altitude profile

also showed signs of unstable approach. At 500 ft the pilot pulled the nose up slightly early to ensure rapid deceleration. The effect of that can be clearly seen in normal acceleration and vertical speed profiles. The anomaly detection algorithms found this flight atypical due to the combined effect of the deviations of several continuous parameters.

This is an interesting example that can be detected by ClusterAD-Flight and MKAD, but may be overlooked by the Exceedance detection method. The exceedance based approach reported Level 1 exceedance events, namely “Speed High in Approach (at 50 ft)”, “Pitch High at Touchdown”, “Path Low in Approach (at 1200 ft)”, “Long Flare Time” and “Approach fast 500 RAD”. The findings of the anomaly detection methods provide a clear picture on the unusual energy management scenario, however the exceedance detection method conclude it as a normal flight.

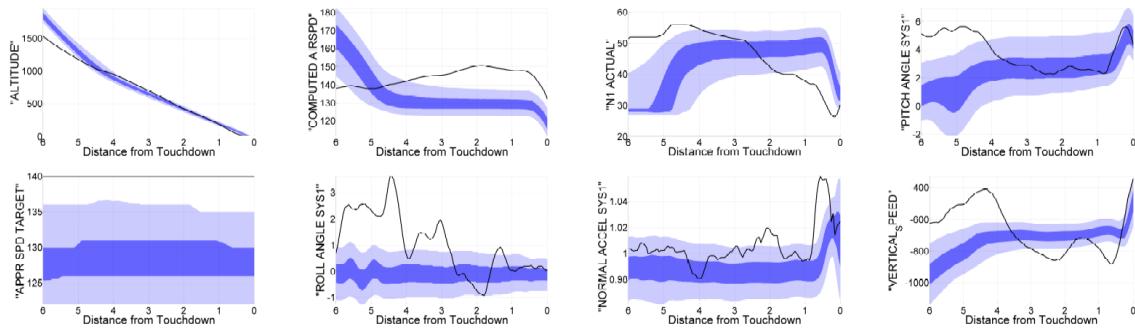


Figure 5.6 High Energy Approach

### *High-Airspeed with Low-Power Approach*

One type of abnormal approaches that can be detected by both ClusterAD-Flight and MKAD was the high-airspeed and low-power approaches. Figure 5.7 is an example of this style of approach. It was a visual landing. The airspeed was always high and the engine was set to idle until 1 nm before touchdown. Procedure calls for the engines to be spooled up for the entire final approach so that instantaneous power adjustments can be made. Other flight parameters also show abnormal patterns compared to the patterns in the majority of flights. For example, the altitude profile was above the normal altitude profile from 5 nm to 1 nm before touchdown, the pitch was relatively low until 2 nm before touchdown; the roll angle had a significant amount of activity at the beginning of the final approach, etc.

Although this flight landed safely, this type of approach is not recommended. The test on this dataset shows that both ClusterAD-Flight and MKAD can catch this type of anomaly. Moreover, Exceedance detection confirmed that this type of anomaly was operationally significant. The exceedance detection identified four Level 3 events, one Level 2 event, and four Level 1 events in the approach part for this flight. The Level 3 events were “Speed High in Approach (at 1000ft)”, “Speed High in Approach (at 500ft)”, “Low Power on Approach” and “Approach Fast 500 RAD”. The Level 2 exceedance was “Pitch High at Touchdown”.

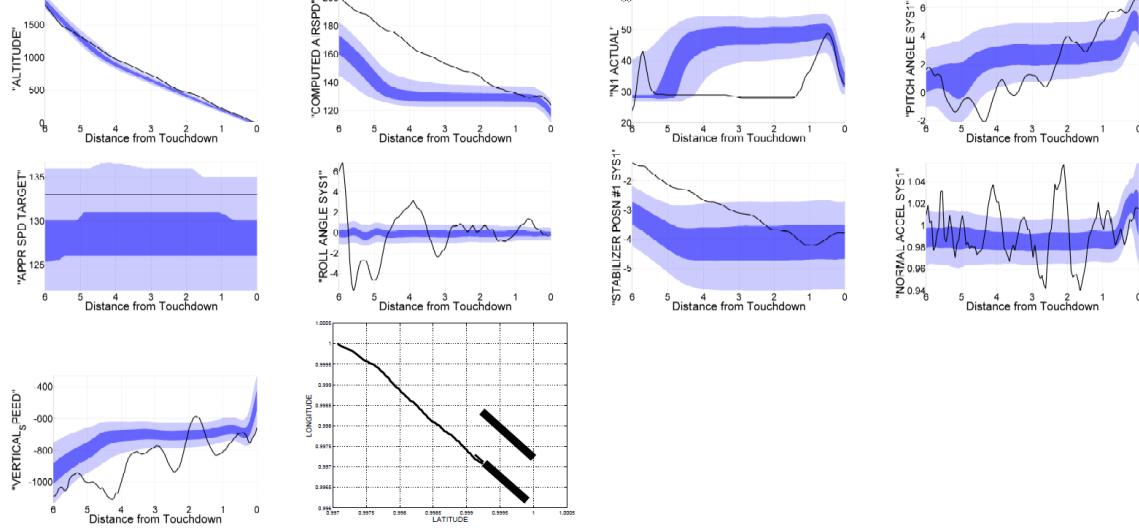


Figure 5.7 High-Airspeed with Low Power

### *Recycling Flight Director*

The main contribution toward the abnormality of this flight came from an atypical event in discrete parameters as a result of change in runway and another error of commission. The first event was related to automation disconnection. This flight was completely hand flown and was initially configured for the right runway. Once the new runway was assigned and the required maneuvering was initiated to align with the left runway, the crew had to recycle the flight directors in order to get to the default modes of Heading and Vertical Speed. The second event was related to mode switching and we are unable to reach any hypothesis on why the pilot would take such an action. The transition from “vertical speed mode” to “open climb mode” around 1500 ft was an inappropriate move by the pilot as the missed approach altitude has already been set and the “open climb mode” will spool up the engines in order to climb to that altitude right away.

However it is clear from Figure 5.9 that the pilot immediately corrected this mistake and continued to hand fly the aircraft appropriately. A level 2 exceedance in “Speed High in Approach (at 1000ft)” was reported, but may have been unrelated to all the forgoing actions.

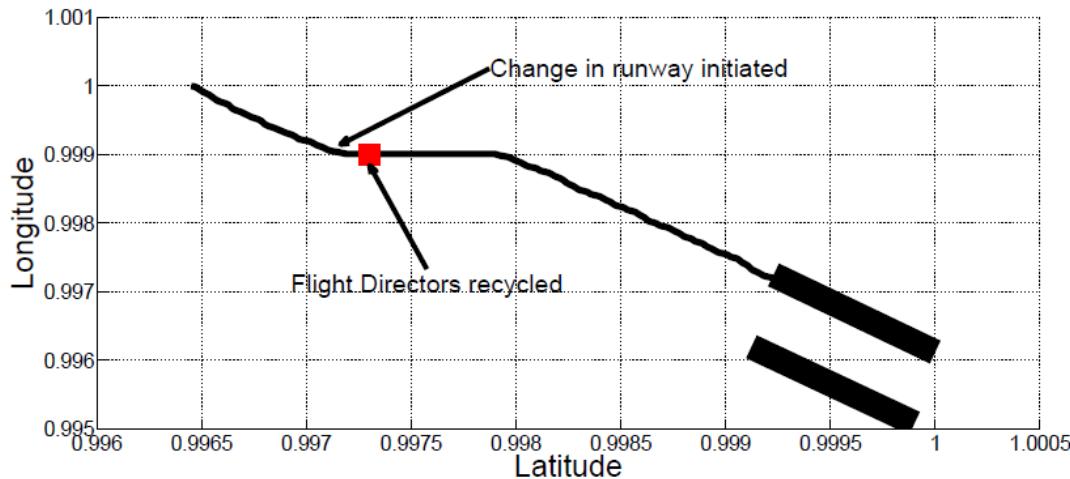


Figure 5.8 Recycling Flight Director – Change in Runway

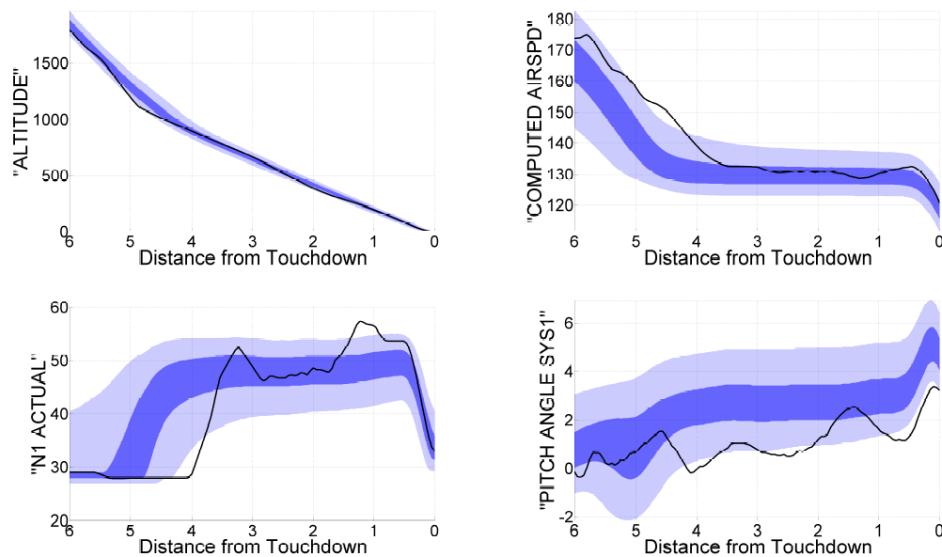


Figure 5.9 Recycling Flight Director

### *Influence of Wind*

Both ClusterAD-Flight and MKAD found this flight (Figure 5.10) as abnormal because of deviations in multiple continuous parameters, combined with various mode transitions. Some of these deviations were not immediately obvious when examining individual parameter plots, however they combined to create an atypical flight. The discrete parameter values showed that

one autopilot (AP 1) was used and later the pilot disconnected the autopilots and proceeded manually with the rest of the approach and landing. Moreover both the flight directors were recycled immediately after that. The first hypothesis was that there might be a change in parallel runway and the pilot had to disconnect the automation while aligning the aircraft to the new runway assigned to him. But the latitude-longitude plot confirmed that there was no “runway change”. The second and a more likely hypothesis was that the switching of the autopilot and flight directors could be part of an auto approach process. The pilot first disconnected the autopilots once the necessary visibility of the runway was achieved, then recycled the flight directors to engage the default modes and later decided to hand fly the aircraft. Another interesting observation is the missing auto flight lateral mode. Further investigations revealed that the “NAV” mode was active throughout the earlier part of the flight and was deactivated right before the final approach. This would not happen and is probably an artifact of the recording process. Any time either an autopilot or a flight director is engaged both a lateral mode and a vertical mode must be in use.

The entire operation was performed in an extremely windy condition. Though wind was not part of the analysis but the wind plots in Figure 5.10 helps to explain atypical fluctuations in some of the parameters like target airspeed, rudder and lateral/normal acceleration. Exceedance based model didn't detect any event for this flight.

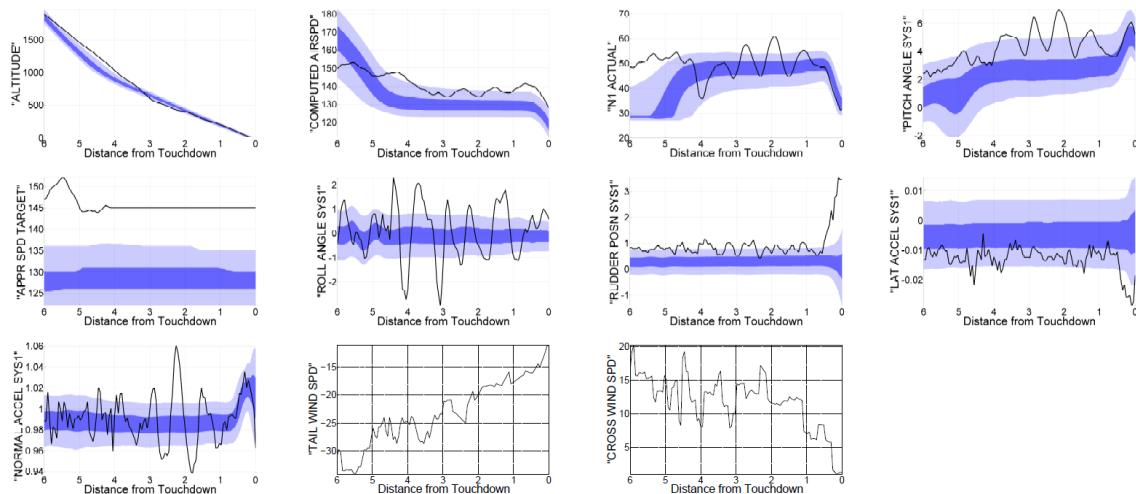


Figure 5.10 Influence of Wind

### 5.2.5 Comparison with Exceedance Detection

Among different levels of exceedance, Level 3 exceedance events are the most severe and raise concerns at airlines. Thus, Exceedance Detection at Level 3 was used as the baseline to compare the performance of ClusterAD-Flight and MKAD. Table 5.8 shows the results of comparing the flights detected by ClusterAD-Flight with Exceedance Level 3, and the flights detected by MKAD with Exceedance Level 3. It is noted that ClusterAD-Flight has detected more flights with Level 3 Exceedance than MKAD on all different detection thresholds.

Table 5.8 Comparison of ClusterAD-Flight and MKAD in Detecting Exceedances

Detection Threshold	Number of flights with Level 3 Exceedance detected by	
	ClusterAD-Flight	MKAD
1%	39	12
3%	93	31
5%	143	53
10%	220	86

The results show that ClusterAD-Flight can be considered as a variation of Exceedance Detection. It looks for deviations between abnormal values and normal values. When most operations are following standards, the normal values are the same as the target values in Exceedance Detection. The advantage of ClusterAD-Flight over Exceedance detection is that ClusterAD-Flight considers all the available flight parameters simultaneously and does not need a pre-specified list of queries.

On the other hand, a number of flights were detected by Exceedance Detection, but not by ClusterAD-Flight or MKAD. Examples of these flights are shown in this section, which are the top two flights with the maximum number of Level 3 exceedance events in the approach phase. Table 5.9 shows the type of exceedance events identified by the exceedance detection method for all three severities.

Table 5.9 Top 2 Flights with Level 3 Exceedance

Flight 1	
Level 3	Pitch Rate High at Landing, Short Flare Time, Tail Strike Risk at Landing
Level 2	Pitch High at Touchdown
Level 1	Height High at Threshold, Short Flare Distance
Flight 2	
Level 3	Speed High in Approach (at 1000 ft), Speed High in Approach (at 1000 ft),

	Low Power on Approach
Level 2	Deviation above Glideslope (1000 ft - 300 ft), Go around
Level 1	Pitch High at Touchdown, Rate of Descent High in Approach (2000 ft - 1000 ft), Approach Fast 500 RAD

Neither ClusterAD-Flight nor MKAD could detect those flights because both algorithms identify abnormal flights by considering all available flight parameters, while Exceedance Detection examine particular flight parameters under certain conditions. The abnormality level of a flight in anomaly detection methods is the combined effect of how abnormal a flight parameter is at an instance, how long the abnormality lasts for that flight parameter, and how many flight parameters are abnormal. The flight parameter plots of Flight 1 and Flight 2 (Figure 5.11 and Figure 5.12) show that most of the parameters are within the blue bands for most of the approach phase. The events detected by Exceedance detection for Flight 1 and Flight 2 are specific deviations at a particular time, such as landing, touchdown, 1000 ft, 500 ft, etc. A short time deviation of a few flight parameters may not be able to bring the flight to top of the abnormal list generated in ClusterAD-Flight and MKAD. Exceedance Detection is designed to search for problems which are foreseen. Instead, the data driven algorithms are created to search for unknown abnormalities. For example, in the case of Flight 1 (Figure 5.11), Exceedance Detection is finely tuned to pick out the high pitch rate, because it was specifically looking for this problem. “Pitch rate” was not included in the parameter list of data driven analysis; so Flight 1 was not picked up by ClusterAD-Flight or MKAD. Another example is the “Speed High in Approach” exceedance events in Flight 2. Although the overall profile of computed airspeed looks normal from the Figure 5.12, there are two small deviations in computed airspeed. The difference between computed airspeed and target airspeed around those deviations resulted in the speed related exceedance events. ClusterAD-Flight and MKAD are not sensitive to such small deviations in a short period. Thus they didn’t find Flight 2 as abnormal.

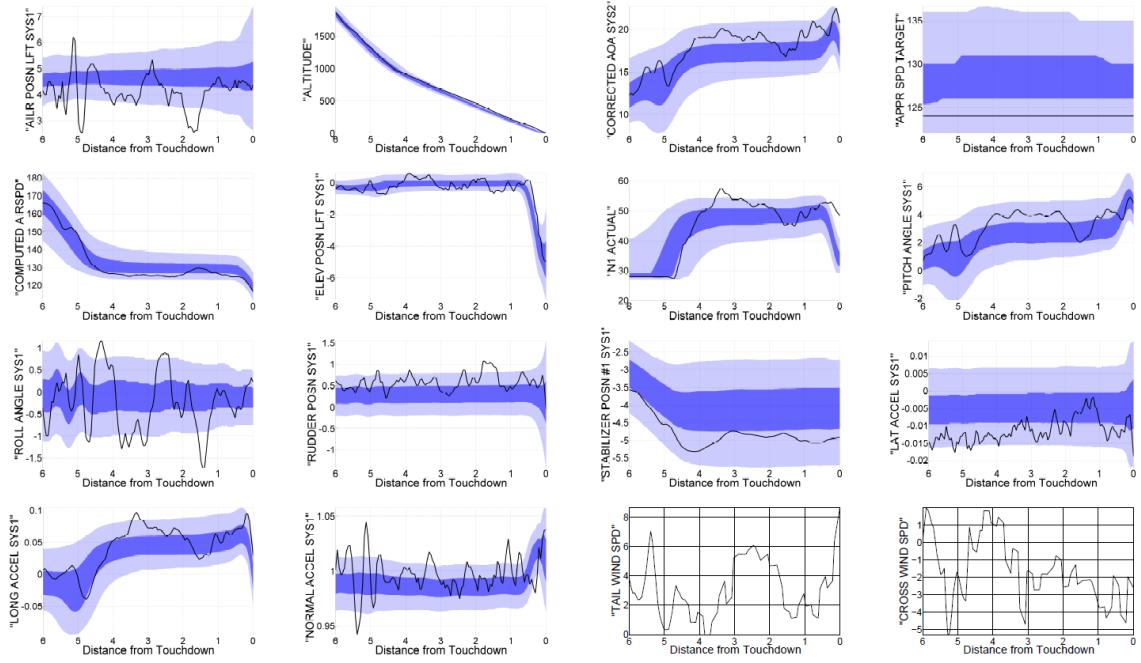


Figure 5.11 Exceedance Detection – Flight 1

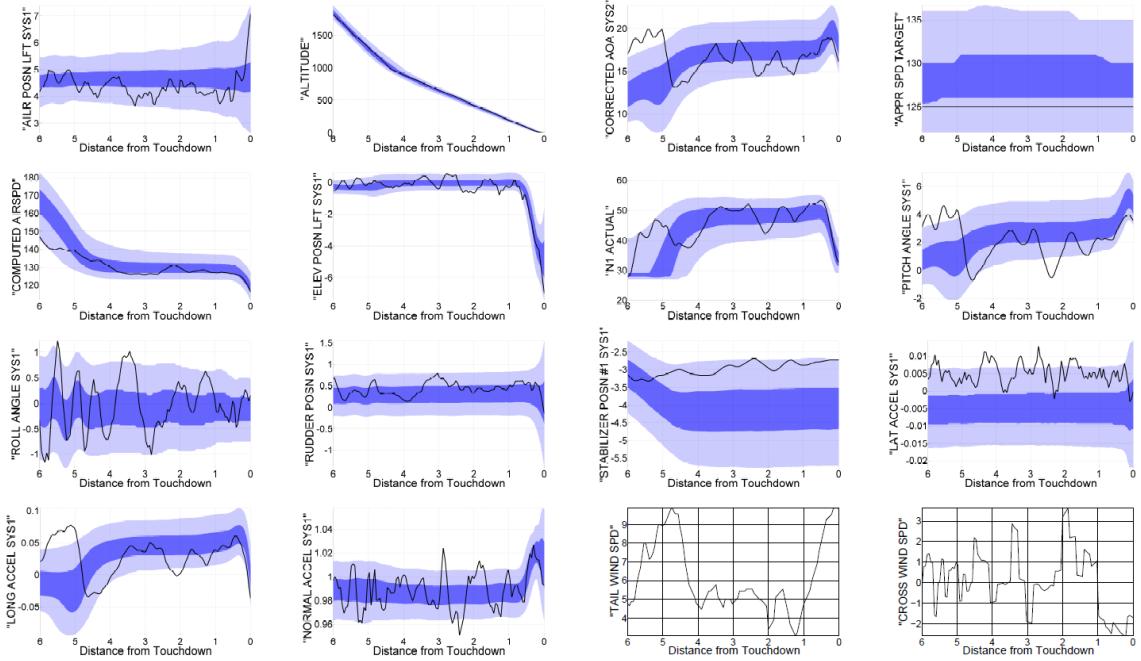


Figure 5.12 Exceedance Detection – Flight 2

### **5.2.6 Study Summary**

A comparison of ClusterAD-Flight and MKAD, and a traditional method, Exceedance Detection, was made on a common aviation data set. Results showed that ClusterAD-Flight was able to detect operationally significant anomalies, which were not captured by Exceedance detection.

There was some overlap among anomalies detected by ClusterAD-Flight, MKAD and Exceedance Detection, yet, each algorithm had its unique strengths. ClusterAD-Flight worked better with continuous flight parameters; MKAD was more sensitive to the sequence of discrete parameters; Exceedance Detection was accurate in examining known issues. Therefore, the strengths from different methods shall be combined.

## **5.3 Evaluation Study II: Comparison of ClusterAD-Flight, ClusterAD-Data Sample, and MKAD in Detecting Exceedances**

The comparison study of ClusterAD-Flight, MKAD and Exceedance Detection showed that both ClusterAD-Flight and MKAD had limited capability to detect Level 3 Exceedance events, which are considered as known issues in the traditional FOQA analysis method. This was one of factors that motivated the development of ClusterAD-Data Sample. After ClusterAD-Data Sample was developed, Evaluation Study II was performed to assess ClusterAD-Data Sample, ClusterAD-Data Sample, MKAD and Exceedance Detection.

### **5.3.1 Comparison Design and Setup**

The objective of this study was to compare ClusterAD-Data Sample, ClusterAD-Flight, MKAD, and Exceedance Detection. The four algorithms were applied on a same set of flight data, Dataset II. This study focused on the agreement with Exceedance Detection because access to the dataset and results were also restricted by the collaboration agreement with NASA. Only aggregated statistical results were available in this study, e.g. number of exceedances, number of flights detected by ClusterAD-Data Sample, and number of commonly detected flights. Detailed information of specific flights was not available.

## **Algorithms**

This study compared anomaly detection algorithms: ClusterAD-Data Sample, ClusterAD-Flight, MKAD, and Exceedance Detection. ClusterAD-Data Sample is also a cluster-based anomaly detection algorithm which is developed in this thesis. It is similar to ClusterAD-Flight but uses a different data transformation technique. Details of the other three algorithms can be found in Section 5.2.2.

## **Algorithm Settings**

All four algorithms can be set at different sensitivity levels for detection.

The sensitivity level of ClusterAD-Data Sample, ClusterAD-Flight and MKAD is set by the “detection threshold” parameter. We used a series of detection thresholds to test algorithm performance on different sensitivity levels (detection threshold = 1%, 3%, 5%, 10%). The comparison among ClusterAD-Data Sample, ClusterAD-Flight and MKAD was always made on the same detection threshold to make the results comparable.

Exceedance Detection detects “exceedance events” when aircraft performance fails to meet target value ranges during specific maneuvers. The industry normally uses three levels to detect exceedance events. Level 1 indicates minor deviation from the performance target, Level 2 indicates moderate deviation, and Level 3 indicates the severest ones. Because Level 3 exceedance events are the issues of most concerns to airlines, the comparisons with Exceedance Detection were focused on Level 3 in this study.

## **Dataset**

The Dataset II, which was used in Evaluation Study I, was used again in this study. It included 25519 A320 flights landing at a standard European airport. The analysis focused on the approach phase (from 6nm before touchdown to touchdown). Details of this dataset can be found in Section 5.2.2.

### **5.3.2 Results**

Among different levels of exceedance, Level 3 exceedance events are the most severe and raise concerns at airlines. Thus, Exceedance Detection at Level 3 was used as the baseline to compare

the performance of ClusterAD-Data Sample, ClusterAD-Flight and MKAD. Figure 5.13 shows the results of comparing the flights detected by ClusterAD-Data Sample, ClusterAD-Flight and MKAD with Exceedance Level 3. ClusterAD-Data Sample detected more flights with Level 3 events than ClusterAD-Flight and MKAD across all detection thresholds, as shown in Figure 5.13. For example, when the detection threshold was set to detect the top 10% abnormal flights, 70% of flights with Level 3 events were identified by ClusterAD-Data Sample, in contrast, only 12% were identified by MKAD and 30% by ClusterAD-Flight.

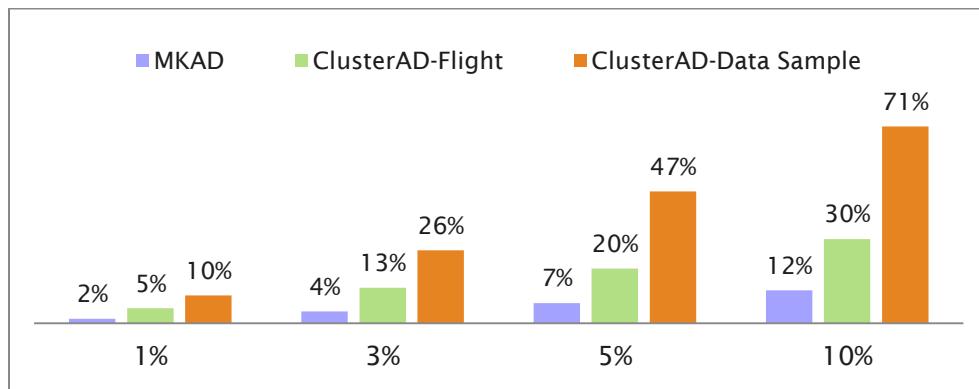


Figure 5.13 Percentage of Flights with Level 3 Exceedance Events Detected by MKAD, ClusterAD-Flight, and ClusterAD-Data Sample

### 5.3.3 Study Summary

In this study, ClusterAD-Data Sample showed improvement in detecting flights with severer exceedance events. This result was expected because the cluster analysis in ClusterAD-Data Sample is based on data samples, which are also the basis of defining exceedance events. However, the agreement with Exceedance Detection measures only one aspect of the detection algorithm. It is not clear whether ClusterAD-Data Sample detects new types of safety issues, which are not included in Exceedance Detection, better or not.

## 5.4 Evaluation Study III: Evaluation of ClusterAD algorithms and Data Visualization Tools with Domain Experts

An evaluation study was performed to evaluate the operational significance of flights detected by ClusterAD-Flight and ClusterAD-Data Sample as perceived by domain experts, and to test the

data visualization tools. In the evaluation, domain experts were asked to review a number of flights detected by anomaly detection algorithms. Data visualization tools described in previous chapter were used to facilitate the review process. The evaluation was designed to obtain domain experts feedback on the data visualization tools and to compare perceived operational significance of flights detected by ClusterAD-Flight and ClusterAD-Data Sample. The comparison between ClusterAD-Flight and ClusterAD-Data Sample and results on data visualization tools are presented in this section.

### **5.4.1 Evaluation Design**

#### **Objective**

The objectives of this evaluation were 1) to demonstrate and test the data visualization tools; 2) to compare perceived operational significance of flights detected by ClusterAD-Flight and ClusterAD-Data Sample.

#### **Dataset**

A FDR dataset, Dataset III, was obtained from an oversea airline with the support of the FAA, which consisted of 10528 A320 flights that originated from or arrived at 36 airports. Each flight has 142 flight parameters sampled at 0.5 Hz or 0.1 Hz depending on the altitude during two flight phases: 1) from takeoff to 10000 ft AGL, 2) from 10000 ft AGL to touchdown. No exceedance information was available for this dataset and data entries were de-identified. This study focused on the approach phase: from 6nm before touchdown to touchdown.

#### **Independent Variables**

With the two objectives, the evaluation was designed to have three independent variables: type of flight parameter plot, availability of flight abnormality visualization, and detection algorithm, as shown in Figure 5.14.

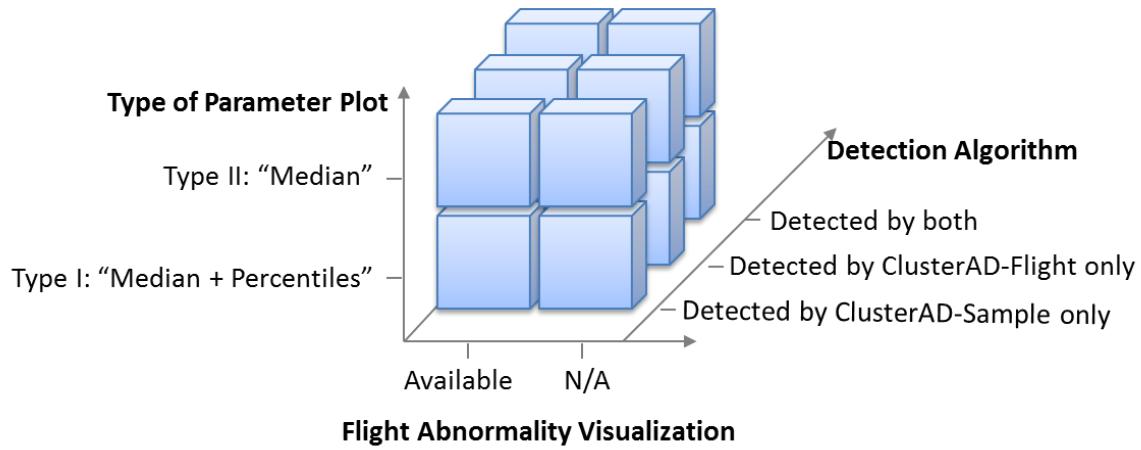


Figure 5.14 Evaluation Design: Independent Variables

The first two of the independent variables were related to the first objective: to demonstrate and test the data visualization tools. Two types of flight parameter plots were tested. Type I of the flight parameter plot displays median and percentiles as reference information, as shown in Figure 5.15. Type II only displays median as the reference, as shown in Figure 5.16. Type I provides more information than Type II, but it might introduce bias from the distribution of the dataset. As to the second independent variable, it was designed to test if there is any benefit from providing the Flight Abnormality Visualization tool Figure 5.17.

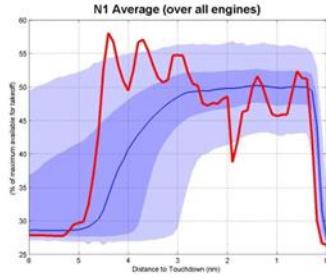


Figure 5.15 Type I Flight Parameter Plot: Median + Percentiles

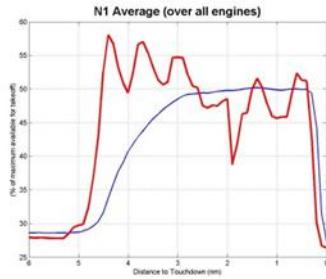


Figure 5.16 Type II Flight Parameter Plot: Median only

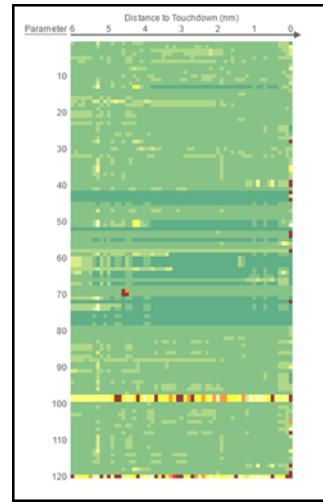


Figure 5.17 Flight Abnormality Visualization

Data visualization tools were presented using one of the four configurations by combining of the first two independent variables at different levels: A) Type I flight parameter plot + Flight abnormality probe, B) Type II flight parameter plot + Flight abnormality probe, C) Type I flight parameter plot only, D) Type II flight parameter plot only, as illustrated in **Error! Reference source not found.**. Each subject experienced all the four configurations in order.

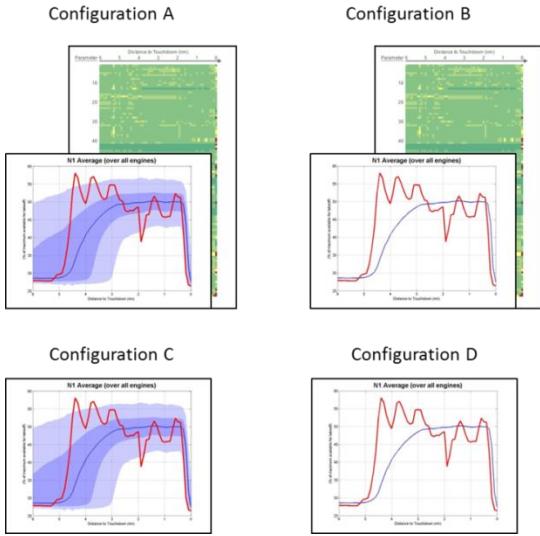


Figure 5.18 Four Configurations of Data Visualization Tools

Lastly, the third independent variable was directly related to the second objective: to compare ClusterAD-Flight and ClusterAD-Data Sample. Flights detected by different anomaly detection algorithms were organized into three groups: 1) commonly detected by ClusterAD-Flight and ClusterAD-Data Sample; 2) detected by ClusterAD-Flight; 3) detected by ClusterAD-Data Sample.

## Flights to Be Reviewed

ClusterAD-Flight and ClusterAD-Data Sample were applied to Dataset III. At each level of detection threshold (top  $x\%$  abnormal flights), there were flights commonly detected by ClusterAD-Data Sample and ClusterAD-Flight, and unique ones detected by either algorithm. The results of ClusterAD-Data Sample and ClusterAD-Flight were expected to be different because they use different detection strategies. Comparing flights detected by the two algorithms, a portion of the flights was always commonly detected by both algorithms at different detection thresholds. The number of common flights detected by both varied from 20 (Detection Threshold = 0.5%), 55 (Detection Threshold = 1%), to 157 (Detection Threshold = 3%), as shown in Figure 5.19. The percentage of common detection kept relatively constant at different detection thresholds.

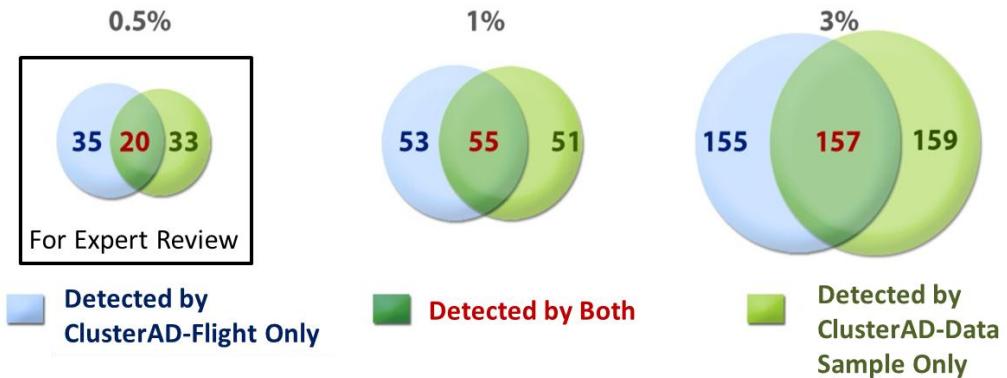


Figure 5.19 Abnormal Flights Detected by ClusterAD-Flight and ClusterAD-Data Sample  
(Dataset: 10528 A320 flights)

Flights were selected from those detected by two algorithms at detection threshold = 0.5%. There were eighty-eight flights were identified by either ClusterAD-Flight or ClusterAD-Data Sample at that detection threshold. These flights were organized into three groups:

- Flights detected by both ClusterAD-Flight and ClusterAD-Data Sample (20 flights)
- Flights detected by ClusterAD-Flight only (35 flights)
- Flights detected by ClusterAD-Data Sample only (33 flights)

Ideally, all these flights would be reviewed by domain experts. However, limited by the number of domain experts and time of each domain expert available in this study, an initial analysis was performed to reduce the number of flights need to be reviewed by domain experts. The initial analysis selected flights with representative data patterns. Since there were a finite number of data patterns observed in each of the three groups, a flight was select to represent a type of data pattern in every group. After the initial analysis, sixteen flights were selected for domain experts to review, as shown in Table 5.10.

Table 5.10 Flights Selected for Expert Review from Abnormal Flights at Detection Threshold = 0.5%

Detected by ClusterAD-Flight & ClusterAD-Data Sample	Detected by ClusterAD-Flight	Detected by ClusterAD-Data Sample
1209099	1209326	1200517
1296837	1264410	1222631
1320982	1314619	1316840
1341150	1349379	1366200
1360984		1373774
1384717		

	1393754		
Total	7	4	5

## Randomization

Every participant was assigned to review all 16 flights, which were ordered into four groups with equal size. Flights in a group were presented using one of the four configurations of data visualization tools. Each subject experienced all the four configurations in order. The order was randomized between subjects to counterbalance the learning effect. The assignment of which flight using which configuration of data visualization tools was also randomized between subjects, so that minimum bias was introduced, not always presenting particular flight features via a configuration of data visualization tools.

## Dependent Variables

The comparison between ClusterAD-Flight and ClusterAD-Data Sample was based on an operational review questionnaire, which aimed to assess each flight's perceived operational characteristics from three aspects: its abnormality level in operational sense, its safety implications, and detailed description about its abnormal behavior when applicable. The operational review questionnaire is shown in Figure 5.20.

<b>1. This flight is abnormal, based on my operational experience.</b>
<input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Strongly disagree      Disagree      Neutral      Agree      Strongly agree
<b>2. This flight represents safety hazards, based on my operational experience.</b>
<input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Strongly disagree      Disagree      Neutral      Agree      Strongly agree
<b>3. Please describe this flight. If it is abnormal, please describe what is abnormal about the flight.</b>
<hr/> <hr/>

Figure 5.20 Operational Review Questionnaire

A standard usability questionnaire, the After-Scenario Questionnaire (ASQ) (Lewis, 1991a, 1991b, 1995), was used to evaluate the data visualization tools, as shown in Figure 5.21.

<b>1. Overall, I am satisfied with the ease of completing the tasks in this scenario.</b>
<input type="radio"/> Strongly disagree <input type="radio"/> Disagree <input type="radio"/> Neutral <input type="radio"/> Agree <input type="radio"/> Strongly agree
<b>2. Overall, I am satisfied with the amount of time it took to complete the tasks in this scenario.</b>
<input type="radio"/> Strongly disagree <input type="radio"/> Disagree <input type="radio"/> Neutral <input type="radio"/> Agree <input type="radio"/> Strongly agree
<b>3. Overall, I am satisfied with the support information (on-line help, messages, documentations) when completing this task.</b>
<input type="radio"/> Strongly disagree <input type="radio"/> Disagree <input type="radio"/> Neutral <input type="radio"/> Agree <input type="radio"/> Strongly agree

Figure 5.21 Data Visualization Tools Evaluation Questionnaire

## 5.4.2 Apparatus, Participants, and Procedure

### Apparatus

The evaluation was administrated online, in order to facilitate access experts at different locations. HTML and JavaScript were used to develop the online test-bed. In the test-bed, flights to be reviewed were presented via data visualization tools, followed by the questionnaires. Limitations of the test bed should be noted. It focuses on the data visualizations tools, instead of the entire expert review process. For example, other functions commonly involved in the review process, e.g. 3D animation of the flight, were not included in the test bed.

### Participants

With the help of the Air Line Pilots Association (ALPA) and professional pilot forums, four domain experts volunteered to participate and completed the evaluation. The four domain experts are airline pilots with more than 48,600 hours of flying experience combined. All domain experts have experience with FOQA program and some are in charge of a FOQA program.

## **Procedure**

A tutorial was given to each participant at the beginning of the evaluation through web conference, which discussed the nature of the evaluation and explained the context and use of the interface. A practice case was given to each participant after the tutorial, which presented a flight with known abnormalities in particular flight parameters. After the participant was confirmed that he/she understood the use of the interface, a link of the review session was sent to the participant.

In the review session, 16 flights detected by ClusterAD algorithms were reviewed by each participant. During the review of each flight, its information was presented in a format as shown in Figure 5.22. Flight basic information was available to all flights; Flight parameter plot was given as either Type I or Type II; the Flight Abnormality Visualization tool was only available to half of the flights. After the flight information presentation, the data visualization tools questionnaire and operational review questionnaire were followed.

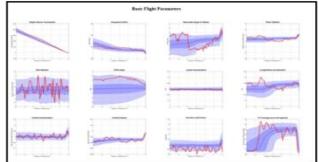
The review session had 16 flights to be reviewed for each participant. The participants did not need to finish all the flights at once. Uncompleted review information was stored, and the review process could be continued from where it was left.

### Flight Basic Information

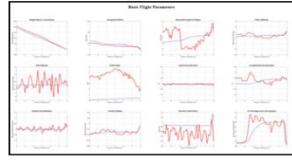
Origin - Destination: EGLL / LHR - LSZH / ZRH  
Month: January      Landing weight: 59653(kg)

### Flight Parameter Plot

Type I



Type II



### Flight Abnormality Visualization

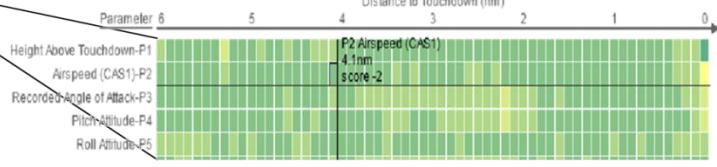
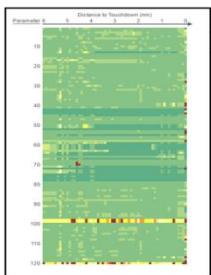


Figure 5.22 Format of Presenting Flight Information

### 5.4.3 Results on Perceived Operational Significance of Abnormal Flights

#### Confounding Factor

Results may be confounded by the fact that each flight was reviewed by four domain experts, each of whom used one of the four configurations of data visualization tools. Perceived operational significance of a same flight could be a result of the difference in visualization tools and the difference between experts. Ideally, each flight should be reviewed by multiple domain experts under exactly the same visualization configurations. Conclusions drawn from this evaluation, especially on the inter-rater agreement, should be treated with caution because of this confounding factor. To eliminate this confounding factor, further evaluations are needed in the next step of this research.

## Results Overview

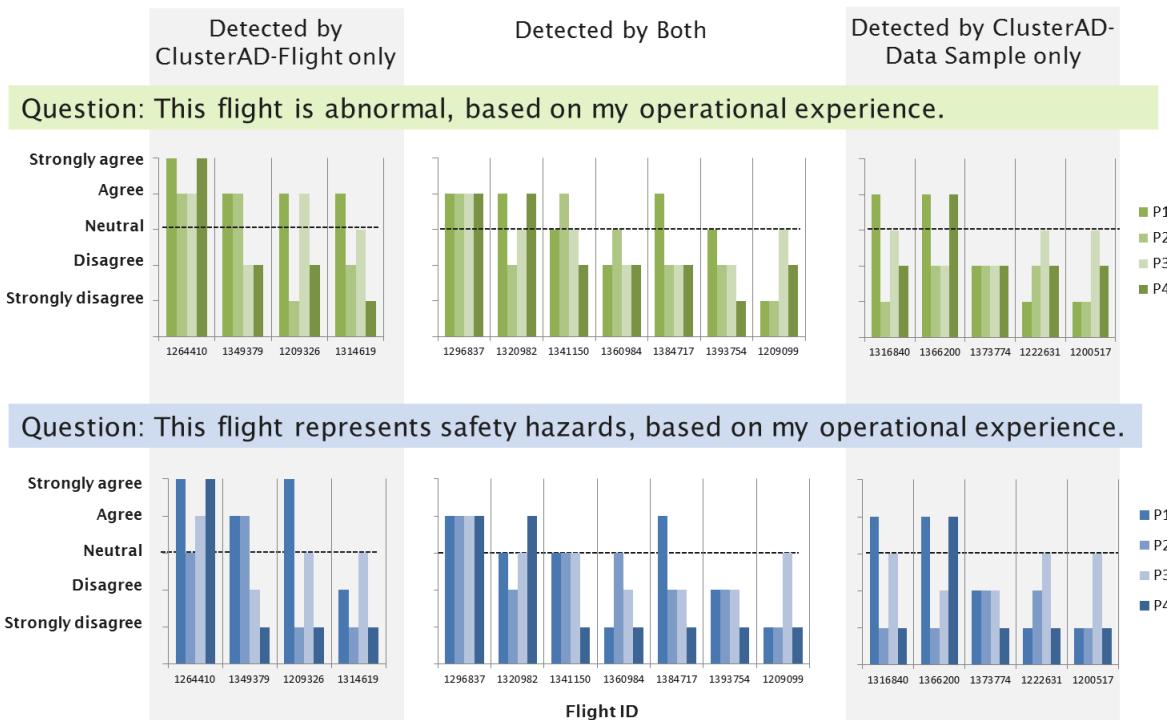


Figure 5.23 Operational Review Results of Question 1 and Question 2

The results of flights' perceived abnormality level and perceived safety implications in the operational review questionnaire across all selected flights are summarized in Figure 5.23. Based on the results, several observations were made:

- 1) The degree of perceived operational abnormality varied across flights. Evaluated by domain experts, some flights were operationally abnormal and representing safety hazards, some were operationally abnormal but not indicating any safety hazards and some were totally benign.
- 2) Flights' perceived "level of abnormality" measured by Question 1 correlates with the perceived "degree of representing safety hazards" by Question 2 based on the ratings by domain experts, and the correlation is statistically significant based on Spearman Rank Correlation Coefficient ( $r_s = 0.90, p = 1 \times 10^{-24}$ ). This indicates that the statement "a flight is operationally abnormal." is perceived similar to the statement "a flight represents safety hazards" by domain experts.

3) The inter-rater agreement varies across flights for both “level of abnormality” (Question 1) and “degree of representing safety hazards” (Question 2). For example, Flight 1296837 was always considered as abnormal and representing safety hazards by all experts, while Flight 1366200 was evaluated as normal and safe by two experts but was considered as abnormal and unsafe by the other two experts. The Krippendorff’s Alpha was calculated as a measure of inter-rater agreement, shown in Table 5.11. The result exhibits that Group 1 flights that were detected by both ClusterAD-Flight and ClusterAD-Data Sample have a higher agreement among experts, Group 2 is lower, and Group 3 is the lowest. Group 3 has the lowest inter-rater agreement, which could be explained by special characteristics of flights in this group. Flights in Group 3 had many different types of instantaneous anomalies, the perceived degree of abnormality depended on how many instantaneous anomalies were identified by domain experts and the level of operational impact of those instantaneous anomalies.

The low inter-rater agreement could also be caused by the difference in data visualization configurations. Since only four domain experts participated in this study and four configurations were used, every flight was reviewed by four domain experts using four different configurations. The disagreement of a flight could be caused by difference in personal opinions, as well as the difference in visualization tools.

Table 5.11 Inter-rater Agreement across Groups

	Flights Detected by Both	Flights Detected by ClusterAD- Flight	Flights Detected by ClusterAD- Data Sample
Question 1: The flight is abnormal	0.38	0.27	0.04
Question 2: The flight represents safety hazards	0.38	0.20	0.10
Measured by Krippendorff’s Alpha: $\alpha = 1$ indicates perfect agreement; $\alpha = 0$ indicates the absence of agreement			

## Comparison between ClusterAD-Flight and ClusterAD-Data Sample

The operational review questionnaire also included a detailed description of flight’s operational characteristics. For each group, a few examples in each group are presented in detail to further

evaluate the strengths and weakness of ClusterAD-Flight and ClusterAD-Data Sample, as shown in Table 5.12.

Table 5.12 Operational Review Summary

Group	Flight	Abnormal?				Hazardous?				Example
		P1	P2	P3	P4	P1	P2	P3	P4	
Detected by Both	1296837	4	4	4	4	4	4	4	4	✓
	1320982	4	2	3	4	3	2	3	4	✓
	1341150	3	4	3	2	3	3	3	2	
	1360984	2	3	2	2	1	3	2	2	
	1384717	4	2	2	2	4	2	2	2	
	1393754	3	2	2	1	2	2	2	1	
	1209099	1	1	3	2	1	1	3	2	✓
Detected by ClusterAD-Flight	1264410	5	4	4	5	5	3	4	5	✓
	1349379	4	4	2	2	4	4	2	2	✓
	1209326	4	1	4	2	5	1	3	2	
	1314619	4	2	3	1	2	1	3	1	
Detected by ClusterAD-Data Sample	1316840	4	1	3	2	4	1	3	2	
	1366200	4	2	2	4	4	1	2	4	✓
	1373774	2	2	2	2	2	2	2	2	✓
	1222631	1	2	3	2	1	2	3	2	
	1200517	1	1	3	2	1	1	3	2	

Legend: 1-Strongly disagree, 2-Disagree, 3-Neutral, 4-Agree, 5-Strongly agree

## Operational characteristics of flights detected by ClusterAD-Flight only

Question: This flight is abnormal, based on my operational experience.

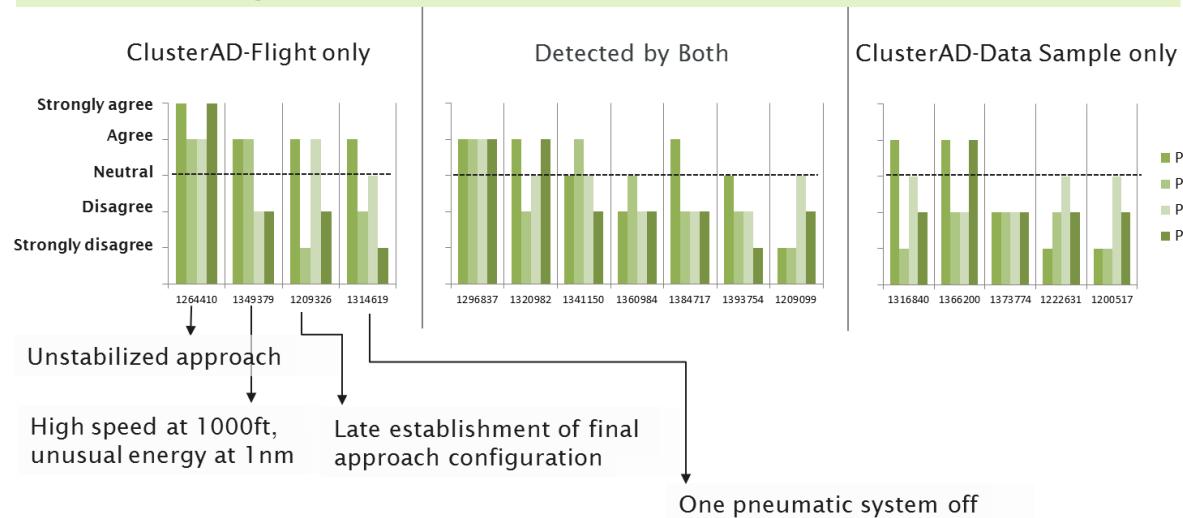


Figure 5.24 Operational Characteristics of Flights Detected by ClusterAD-Flight Only

All flights in this group were considered as operationally abnormal by at least one expert. The operational characteristics are summarized in Figure 5.24. Two examples are presented and discussed: one was agreed by all experts as operationally abnormal (Flight 1264410), the other one was perceived opposite among experts regarding its degree of abnormality.

### Flight 1264410 - High, fast, and high-rate-of-descent unstabilized approach.

This flight was rated as operationally abnormal and representing safety hazards by all domain experts. It was above the glideslope profile until 2nm before touchdown with high airspeed and high rate of descent. Engine thrust was at idle until 1.3nm from touchdown. The vertical speed was almost 4000fpm at 3000ft HAT, and it reduced to 1000fpm at 1000ft HAT. One expert noted that the roll attitude profile indicated turns on final approach, which might be intended to increase distance to the runway to provide room to lose altitude. All observations indicate this was an unstabilized approach.

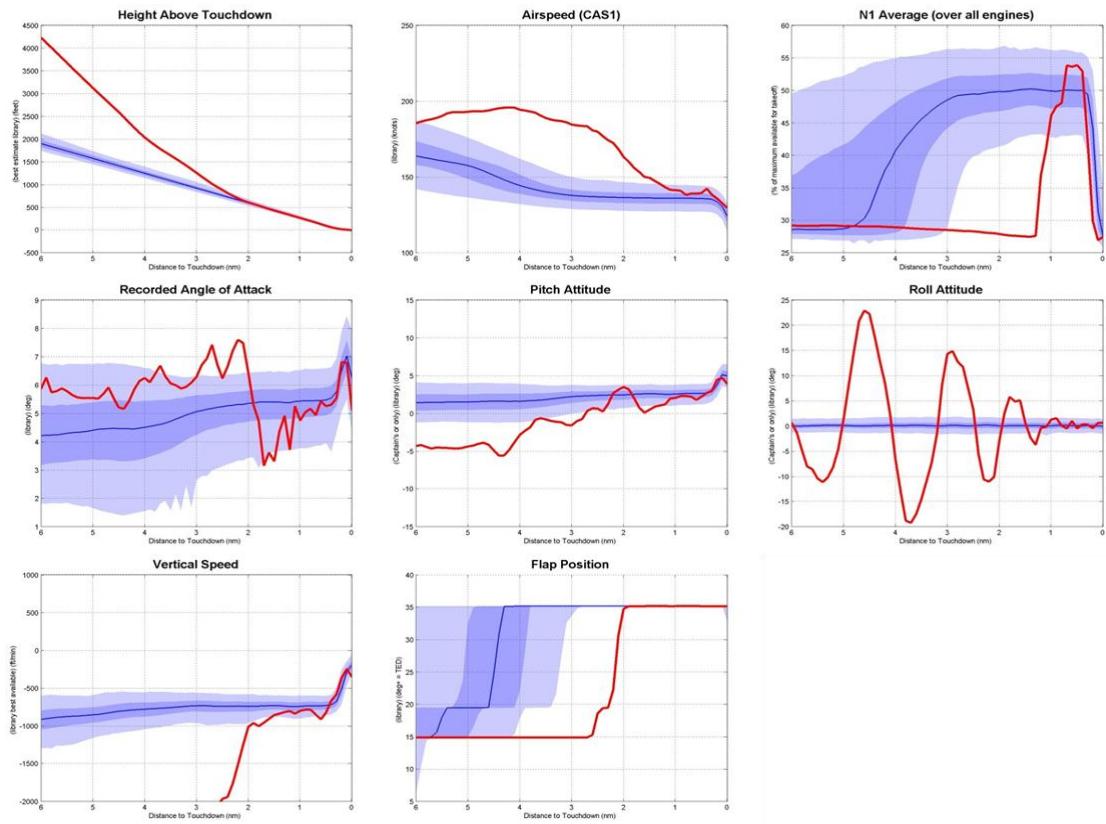


Figure 5.25 Flight 1264410 - High, fast, and high-rate-of-descent unstabilized approach

## Flight 1349379 - unusual energy profile

Two domain experts thought this flight as normal, while other two experts found abnormality in this flight. One expert identified abnormalities in the flap setting and airspeed - "Coming in with Flaps1 by 6 nm instead of the usual Flaps 2" and "airspeed high at 1000ft HAT." Another expert identified abnormalities around 2nm before touchdown. The thrust had been left at idle until 2nm and then increased to normal setting. The expert commented the application of thrust was a bit late in his experience, permitting the Calibrated airspeed (CAS) and groundspeed (GS) reduced by 10kts, which might be below Vapp, as CAS and GS was increased and thrust responded rapidly at 1nm. He suggested that a call to the crew using standard FOQA principles would be required to clarify what was occurring between 3nm and touchdown.

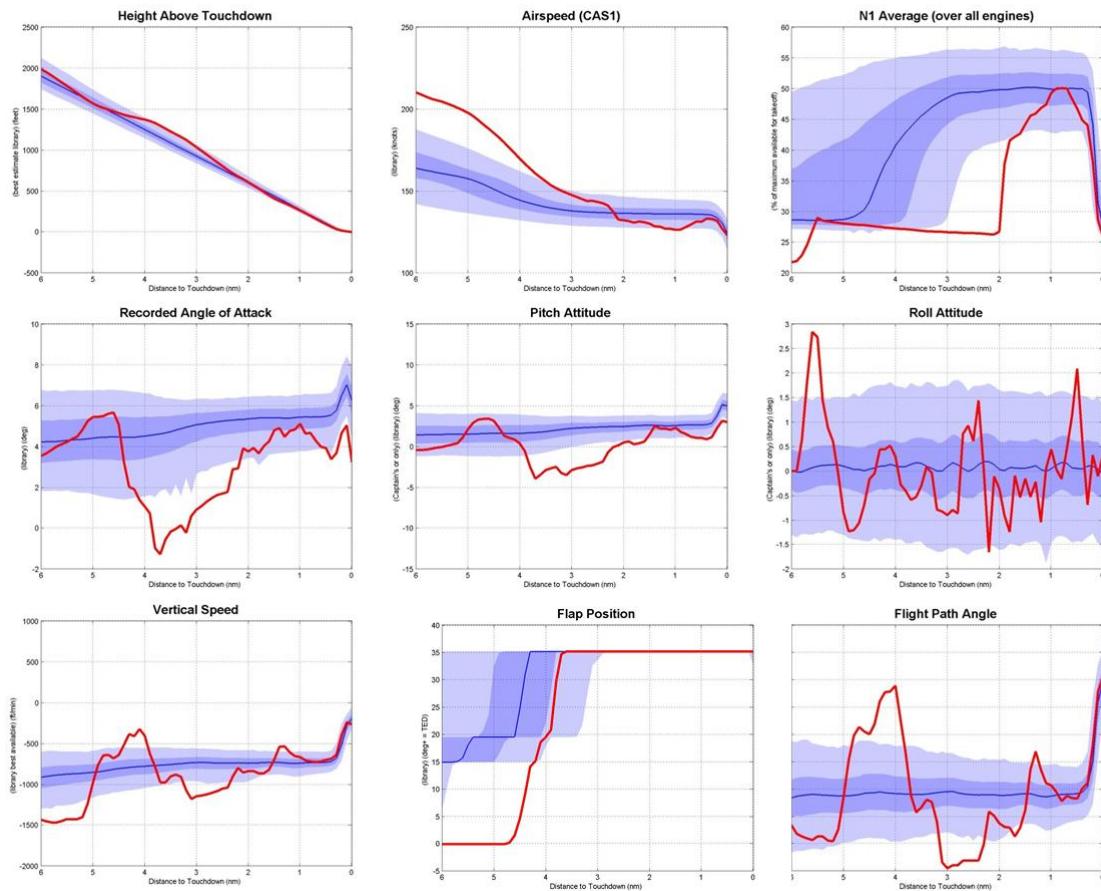


Figure 5.26 Flight 1349379 - unusual energy state at 1nm before touchdown

## Operational characteristics of flights detected by ClusterAD-Data Sample only

Question: This flight is abnormal, based on my operational experience.

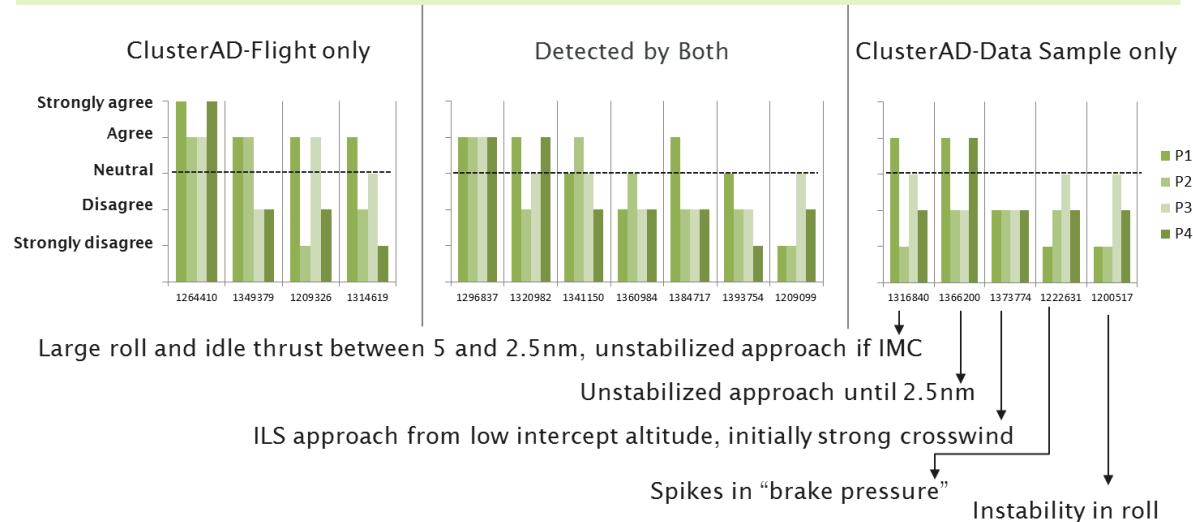


Figure 5.27 Operational Characteristics of Flights Detected by ClusterAD-Data Sample Only

Operational characteristics of flights in this group are summarized in Figure 5.27. None of the flights in the group were agreed by four domain experts. One example with conflicting ratings from different experts is Flight 1366200. Another general observation in this group is that flights exhibited instantaneous anomalies in data patterns, but not all were considered as operationally abnormal. A typical example is Flight 1373774, whose abnormality in wind related parameters was not considered as operationally abnormal by domain experts.

### Flight 1366200 - manual flight with no flight director

This flight was considered as operationally abnormal by two experts, but normal by the other two experts. Experts who considered it abnormal commented this flight came in at correct speed but slightly high, power was idle until about 2nm, roll was instable, and it was not stabilized at 1000ft HAT. One interesting observation was that experts who rated it as normal found the vertical speed value abnormal between 4.3nm and 2.2nm and brake pressure abnormal between 6nm and 5nm, but still thought the flight was operationally normal. One expert explained the vertical speed returned to stabilized values at about 700ft above airport elevation and the brake

pressure spike may be a Landing Gear Control Interface Unit (LGCIU) system test as there are no reasons for brake application at this point in the flight, nor does such application affect the flight.

This example shows that abnormality in data patterns was detected by ClusterAD-Data Sample and confirmed by domain experts. However, the abnormal data pattern does not always indicate abnormality by operational standards.

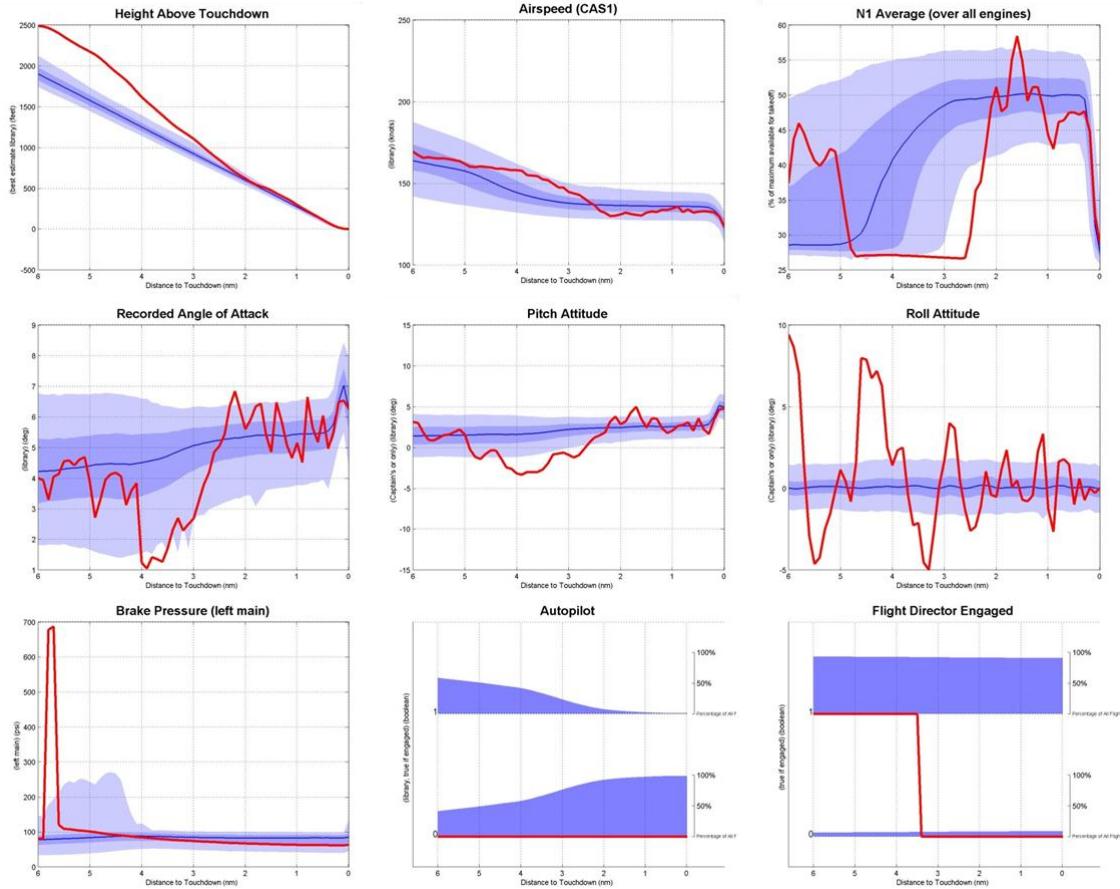


Figure 5.28 Flight 1366200 - manual flight with no flight director

## Flight 1373774 - high wind approach

All domain experts agreed that this flight was a normal approach with an initially-strong headwind and crosswind. The engine thrust had abnormal data pattern between 6nm and 3nm before touchdown. It was because the glide slope was intercepted from below and flap was taken at 4nm. This pattern was uncommon in this dataset, but it was considered as a very good fuel-saving, noise-reduced decelerated approach technique by a domain expert. This flight is a typical example of flights with unusual data patterns but without operational abnormality.

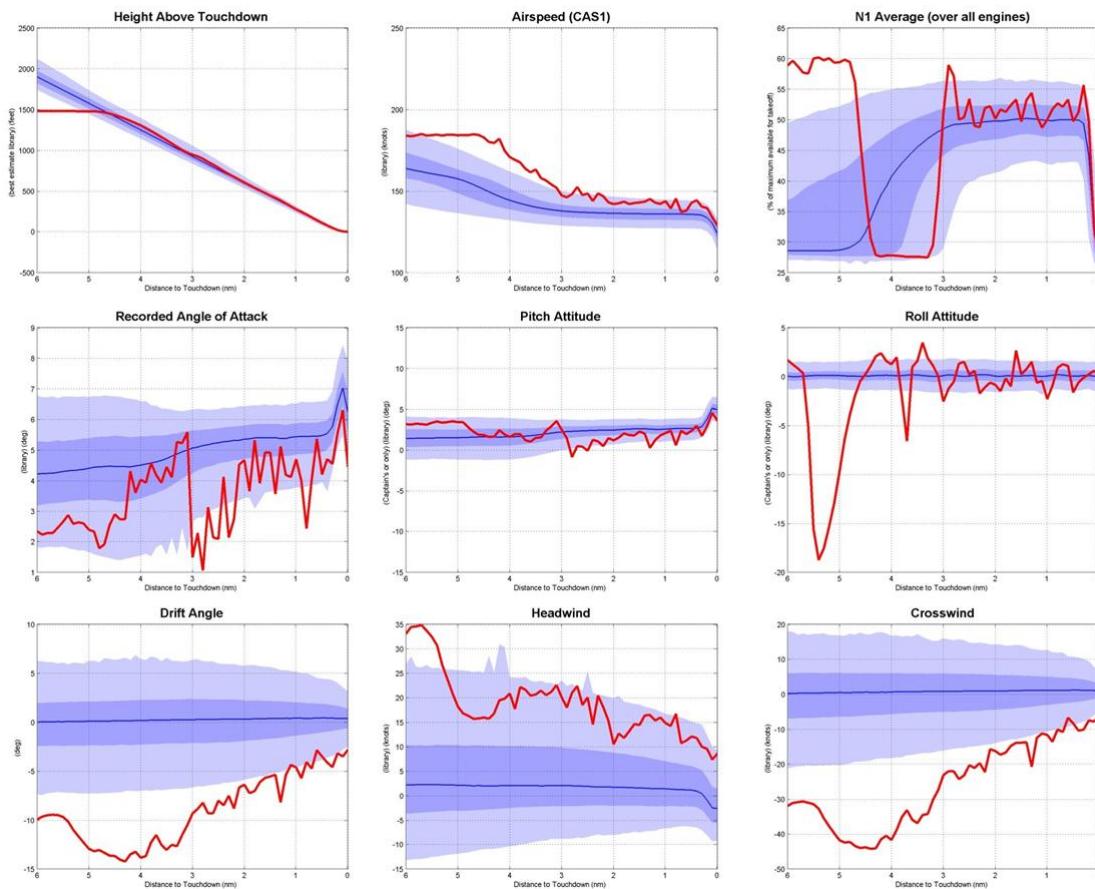


Figure 5.29 Flight 1373774 - high wind approach

## Operational characteristics of flights detected by both ClusterAD-Flight and ClusterAD-Data Sample

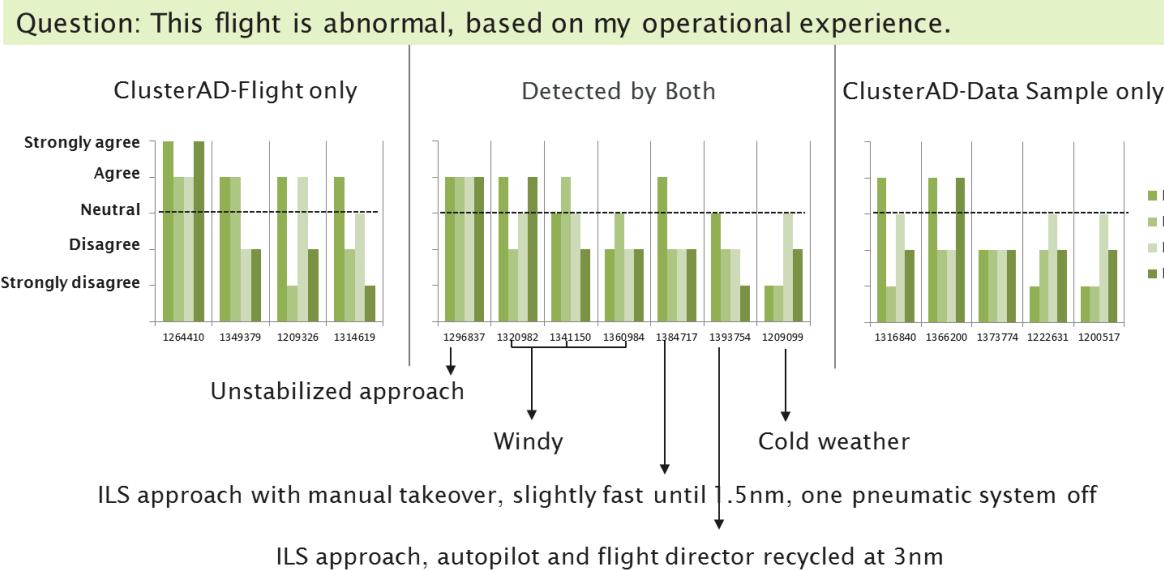


Figure 5.30 Operational Characteristics of Flights Detected by Both ClusterAD-Flight and ClusterAD-Data Sample

Operational characteristics of flights in this group are summarized in Figure 5.30. Three flights were chosen as examples to be discussed in this group: one example of strong indication of operational abnormality and hazardousness (Flight 1296837), one example of conflicting ratings (Flight 1320982), and one example of no identified operational abnormality and hazardousness (Flight 1209099).

### Flight 1296837 - unstabilized high energy approach

Flight 1296837 was reviewed as operationally abnormal and representing safety hazards by all four domain experts. Information of key flight parameters of this flight is shown in Figure 5.31. This flight was commented as an unstabilized approach with high speed, late flap configuration and idle thrust until very short final. Its behavior was away from "normal" in terms of expected stabilized approach criteria at 1000 ft. One expert stated, "As a pilot and also as an operator I would consider this approach as requiring a go-around and it is an argument for the 1000ft HAT stabilized requirement but not all pilots or operators see it this way."

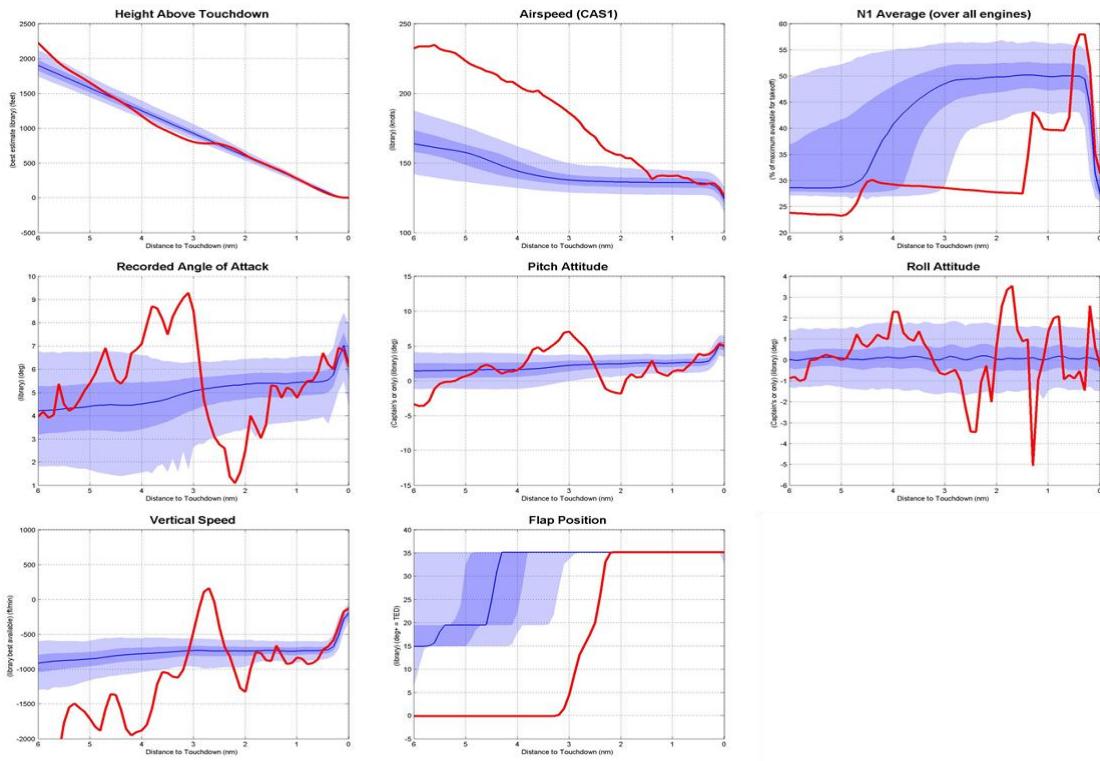


Figure 5.31 Flight 1296837 - unstabilized high energy approach

## Flight 1320982 – Strong tailwind and late localizer intercept approach

Two experts considered this flight as operationally abnormal, while the other two experts didn't agree. Only one expert thought this flight represented safety hazards. Experts who rated the flight as abnormal commented "Strange level off by 3 nm, late localizer intercept (less than 3 miles) but no glideslope hold mode." and "Thrust was at idle until 2.5 miles on final. If we are in IMC, it would mean an unstabilized approach." On the other hand, experts who viewed this flight as normal said "This approach was conducted with a strong tailwind but was close to stabilized by the 500ft point." But expressed concerns on airspeed control between 4 and 3nm on the approach.

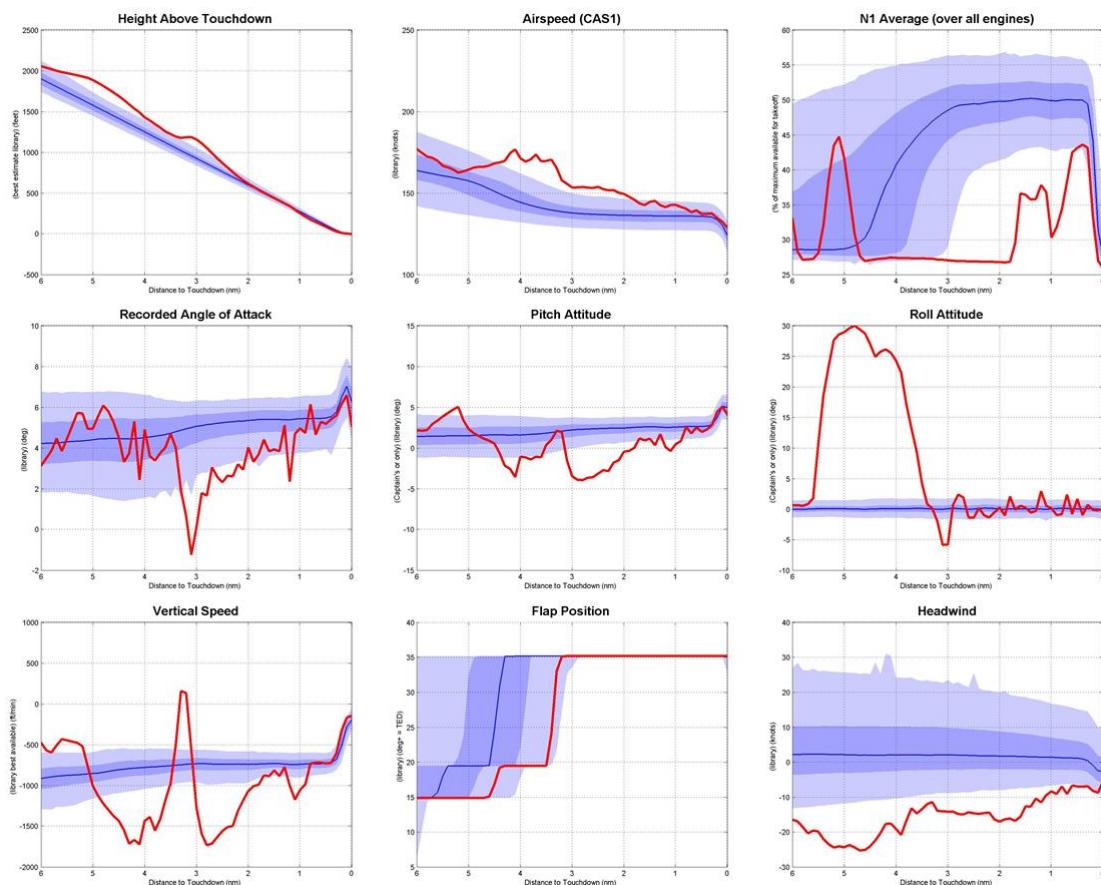


Figure 5.32 Flight 1320982 - Strong tailwind & late localizer intercept approach

## Flight 1209099 - manually-flown approach in cold weather

All experts considered this flight as operationally normal. The flight was a manually-flown approach in cold weather based on the flight parameter information as shown in Figure 5.33. It was detected by both ClusterAD-Flight and ClusterAD-Data Sample because of the high air pressure. This is an example of flights with abnormal data pattern, but operationally normal.

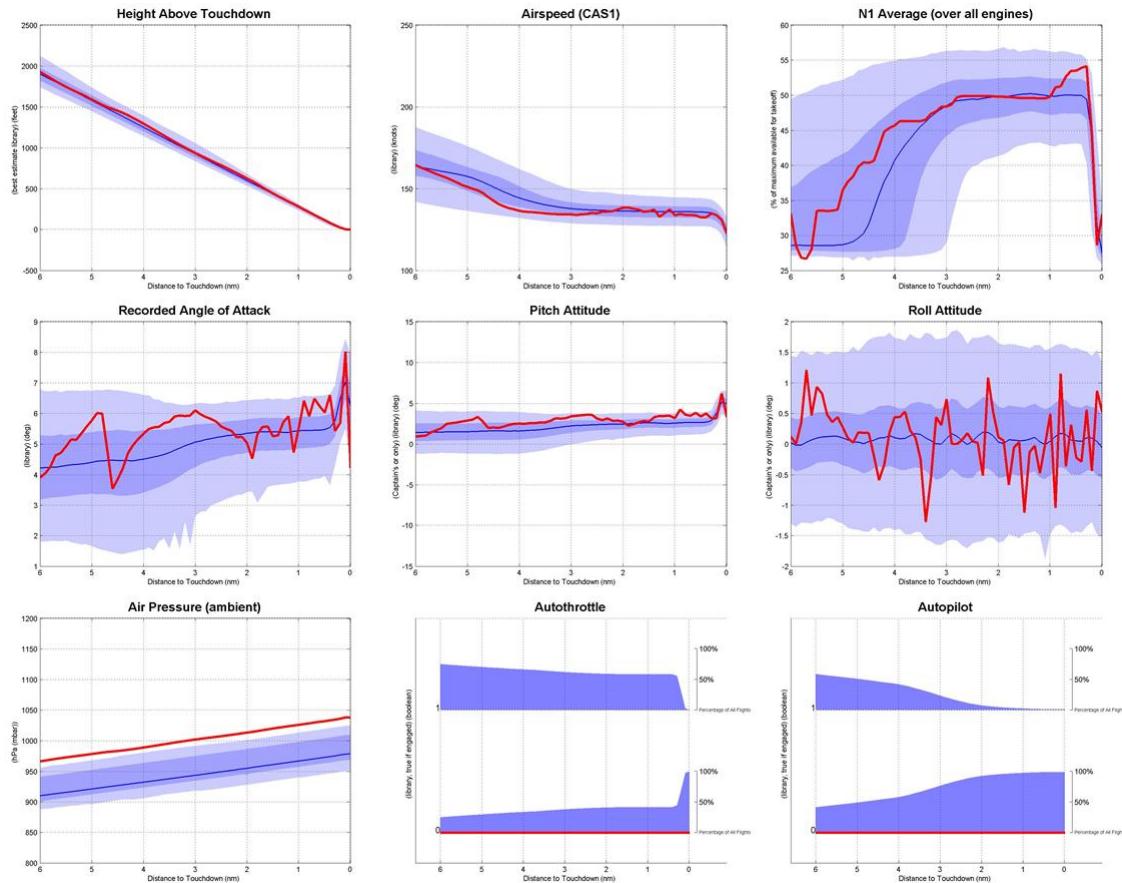


Figure 5.33 Flight 1209099 - manually-flown approach in cold weather

In summary, ClusterAD-Flight and ClusterAD-Data Sample were able to detect flights that were operationally abnormal. However, the type of anomalies detected differed. Flights detected by ClusterAD-Flight exhibited more pattern-based anomalies. These anomalies had abnormal data patterns over the entire phase of flight and were more likely to be recognized as operationally abnormal. In contrast, flights detected by ClusterAD-Data Sample exhibited instantaneous anomalies. The abnormal flight parameters had spikes in their data-series. However, these instantaneous anomalies were not perceived as abnormal by operational standards. Flights detected by both ClusterAD-Flight and ClusterAD-Data Sample were associated with a mix of pattern-based anomalies and instantaneous anomalies. These flights also had a higher inter-rater agreement, which indicates that a hybrid method with both ClusterAD-Flight and ClusterAD-Data Sample might be more reliable.

#### 5.4.4 Results on Data Visualization Tools

Using the standard usability questionnaire ASQ, three aspects of the data visualization tools were evaluated: ease of use, time required, and supporting information. Results are summarized in Figure 5.34, Figure 5.35 and Figure 5.36. Since the sample size is limited, no statistical tests were conducted. Only observations on the raw data are discussed here.

The Flight Abnormality Visualization made the reviewing task easier and less time consuming. As shown in Figure 5.34 and Figure 5.35, Configuration A and B had less responses in “Disagree” regarding the question: “I am satisfied with the ease of completing the tasks” and the question: ”I am satisfied with the amount of time it took to complete the tasks.”

The different types of flight parameter plot did not have an obvious impact on the ease of use and the time required. However, comparing Configuration A and Configuration B, some different opinions towards Configuration B were observed. It showed that some domain experts prefer a simpler graphic with only median plotted in Type II flight parameter plots, while some domain experts would like rich information on percentiles in Type I flight parameter plots.

The supporting documents for the data visualization tools were perceived equally helpful across difference configurations.

**Q: Overall, I am satisfied with the ease of completing the tasks in this scenario.**

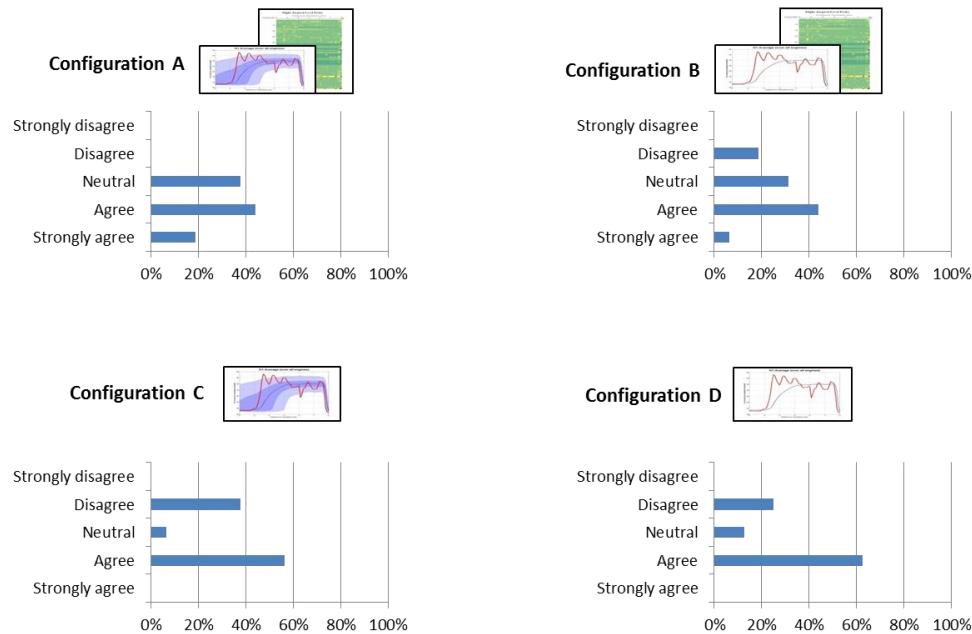


Figure 5.34 Results on “Ease of Use”

**Q: Overall, I am satisfied with the amount of time it took to complete the tasks in this scenario.**

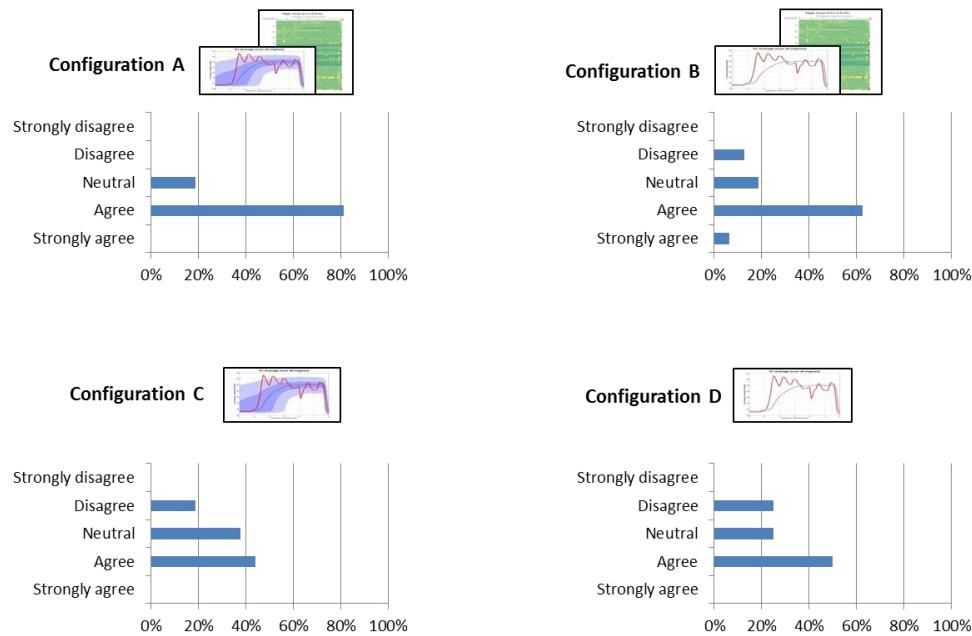


Figure 5.35 Results of “Time Required”

**Q: Overall, I am satisfied with the support information (on-line help, messages, documentations) when completing this task.**

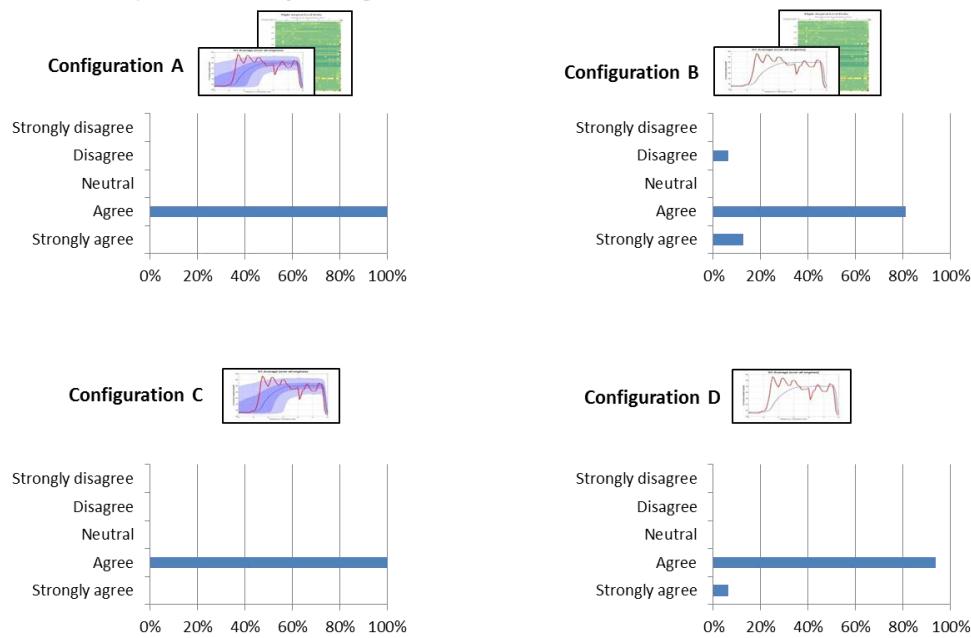


Figure 5.36 Results on “Supporting Information”

### 5.4.5 Study Summary

The results showed that both ClusterAD-Flight and ClusterAD-Data Sample were able to detect flights with operational significance, evaluated by domain experts. Some flights were agreed by all domain experts that they exhibited safety hazards. Thus, cluster-based anomaly detection algorithms could be used to detect flights with potential risks.

The strengths and weakness of ClusterAD-Flight and ClusterAD-Data Sample were revealed from the operational characteristics of flights detected by difference algorithms. ClusterAD-Flight was more sensitive to pattern-based anomalies with abnormal data patterns over the entire phase of flight. ClusterAD-Data Sample was more sensitive to instantaneous anomalies, which were often associated with spikes in the data-series of flight parameters.

Flights detected by ClusterAD-Flight were more likely to be perceived as operationally abnormal by domain experts. In contrast, flights detected by ClusterAD-Data Sample with instantaneous anomalies were less likely to be agreed as abnormal by all domain experts.

Data visualization tools were shown to be effective in facilitating the expert review process. Especially, the benefit brought by the Flight Abnormality Visualization tool was significant. The Flight Abnormality Visualization tool helped domain experts quickly identify the resources of anomalies, which flight parameters at what time contributed to the abnormality in a flight. Subjective evaluations showed that the review process was easier and less time consuming with the Flight Abnormality Visualization tool. This tool was enabled by ClusterAD-Data Sample.

Thus, a hybrid method with both ClusterAD-Flight and ClusterAD-Data Sample is expected to perform better. Flights detected by both ClusterAD-Flight and ClusterAD-Data Sample were associated with a mix of pattern-based anomalies and instantaneous anomalies. These flights also had a higher inter-rater agreement. ClusterAD-Flight and ClusterAD-Data Sample could complement each other by detecting different types of anomalies. More importantly, both algorithms could complement existing methods by identifying emerging issues that were not accounted for in the past.

# Chapter 6

# Conclusions and Future Work

## 6.1 Conclusions

This thesis developed a new anomaly detection approach to support proactive safety management using airline's daily operational FDR data. The new approach includes two components: cluster-based anomaly detection algorithms and domain experts review. The former identifies flights with abnormal data patterns from a large number of routine flights, while the latter evaluates detected abnormal flights to determine operational significance.

Two algorithms for cluster-based anomaly detection were developed to explore two types of cluster techniques: ClusterAD-Flight and ClusterAD-Data Sample. Both algorithms assume that a finite number of normal patterns exist in the FDR data of routine flights. These normal patterns can be recognized by cluster analysis, which are superior to existing methods since cluster-based anomaly detection algorithms can detect unknown issues without a specific definition of what "abnormal" is in advance. ClusterAD-Flight converts data of a flight for a specific phase into a form that is applicable for cluster analysis, while ClusterAD-Data Sample transforms data samples.

In addition, new data visualization tools were developed to facilitate review processes. Flight Parameter Plots generate informative graphics to present raw FDR data in front of domain experts; Flight Abnormality Visualization can assist domain experts in identifying sources of anomalies quickly. These visualization tools have been confirmed by domain experts as effective in supporting the review process.

Evaluation studies were conducted using airline FDR data. ClusterAD-Flight and ClusterAD-Data Sample were compared with Exceedance Detection, the current method in use by airlines, and MKAD, another anomaly detection algorithm developed at NASA, using a dataset of 25519 A320 flights. An evaluation of the entire detection approach including detection algorithms and data visualization tools was conducted with domain experts using a dataset of 10,528 A320 flights.

These evaluation studies resulted in several findings:

First, the cluster-based anomaly detection approach Sample was able to identify operationally significant anomalies that beyond the capacities of current methods; they could complement the current method of exceedance detection in order to identify new types of anomalies.

Second, the data visualization tools, especially Flight Abnormality Visualization, were found to be effective in supporting the review process. Flight Abnormality Visualization enabled domain experts to quickly identify the resources of anomalies, which reduced the time needed to review a flight.

Third, different data transformation techniques detected different types of anomalies. ClusterAD-Flight was more sensitive to pattern-based anomalies, which were abnormal patterns over a period of time and more likely to be perceived as operationally abnormal by domain experts. ClusterAD-Data Sample detected instantaneous anomalies better, which were less likely to be perceived as abnormal by domain experts. MKAD performed better at detecting abnormal sequences in discrete parameters.

Forth, standard operations of flights could be identified by cluster analysis. Multiple standard operations during takeoff phase could be identified by ClusterAD-Flight. Moreover, aircraft operational modes across flight phases could be identified by ClusterAD-Data Sample.

Lastly, a hybrid method with both ClusterAD-Flight and ClusterAD-Data Sample is expected to perform better. Flights detected by both ClusterAD-Flight and ClusterAD-Data Sample were associated with a mix of pattern-based anomalies and instantaneous anomalies. These flights also had a higher inter-rater agreement. ClusterAD-Flight and ClusterAD-Data Sample could complement each other by detecting different types of anomalies, and the two could work together to complement existing methods by identifying emerging issues that were not accounted for in the past.

In summary, the new cluster-based anomaly detection approach developed in this thesis provides a promising way to detect abnormal flights. It combines the strength of cluster-based algorithm with expert review and it can expose unknown safety concerns. This new approach can help airlines detect early signs of safety degradations, reveal latent risks, deploy predictive maintenance, and train staff accordingly.

## 6.2 Recommendations for Future Work

The new approach developed in this thesis is yet to be implemented in real-world operations and more research efforts are needed before it is integrated with current FOQA program. A set of comprehensive tools are needed to support expert review. Details of a review protocol need to be developed to address the subjective variations across domain experts.

The research can be extended to make the new approach applicable to all phases of flight. Currently, ClusterAD-Flight is limited to takeoff and approach phase and ClusterAD-Data Sample has only been tested to these two phases as well. Given the diversity of temporal patterns in other flight phases, this limit can be overcome by anchoring raw FDR data and making data patterns of different flights comparable.

Another direction is to develop reinforcement learning capabilities, where initial reviews can inform recurrent reviews with established baselines. Assuming the new approach has been adopted by airlines with reinforcement learning capability, review results from domain experts can be dynamically incorporated into baselines. In this way, abnormal flights of identical symptoms can be categorized automatically without repeated expert review.

A third direction is to analyze data from sources other than FDR data, such as maintenance records, pilot voluntary reports, and weather reports. Combing information from multiple sources allows airlines to upgrade anomaly detection method into a diagnostic system. It can easier identify abnormality with operational significance as well as its causes, thus answering the “what” and the “why” questions at the same time.

# Reference

- Abraham, B. (1989). Outlier Detection and Time Series Modeling. *Technometrics*, 31(2), 241. doi:10.2307/1268821
- Abraham, B., & Box, G. E. P. (1979). Bayesian Analysis of Some Outlier Problems in Time Series. *Biometrika*, 66(2), 229– 237.
- Aerobyes Ltd. (n.d.). Aerobyes - FDM/FOQA solutions. Retrieved February 2, 2013, from <http://www.aerobyes.co.uk/>
- Allan, J., Carbonell, J., & Doddington, G. (1998). Topic detection and tracking pilot study: Final report. *DARPA Broadcast News Transcription and Understanding Workshop*. (pp. 194– 218).
- Amidan, B., & Ferryman, T. (2000). APMS SVD methodology and implementation. *Battelle Report, PNWD-3026*.
- Amidan, B. G., & Ferryman, T. a. (2005). Atypical event and typical pattern detection within complex systems. *2005 IEEE Aerospace Conference* (pp. 3620– 3631). IEEE. doi:10.1109/AERO.2005.1559667
- Ankerst, M., Breunig, M. M., Kriegel, H. P., & Sander, J. (1999). OPTICS: ordering points to identify the clustering structure. *ACM SIGMOD Record* (Vol. 28, pp. 49– 60). ACM.
- Anscombe, F. J., & Guttman, I. (1960). Rejection of Outliers. *Technometrics*, 2(2), 123– 147.
- Baker, L. D., Hofmann, T., McCallum, A. K., & Yang, Y. (1999). A Hierarchical Probabilistic Model for Novelty Detection in Text. *NIPS' 99*.
- Bay, S. D., & Schwabacher, M. (2003). Mining distance-based outliers in near linear time with randomization and a simple pruning rule. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ' 03*, 29. doi:10.1145/956750.956758
- Bianco, A. M., Garcia Ben, M., Martinez, E. J., & Yohai, V. J. (2001). Outlier Detection in Regression Models with ARIMA Errors using Robust Estimates. *Journal of Forecasting*, 20(8), 565– 579. doi:10.1002/for.768
- Blender, R., Fraedrich, K., & Lunkeit, F. (1997). Identification of cyclone-track regimes in the North Atlantic. *Quarterly Journal of the Royal Meteorological Society*, 123(539), 727– 741. doi:10.1256/smsqj.53909
- Boeing Commercial Airplanes. (2012). *Statistical Summary of Commercial Jet Airplane Accidents. Worldwide Operations*.
- Bolton, R., & Hand, D. (2001). Unsupervised Profiling Methods for Fraud Detection. *Credit Scoring and Credit Control VII* (pp. 5– 7).
- Boussemart, Y., Las Fargeas, J. C., Cummings, M., & Roy, N. (2009). Comparing Learning Techniques for Hidden Markov Models of Human Supervisory Control Behavior. *AIAA*

*Infotech@Aerospace Conference*. Seattle, Washington: American Institute of Aeronautics and Astronautics.

- Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000). LOF: identifying density-based local outliers. *ACM SIGMOD Record*, 29(2), 93– 104.
- Budalakoti, S., Srivastava, A. N., & Akella, R. (2006). Discovering atypical flights in sequences of discrete flight parameters. *Aerospace Conference, 2006 IEEE* (pp. 1– 8). IEEE. doi:10.1109/AERO.2006.1656109
- Budalakoti, Suratna, Srivastava, A. N., & Otey, M. E. (2008). Anomaly detection and diagnosis algorithms for discrete symbol sequences with applications to airline safety. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 39(1), 101– 113. doi:10.1109/TSMCC.2008.2007248
- Byers, S., & Raftery, A. E. (1998). Nearest-neighbor clutter removal for estimating features in spatial point processes. *Journal of the American Statistical Association*, 93(442), 577– 584. doi:10.2307/2670109
- Cabrera, J., & Lewis, L. (2001). Detection and classification of intrusions and faults using sequences of system calls. *Acm sigmod record*, 30(4), 25. doi:10.1145/604264.604269
- CAE Flightscape. (n.d.-a). Flightscape INSIGHT™ ANALYSIS. Retrieved February 2, 2013, from <http://www.flightscape.com/product/service/insight-analysis/>
- CAE Flightscape. (n.d.-b). Flightscape INSIGHT™ FDM. Retrieved February 2, 1031, from <http://www.flightscape.com/product/service/insight-fdm/>
- Campbell, N. (2003). Flight Data Analysis – An Airline Perspective. *Australian and New Zealand Societies of Air Safety Investigators Conference (ANZSASI) Seminar 2003*. Maroochydore.
- Campbell, N. A. H. (2007). The Evolution of Flight Data Analysis. *Australian and New Zealand Societies of Air Safety Investigators Conference (ANZSASI) Seminar 2007*. Wellington.
- Chan, P. K., & Mahoney, M. V. (2005). Modeling multiple time series for anomaly detection. *Fifth IEEE International Conference on Data Mining (ICDM' 05)*, 90– 97. doi:10.1109/ICDM.2005.101
- Chandola, V., Banerjee, A., & Kumar, V. (2009). Outlier detection: A survey. *ACM Computing Surveys*, 1– 72.
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, 297, 273– 297.
- 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. *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 47– 56). ACM.
- Dempster, A., Laird, N., & Rubin, D. (1977). Maximum Likelihood fromIncomplete Data via the EMAlgorithm. *Journal of the Royal Statistical Society*, 39(1), 1– 38.
- Desforges, M., Jacob, P., & Cooper, J. (1998). Applications of probability density estimation to the detection of abnormal conditions in engineering. *Proceedings of the Institution of*

*Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, 212(8), 687– 703.  
doi:10.1243/0954406981521448

- Endler, D. (1998). Intrusion detection. Applying machine learning to Solaris audit data. *Proceedings. of 14th Annual Computer Security Applications Conference* (pp. 268–279). IEEE. doi:10.1109/CSAC.1998.738647
- Eskin, E. (2000). Anomaly detection over noisy data using learned probability distributions. *Proceedings of the Seventeenth International Conference on Machine Learning*. San Francisco, CA: Morgan Kaufmann Inc.
- Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the 2nd international conference on knowledge discovery and data mining* (Vol. 1996, pp. 226–231). Portland: AAAI Press.
- Federal Aviation Administration. (2004, April 1). Advisory Circular. 120-82 Flight Operational Quality Assurance. *Federal Aviation Administration*. Washington, DC.  
doi:10.1177/004728757301200242
- Federal Aviation Administration. (2006). Advisory Circular: Introduction to Safety Management Systems for Air Operators (AC120-92).
- Federal Aviation Administration. (2010a). Airline Safety and Federal Aviation Administration Extension Act of 2010.
- Federal Aviation Administration. (2010b). Advisory Circular: Safety Management Systems for Aviation Service Providers (AC120-92A).
- Federal Aviation Administration. (2011). *14 CFR 121, Appendix M - Airplane Flight Recorder Specifications*.
- Federal Aviation Administration. (2013). FAA and Industry Are Advancing The Airline Safety Act, But Challenges Remain to Achieve Its Full Measure (Audit Report AV-2013-037).
- Fox, A. J. (1972). Outliers in time series. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(3), 350–363.
- Gaber, M. M., Zaslavsky, A., & Krishnaswamy, S. (2005). Mining data streams: a review. *ACM Sigmod Record*, 34(2), 18–26.
- Gariel, M., Srivastava, A. N., & Feron, E. (2010). Trajectory Clustering and an Application to Airspace Monitoring. *Arxiv preprint arXiv:1001.5007*, 1–15.
- Gibbons, R. D., Bhaumik, D. K., & Aryal, S. (2009). *Statistical Methods for Groundwater Monitoring*. John Wiley & Sons, Inc.
- Golay, X., Kollias, S., Stoll, G., Meier, D., Valavanis, a, & Boesiger, P. (1998). A new correlation-based fuzzy logic clustering algorithm for fMRI. *Magnetic resonance in medicine : official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine*, 40(2), 249–60.
- Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics*, 11(1), 1–21.

- Guha, S., Rastogi, R., & Shim, K. (1998). CURE: an efficient clustering algorithm for large databases. *ACM SIGMOD Record* (Vol. 27, pp. 73–84). ACM.
- Hinneburg, A., & Keim, D. A. (1998). An efficient approach to clustering in large multimedia databases with noise. *Proceedings of 1998 Int. Conf. Knowledge Discovery and Data Mining (KDD'98)* (pp. 58–65). New York, NY.
- Hodge, V., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2), 85–126. doi:10.1023/B:AIRE.0000045502.10941.a9
- Hofmeyr, S. A., Forrest, S., & Somayaji, A. (1998). Intrusion detection using sequences of system calls. *Journal of Computer Security*, 6(3), 151–180.
- Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6), 417–441. doi:10.1037/h0071325
- Ihler, A., Hutchins, J., & Smyth, P. (2006). Adaptive event detection with time-varying poisson processes. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining KDD 06*, 207. doi:10.1145/1150402.1150428
- Ilgun, K., Kemmerer, R. A., & Porras, P. A. (1995). State Transition Analysis: A Rule-Based Intrusion Detection Approach. *IEEE Transactions on Software Engineering*, 21(3), 181–199. doi:10.1109/32.372146
- Iverson, D. L. (2004). Inductive system health monitoring. *Proceedings of The 2004 International Conference on Artificial Intelligence (IC-AI04)*.
- Jain, R., & Chilamtac, I. (1985). The P2 Algorithm for Dynamic Calculation of Quantiles and Histograms Without Storing Observations. *Communications of the ACM*, 28, 1076–1085.
- Keogh, E., Lonardi, S., & Ratanamahatana, C. A. (2004). Towards parameter-free data mining. *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '04*, 206. doi:10.1145/1014052.1014077
- Kriegel, H.-P., Kröger, P., & Zimek, A. (2009). Clustering high-dimensional data: A Survey on Subspace Clustering, Pattern-Based Clustering, and Correlation Clustering. *ACM Transactions on Knowledge Discovery from Data*, 3(1), 1–58. doi:10.1145/1497577.1497578
- Kumar, M., & Woo, J. (2002). Clustering Seasonality Patterns in the Presence of Errors Clustering Seasonality Patterns in the Presence of Errors. *the Eighth ACM International Conference on Knowledge Discovery and Data Mining* (pp. 557–563).
- Larder, B., & Summerhayes, N. (2004). Application of Smiths Aerospace Data Mining Algorithms to British Airways 777 and 747 FDM Data. *Washington, DC: FAA, Global Aviation Information Network*, (December).
- Laurikkala, J., Juhola, M., & Kentala, E. (2000). Informal identification of outliers in medical data. *The Fifth International Workshop on Intelligent Data Analysis in Medicine and Pharmacology*. Citeseer.
- Lewis, J. R. (1991a). Psychometric evaluation of an after-scenario questionnaire for computer usability studies: The ASQ. *SIGCHI Bulletin*, 23, 78–81.

- Lewis, J. R. (1991b). An after-scenario questionnaire for usability studies: psychometric evaluation over three trials. *SIGCHI Bulletin*, 23, 79.
- Lewis, J. R. (1995). IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction*, 7, 57–78.
- Liao, T. W. (2005). Clustering of time series data—a survey. *Pattern Recognition*, 38(11), 1857–1874. doi:10.1016/j.patcog.2005.01.025
- Logan, T. J. (2008). Error prevention as developed in airlines. *International journal of radiation oncology, biology, physics*, 71(1 Suppl), S178–81. doi:10.1016/j.ijrobp.2007.09.040
- Maille, N. P., & Statler, I. C. (2009). *Comparative Analyses of Operational Flights with AirFASE and The Morning Report Tools* (No. NASA/TM-2009-215379). Moffett Field, California.
- McLachlan, G. J., & Basford, K. E. (1988). *Mixture Models: Inference and Applications to Clustering*. New York: Marcel Dekker.
- Möller-Levet, C., Klawonn, F., Cho, K. H., & Wolkenhauer, O. (2003). Fuzzy clustering of short time-series and unevenly distributed sampling points. *LNCS, Proceedings of the IDA2003* (pp. 330–340). Springer.
- Ng, R. T., & Han, J. (1994). Efficient and Effective Clustering Data Mining Methods for Spatial Data Mining. *the 20th International Conference on Very Large Data Bases* (pp. 144–155). Chile.
- NTSB. (2009). FDR Group Chairman's Factual Report, DCA09MA027.
- Pearson, K. (1896). Mathematical contributions to the theory of evolution. III. Regression, heredity and panmixia. *Philosophical Transactions of the Royal Society of London. Series A*, 187, 253–318.
- Reynolds, D. A. (2008). Gaussian Mixture Models. *Encyclopedia of Biometric Recognition, Springer*, (2), 659–663. doi:10.1007/978-0-387-73003-5\_196
- Rodgers, J. L., & Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1), 59–66.
- Rosenkrans, W. (2008, January). Fade-Free Memory. *Aerosafe World*, (January), 47–48.
- Rosner, B. (1983). Percentage Points for a Generalized ESD Many-Outlier Procedure. *Technometrics*, 25(2), 165–172.
- Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. John Wiley & Sons.
- Sagem. (n.d.). Sagem Analysis Ground Station. Retrieved February 2, 2013, from <http://ags.sagem-ds.com/en/site.php?spage=02010200>
- Sagem. (2013). Flight Data Interface and Management Unit.

- Schwabacher, M., & Oza, N. (2007). Unsupervised Anomaly Detection for Liquid-Fueled Rocket. *Propulsion Health Monitoring, AIAA InfoTech Aerospace Conference, Rohnert Park* (Vol. 6, pp. 464–482). doi:10.2514/1.42783
- Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2), 461–464.
- Sekar, R., Gupta, A., Frullo, J., Shanbhag, T., Tiwari, A., Yang, H., & Zhou, S. (2002). Specification-based anomaly detection: a new approach for detecting network intrusions. *Proceedings of the 9th ACM conference on Computer and communications security*, 265–274. doi:10.1145/586110.586146
- Smyth, P. (1994). Markov monitoring with unknown states. *IEEE Journal on Selected Areas in Communications*, 12(9), 1600–1612. doi:10.1109/49.339929
- Srivastava, A. N. (2005). Discovering system health anomalies using data mining techniques. *Proceedings of the 2005 Joint Army Navy NASA Airforce Conference on Propulsion* (pp. 1–11).
- Stefansky, W. (1972). Rejecting outliers in factorial designs. *Technometrics*, 14(2), 469–479.
- Stigler, S. M. (1989). Francis Galton's Account of the Invention of Correlation. *Statistical Science*, 4(2), 73–79.
- Tax, D. M. J., & Duin, R. P. W. (1999). Support vector domain description. *Pattern Recognition Letters*, 20(11-13), 1191–1199. doi:10.1016/S0167-8655(99)00087-2
- Teledyne Controls. (2011). *Wireless GroundLink System*.
- Treder, B., & Craine, B. (2005). *Application of Insightful Corporation's Data Mining Algorithms to FOQA Data at JetBlue Airways: A Technology Demonstration in Partnership with the Federal Aviation Administration and the Global Aviation Information Network (GAIN)*. *Global Aviation Information Network*. Seattle, WA.
- Vlachos, M., Lin, J., Keogh, E., & Gunopulos, D. (2003). A wavelet-based anytime algorithm for k-means clustering of time series. In *Proc. Workshop on Clustering High Dimensionality Data and Its Applications*.
- Warrender, C., Forrest, S., & Pearlmuter, B. (1999). Detecting Intrusions using System Calls: Alternative Data Models. *IEEE Symposium on Security and Privacy*, pages(c), 133–145. doi:10.1109/SECPRI.1999.766910
- Zhang, T., Ramakrishnan, R., & Livny, M. (1996). BIRCH: an efficient data clustering method for very large databases. *ACM SIGMOD Record* (Vol. 25, pp. 103–114). ACM.