Knowledge Discovery Within ADS-B Data from Routine Helicopter Tour Operations

Hsiang-Jui Chin*, Alexia P. Payan†, and Dimitri N. Mavris‡
School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, GA, 30332-0150, USA

Charles C. Johnson§
Federal Aviation Administration, Atlantic City International Airport, NJ 08405, USA

Knowledge discovery or data mining techniques are widely used for anomaly detection in the commercial aviation domain to retrospectively improve operational safety. However, in the general aviation domain, especially for rotorcraft, analyses of flight data records for anomaly detection are not as prevalent. In this study, ADS-B data from a helicopter tour operator will be used to develop a prototype framework for uncovering patterns from routine flights. The ADS-B data contains two types of information: 1) time series of various flight parameters and 2) trajectory parameters. Various knowledge discovery techniques able to handle the aforementioned data types are explored and a few promising methods are applied to the ADS-B data of a helicopter tour operator in Hawaii. From the clustering results, patterns in the flight data records can be observed and can then be used by Subject-Matter Experts (SMEs) to facilitate the detection of anomalies. With this framework in place, rotorcraft operators will be able to analyze their routine flight data to not only monitor the safety of their operations but also to acquire knowledge on their operational patterns.

I. Nomenclature

$ADS-B$	=	Automatic Dependent Surveillance – Broadcast
$ASIAS$	=	Aviation Safety Information Analysis and Sharing
$CLARA$	=	Clustering LARge Application
$ClusterAD$	=	Clustering Anomaly Detection
$DBSCAN$	=	Density-Based Spatial Clustering of Applications with Noise
$DMKD$	=	Data Mining and Knowledge Discovery
ED	=	Exceedance Detection
EMS	=	Emergency Medical Services
FAA	=	Federal Aviation Administration
$FOQA$	=	Flight Operations Quality Assurance
$FPCA$	=	Functional Principal Component Analysis
HAI	=	Helicopter Association International
$HFDM$	=	Helicopter Flight Data Monitoring
HOG	=	Histogram of Oriented Gradient
IMS	=	Inductive Monitoring System
MSL	=	Mean Sea Level
$nLCS$	=	Normalized Longest Common Subsequences
$NTSB$	=	National Transportation Safety Board
$PEGASAS$	=	Partnership to Enhance General Aviation Safety, Accessibility and Sustainability

*Graduate Research Assistant, Aerospace Systems Design Laboratory, School of Aerospace Engineering, 270 Ferst Drive, Atlanta, GA, 30332-0150, AIAA Member.

†Research Engineer II, Aerospace Systems Design Laboratory, School of Aerospace Engineering, 270 Ferst Drive, Atlanta, GA, 30332-0150, AIAA Member.

‡S.P. Langley NIA Distinguished Regents Professor, Boeing Regents Professor of Advanced Aerospace Systems Analysis, School of Aerospace Engineering, Director Aerospace Systems Design Laboratory, 270 Ferst Drive, Atlanta, GA, 30332-0150, AIAA Associate Fellow.

§General Engineer, System Safety Section, Aviation Research Division, NextGen WJHTC Office, Federal Aviation Administration, William J. Hughes Technical Center, Atlantic City International Airport, NJ 08405.

SDF	=	Symbolic Dynamic Filtering
SME	=	Subject-Matter Expert
SPC	=	Statistical Process Control
$USHST$	=	Unites States Helicopter Safety Team

II. Introduction

Historically, rotary-wing aircraft have experienced higher accident and incident rates compared to fixed-wing aircraft, both in the commercial and the general aviation domains. In addition, safety reports from the U.S. Helicopter Safety Team (USHST) show that, although the number of fatal helicopter accidents has been decreasing steadily since the 1990s, it has started increasing again since 2015 [1]. Several studies of helicopter accidents have actually shown that between 60% and 85% of the rotorcraft accidents analyzed were due to pilots' actions and errors [2],[3]. As such, in order to improve the safety of rotorcraft operations, one really needs to understand what is happening in the cockpit. In fact, the National Transportation Safety Board (NTSB) has been recommending the use of flight data recorders onboard helicopters since 2015*. However, a large proportion of helicopters, mostly those belonging to one individual or to small-size companies, and operating across various missions, still do not have flight data recorders onboard.

In the United States, several entities such as the USHST and the Helicopter Association International (HAI) are collaborating with universities, government, and/or industry partners to raise awareness about the need for information systems based on flight data records, called Helicopter Flight Data Monitoring (HFDM) systems, to improve the safety of helicopter operations independently of the mission and the size of the operation. One such collaboration effort involves HAI, the Federal Aviation Administration (FAA) and several university and industry partners through the Partnership to Enhance General Aviation Safety, Accessibility and Sustainability (PEGASAS). More specifically, the goal of this FAA Center of Excellence is to develop an HFDM program for rotary-wing aviation similar to the Aviation Safety Information Analysis and Sharing (ASIAS) developed for fixed-wing commercial aviation†. In this HFDM program for ASIAS, participating helicopter operators voluntarily provide routine flight data records to HAI on the premises of a Memorandum of Understanding (MOU). The data is then stored in a secure database, de-identified, and provided to the PEGASAS research team at the Georgia Institute of Technology who is in charge of prototyping data analysis capabilities. Such capabilities include risk assessments, rotorcraft performance models, parameter exceedance analyses, safety metrics, and various visualizations of interest to operators involved [4],[5],[6],[7],[8]. The resulting prototype algorithms are then implemented on an online platform hosted by HAI where the operators enrolled in the program can run the various analyses on their flight data records and get insight into the safety of their operations with respect to their own standard operating procedures or other rotorcraft operators within the same mission segment or across the industry. This process is illustrated in Fig. 1.

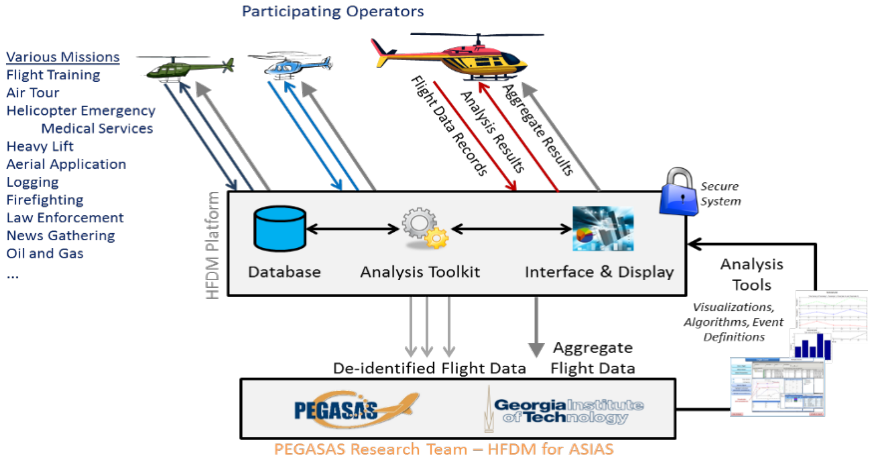


Fig. 1 HFDM for ASIAS Concept of Operations

*https://www.ntsb.gov/safety/mwl/Pages/mwl_archive.aspx

†https://www.pegasas.aero/projects/project-2-rotorcraft-asias

Although exceedance-based analyses allow operators to retrospectively detect anomalies in their operations, they require the pre-definition of flight parameter thresholds. In this study, the goal is to be able to identify patterns in flight data records without specifying parameter thresholds. Eventually, such pattern recognition may be translated into anomaly detection after subject-matter experts (SME) provide some input into what is an anomaly and what is not. In this study, we will use Automatic Dependent Surveillance – Broadcast (ADS-B) data from a helicopter tour operator and group flight data records into several categories based on features that may be extracted from the data. In the following sections, we will briefly go over some literature relevant to the detection of anomalies in flight data records. Based on several observations, a process for analyzing ADS-B data in the context of knowledge discovery will be proposed and various candidate techniques used in the process will be introduced in the methodology section. To test the validity of the proposed approach, an ADS-B dataset from a tour operator will be used and cleaned, and several experiments applied to both the altitude and trajectory data will be performed. The more relevant techniques that facilitate knowledge discovery in routine flight data records will be summarized in the last section and future avenues for research will be suggested.

III. Literature review

In this section, we will review various techniques for anomaly detection in flight data records from commercial fixed-wing aviation. To the best of the authors' knowledge, there are not many data mining methods applied to flight data records from general aviation, especially rotary-wing, potentially due to their heterogeneous nature and limited availability. Since flight data records are essentially a multivariate time series, it is worthwhile reviewing some recent advancements in time-series clustering.

1. Anomaly detection techniques for flight data records in commercial fixed-wing aviation

Amidan et al. [9] developed a tool named "Morning report" which uses mathematical signatures with k-means clustering to find anomalies. First, the flight data records are truncated into homogeneous segments, which correspond to phases of flight. Then, for a specific flight phase, the multivariate time series are transformed into mathematical signatures clustered using the k-mean algorithm and given an atypicality score based on principal component analysis. A metric called "global atypicality score" was proposed to address the multiple flight phases in a standard flight sequence. The combination of the p-value of the atypicality score and the cluster membership score determines whether the flight of interest is in a larger or a smaller cluster. All the flights are finally ranked based on this metric so subject-matter experts can identify the root cause of potential anomalies. Budalakoti et al. [10] proposed a method called "SequenceMiner" to detect anomalies within discrete flight data, such as switch signals from the cockpit. The distance measure used to compare two sequences was the "normalized longest common subsequence" (nLCS) and outliers were detected using Clustering LARge Application (CLARA). Das et al. [11] developed a framework named "data mining and knowledge discovery" (DMKD) which combined Symbolic Dynamic Filtering (SDF) for feature construction and iOrca for outlier detection. The SDF is used to transform continuous data into discrete data which is then fed into iOrca to retrieve the anomalies. Iverson [12] used a statistically oriented and data-driven approach for developing the "Inductive Monitoring System" (IMS). IMS is a supervised learning method where training data from nominal operations are required to construct a knowledge database, i.e. a set of nominal data clusters. Once constructed, the database can be queried to determine if a newly observed data point belongs to one of the nominal clusters or not. If it does not, then it is anomalous. However, without the availability of a normal dataset, it would be difficult to construct the knowledge database. Das et al. [13] developed a method called "Multiple Kernels Anomaly Detection" (MKAD) which uses a one-class support vector machine (SVM) with kernel functions for both continuous and discrete data. They found that more anomalous flights can be retrieved if both types of data are included in the analysis. Li et al. [14] proposed an approach named "ClusterAD" which features principal component analysis for feature extraction and DBSCAN for clustering analysis. This methodology was compared with the exceedance detection (ED) and MKAD. Results showed that ClusterAD and MKAD outperform ED in identifying operational significant anomalies. Further, ClusterAD worked well for continuous type of flight data while MKAD was better able to deal with discrete sequence data. Puranik et al. [15] worked on flight data records for fixed-wing general aviation and used energy-based metrics to represent the flights. The proposed method was able to detect anomalies in both artificial and real flights in the approach phase. Chu et al. [16] used a method based on the idea of multivariate statistical process control (SPC) to find anomalies for flights in the cruise phase. The Hotelling T^2 statistics was used as the metric for the control chart. To test the methodology, simulated flights with fault injection were analyzed and it was shown that 80% of the anomalous flights were identified.

2. Time-series clustering

In [17], Liao did a comprehensive survey of techniques used in time-series clustering. Several standard methods used for static data clustering were reviewed and different routes for migrating from static data to time-series clustering were explained. Unless the raw data of the time series is used directly, additional steps such as feature extraction or time-series modeling are required to gain representative information from the time series before clustering. In [18], Aghabozorgi et al. carried out a detailed review on time-series clustering applications in a variety of domains such as biology and finance. As the authors pointed out, there were two main directions that researchers had placed their efforts on. One was to transform high-dimensional data into low-dimensional data so that traditional clustering algorithms can be directly applied. The other was to measure the similarity between two time series. Both the representation methods and the distance measures mentioned in the review were summarized in [18].

3. Observations

The previous review of anomaly detection methods in flight data records showed that the majority of the analyses focused on flight data coming from fixed-wing commercial aviation rather than from general aviation, let alone from the rotary-wing domain. In general, fixed-wing aircraft flights follow a more routine pattern compared to rotary-wing general aviation flights. Indeed, due to the ability of rotorcraft to hover and their capability to perform diverse operations, it is challenging to identify nominal and anomalous patterns in the corresponding flight data records. Furthermore, the aforementioned analyses were performed on the basis of the same phases of flight. The analysis would start with grouping flights or flight segments based on their commonality and then anomalous flights would be identified if they are dissimilar to the norm. The notion of 'homogeneous segment' is not restricted to only the concept of phases of flight but can be extended to other contexts like the type of operations. For example, approach segments from a helicopter tour operator and approach segments form an helicopter Emergency Medical Services (EMS) operator may exhibit very dissimilar patterns due to the difference in the type of missions performed and/or standard operating procedures of the given operators.

There are two main components observed in the literature for the process of flight data analysis: 1) feature extraction; and 2) supervised or unsupervised learning. The goal of the feature extraction step may be to reduce the dimensionality of the original variable space, to generalize the original variables as metrics, or to transform the original continuous variables into discrete variables. Depending on the availability of nominal flight data records, supervised or unsupervised techniques can be selected for the prediction or pattern discovery step. From the perspective of supervised learning, anomaly detection is typically built under the assumptions that the nominal operation is known and a decent amount of nominal data is available. If these assumptions are valid, supervised learning is typically able to catch anomalous flights accurately without the intervention of SMEs. However, without prior information about nominal operations, unsupervised learning techniques may be adopted to discover knowledge in the data. This is a more difficult task compared to the use of supervised learning methods, especially in the case of data scarcity. Results from the application of unsupervised learning techniques might not be directly linked to safety hazards. Typically, SME feedback is required in those cases to close the loop and verify that the identified flights are truly anomalous.

IV. Methodology

Based on insights from the reviewed literature, we proposed the process shown in Fig. 2 to discover knowledge (and eventually anomalies) in flight data records from helicopter operations. In this study, this methodology will be applied to ADS-B-type data. The first step of the process is 'data pre-processing' which is optional if the data received is well-organized, is not too noisy, and does not contain errors such as missing parameter values, Not-A-Numbers (NANs), or empty data. If some data is missing or incomplete, statistical inference or data cleaning is used in this step. Then, flight data records may be grouped into three different categories: 1) time-series data; 2) trajectory data; and 3) sequence data. The first two categories apply to continuous data while the third one applies to discrete data. Flight parameters changing with time such as *altitude* can be considered a time series. The trajectory data is a combination of latitude and longitude (GPS) positions without the time information. The sequence data refers to a sequence of categorical variables like the switch signals in the cockpit or phases of flight. Due to the absence of sequence data in the available ADS-B dataset for this study, we will only focus on the time series and trajectory data in the remaining of this paper. Regarding the feature extraction step, different approaches will be applied to get a low-dimensional representation of the time series and trajectory data. Once those features have been identified, standard clustering analysis techniques may potentially be applied to find patterns in helicopter flight data records. The methods used in this study will be explained

briefly in following subsections.

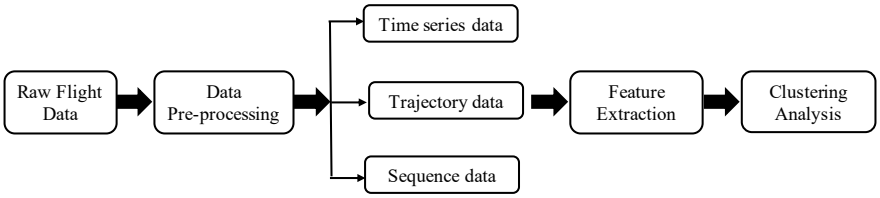


Fig. 2 Process flow for helicopter flight data analysis

A. Feature extraction

1. B-spline

B-spline is a method used for curve-fitting, where the "B" stands for the *basis*. This method constructs a piecewise polynomial function that fits to a number of input points. When compared with the conventional cubic splines, B-splines have the advantage of avoiding the multi-collinearity issue of the basis functions. A curve $S(t)$ can be approximated by its n^{th} order B-spline representation as in Eq. (1).

$$S(t) \approx \sum_i \alpha_i B_{i,n}(t) \tag{1}$$

where i is a variable related to the number of knots selected in the time domain, α_i are the B-spline coefficients, and $B_{i,n}(t)$ are the B-spline basis functions.
The B-spline basis functions $B_{i,n}(t)$ can be calculated using *Cox-de Boor recursive formula* as in Eq. (2) and Eq. (3).

$$B_{i,1}(t) = \begin{cases} 1, & \text{if } t_i \le t < t_{i+1} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

$$B_{i,k+1}(t) = \frac{t - t_i}{t_{i+k} - t_i} B_{i,k}(t) + \frac{t_{i+k+1} - t}{t_{i+k+1} - t_{i+1}} B_{i+1,k}(t) \tag{3}$$

The B-spline coefficients α_i can be estimated using the least-square formula and they can serve as the low-dimensional features of the time series.

2. Functional principal component analysis

Assume the observed time series can be expressed as in Eq. (4).

$$S_i(t) = \mu(t) + \epsilon_i(t) \tag{4}$$

where $S_i(t)$ is the i^{th} observed signal, $\mu(t)$ is the mean signal, and $\epsilon_i(t)$ is the residual of the i^{th} observed signal.
Through Karhunen-Loeve theorem, the residual can be represented as an infinite sum of multiplications between functional principal component (FPC) scores and their corresponding eigenfunctions, which is expressed in Eq. (5).

$$\epsilon_i(t) = \sum_{k=1}^{\infty} \xi_{ik} \phi_k(t) \tag{5}$$

where ξ_{ik} are the FPC scores and $\phi_k(t)$ are the eigenfunctions.
In the FPCA, a continuous curve or profile, which is essentially an infinite-dimensional entity, can be approximately represented by several FPC scores through the eigen-decomposition of the covariance function. The eigen-decomposition is shown in Eq. (6).

$$\int_0^M \hat{C}(t, t') \hat{\phi}_k(t) dt = \hat{\lambda}_k \hat{\phi}_k(t') \tag{6}$$

where $\hat{C}(t, t')$ is the estimated covariance function, $\hat{\lambda}_k$ is the k^{th} eigenvalue, and $\hat{\phi}_k(t)$ is the eigenfunction corresponding to the k^{th} eigenvalue.
With the estimated eigenfunction $\hat{\phi}_k(t)$, the FPC scores for any new observed curves can be calculated using Eq. (7).

$$\xi_{ik} = \int_0^M \epsilon_i(t)\hat{\phi}_k(t)dt \tag{7}$$

The ξ_{ik}'s are used as the low-dimensional representation of the time series data.

3. Histogram of oriented gradient

The Histogram of Oriented Gradient (HOG) is a popular feature descriptor of images used in the field of computer vision that was developed by Dalal and Triggs [19]. It extracts relevant information out of a high-dimensional image while keeping the dimension of the feature relatively low. This method had been applied to many object detection tasks including handwritten digits recognition [20],[21]. Since the graphical representations of the trajectory data (i.e. flight paths in the latitude/longitude domain) look in fact quite similar to images of handwritten digits, this method can potentially be used for extraction of features from the trajectory data. To construct the HOG feature, the image is first separated into several cells in a grid fashion. Each cell contains several pixels and for each pixel, the magnitude and the orientation of the gradient to the surrounding pixels are calculated and stored. Then, the gradient information is grouped using a histogram and this histogram is the feature representing the corresponding portion of the image. Once all the HOG features are retrieved for all the cells in the image, a block which contains multiple cells slides through the image to gather aggregate information of the HOG features.

B. Clustering analysis

1. K-means clustering

K-means clustering is an iterative algorithm for finding an appropriate grouping of the data. With prior knowledge of the number of clusters k in mind, it initializes k cluster centers randomly in the feature space. Each data point is assigned to the closest cluster center using a given distance measure and the location of the cluster center is then updated based on the relative positions of all the points in that specific cluster to the current center. This two-step process is repeated until no change in cluster assignment is observed or the maximum number of iterations is reached. Although it is a rather efficient algorithm, there are some issues with this approach. First, the user has to specify the number of clusters in advance and sometimes it is difficult to pick the right k number especially in high-dimensional spaces. Second, the result of the clustering may not be deterministic if no clear pattern can be observed in the dataset. Due to the randomness in the aforementioned initialization process, cluster assignment using k-means clustering might change every time the algorithm is run.

2. Hierarchical clustering

Hierarchical clustering is a sequential process which can either start from a state in which each data point serves as its own cluster or a state in which all data points are in one cluster. The bottom-up method is called *agglomerative clustering* while the top-down method is called *divisive clustering*. For the agglomerative clustering, the decision on combining two different clusters is based on a measure called "linkage". If two clusters are close to each other compared to other existing clusters in terms of "linkage", they are merged as one cluster. Kassambara and Alboukadel [22] summarize popular linkages used in practice such as the complete linkage and Ward's methods. The result of the hierarchical clustering is typically represented in a dendrogram, which is a tree-like structure showing the progression of the clusters merging process. To form clusters in the dataset, the dengrogram is cut based on the chosen number of clusters. There are different metrics for picking the optimal number of clusters and they will be mentioned in the following subsection.

3. Clustering tendency

Before a clustering analysis can be performed, it is essential to determine the existence of cluster-like patterns in the data. Hopkins et al. [23] proposed a metric named *Hopkins statistic H* to distinguish patterns observed in the data. If H is larger than 0.5 and close to 1, it means that the data is highly clusterable. If H is closer to 0.5, the data would appear as randomly distributed.

4. Indices for finding the optimal number of clusters

There are various techniques for finding the optimal number of clusters. For example, the elbow method is a heuristic approach which defines the location where the slope of the "within-group" variation curve changes abruptly as the optimal number of clusters. Gap statistics [24], which is based on statistical testing, is another method for obtaining the optimal number of clusters. Although there exists many indices for determining the number of clusters, we decided to use *average silhouette* [25] for its simplicity and interpretability. Conceptually, this index minimizes the intra-cluster variation while also maximizing the inter-cluster variation. The silhouette of a data point i can be expressed as in Eq. (8).

$$s(i) = \frac{b(i) - a(i)}{\max\left[a(i), b(i)\right]} \tag{8}$$

where $a(i)$ is the average distance to the points within the same cluster and $b(i)$ is the distance to the nearest cluster. Larger values of the average silhouette correspond to data points inside each cluster being more compact and clusters being farther apart from each other.

V. Application

In this section, we will briefly describe the flight data used in this study and mention the challenges and assumptions related to this dataset. A pre-processing procedure suitable for the dataset will be discussed: it consists in organizing the data into a consistent structure before conducting the analysis. The analysis will focus on one-dimensional profile data such as altitude and trajectory data (GPS position). Both synthetic and real flight data will be used to test the methods mentioned in previous sections.

A. Description of the flight data records

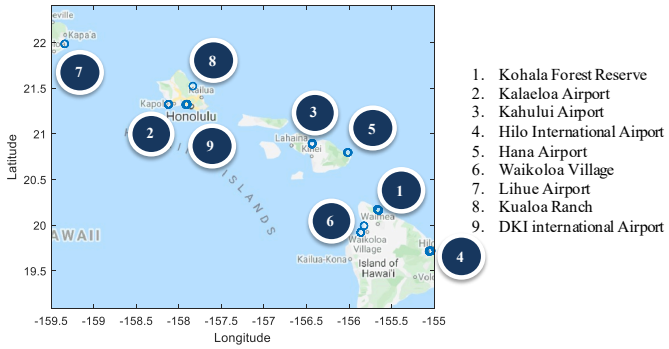
In this study, we use ADS-B data from a tour operator. Typical missions for this operator are sightseeing and air taxi. The ADS-B dataset contains a year-worth of operations across four helicopters. The variables included in the ADS-B dataset are latitude, longitude, altitude (MSL), and airspeed components in the north / east / up coordinate system. The data is recorded on daily basis but not the fleet is not always operating at the same time. Several issues were identified in this dataset. First, the resolution of the altitude data is rather low. The altitude data is recorded in multiples of 25 feet, thus small changes in altitude (less than 25 feet) would not be detected. The altitude profile resembles a step function rather than a smooth curve. Second, the sampling rate of the data records is not constant which means that the time step between each timestamp is not the same for all the recorded parameters. This might be caused by some interferences between the broadcasting device and the ground station or by uncalibrated devices. This leads to the third issue which is that of incomplete flight data records. A typical flight would takeoff from the runway / helipad and land at the same spot where the flight started or another airport based on its mission. However, some flights in the dataset would start or end at cruising altitude and they are thus most probably not completely recorded. The last issue is that there are in fact multiple flights or missions recorded in a single file for a full day of operations. Therefore, a pre-processing step is required to truncate the data into single flights/missions rather than directly analyze the day-based data structure.

Some assumptions were made regarding the ADS-B dataset. First, the data received is coming from calibrated devices and the errors associated with the data are minimal. Second, if an appropriate grouping is found for the truncated dataset, flight data in each group should reflect the true operations of the helicopter fleet of interest.

B. Pre-processing

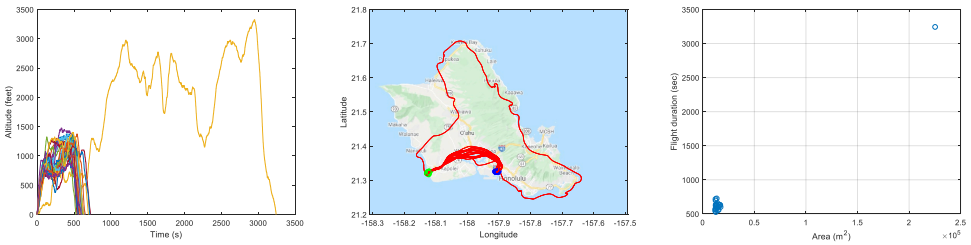
In this section, we will explain how the ADS-B dataset was pre-processed before proceeding to the analysis step. There should be a certain degree of flexibility for wrangling the data. In this section, we provide a general pre-processing procedure based on our own perspective. The procedure consists of two steps: 1) data filtering and segmentation; 2) grouping of flights with similar patterns. For the data filtering and segmentation, we first sort the data in chronological order and then filter out low altitude operations (less than 100 feet) and short duration flights (less than 3 minutes). To account for cases where multiple flights/missions are lumped together in the same file, we segment the files into different entities if there exists a time gap larger than 3 minutes between what appears to be two successive missions. The completeness of the segmented entities is ensured by examining if the flight starts/ends at or close to facilities that are suitable for takeoff/landing. Incomplete flights are discarded since it is difficult to infer what happened during flight for the missing data points. All the flights within the complete set are checked to make sure that there is not any time step larger than a threshold value. In this study, a threshold value of 10 seconds was selected and any flights with

a time step larger than 10 seconds is not included in the final set of flights for the analyses. In order to group flights with similar patterns, the pairs of takeoff and landing sites are compared, and flights having the same pair of start and end points are grouped together. Indeed, it is quite natural to group flights based on this feature for air tour/air taxi operations. Fig. 3 shows that there are 9 distinct locations for takeoff and landing sites for this dataset. A transition table can be created to count the frequency of flights from one site to another. Fig. 4 shows data traces of a group of 25 flights from the Daniel K. Inouye International airport (DKI/HNL) to the Kalaeloa airport (JRF). At first glance, it is obvious that there are two patterns associated with this group. One is a longer flight that toured around the island and the other is an air-taxi like transport between the two airports. They can be easily separated if we project the flights in two features, namely *covered area* and *flight duration*. In this scenario, we will exclude the longer flight and proceed with the remaining 24 flights for the altitude and trajectory analysis.



Takeoff sites \ Landing sites	1	2	3	4	5	6	7	8	9
1	0	0	0	0	0	63	0	0	0
2	0	1	0	0	0	0	0	0	52
3	0	0	44	0	15	4	0	0	1
4	6	0	0	73	0	5	0	0	0
5	0	0	19	0	0	0	0	0	0
6	6	0	1	8	0	18	0	0	0
7	0	0	0	0	0	0	3	0	1
8	0	0	0	0	0	0	0	0	3
9	0	25	0	0	0	0	1	1	15

Fig. 3 Flights grouped by takeoff and landing site pairs



Altitude traces **Flight trajectories** **Area versus flight duration**

Fig. 4 Flights from Daniel K. Inouye International airport to Kalaeloa airport

C. Altitude data analysis

1. Synthetic data

Before tackling the real flight altitude data, synthetic altitude traces were generated to test the methods mentioned in the previous sections. Three dissimilar altitude samples were drawn from the dataset and the kriging method with different numbers of knots was applied to generate the synthetic altitude data. With this approach, the synthetic altitude is guaranteed to pass through the knot points in a similar way as an interpolation method while creating small perturbations from the sample altitude away from the knot points. In Fig. 5, synthetic altitude traces for different groups

of data are displayed in different colors (blue, red, and black) and only small variations are observed within the same family of curves. Two feature extraction methods, B-spline and FPCA, were used to reduce the dimensionality of the original dataset. The extracted features are visualized in Fig. 6. The Hopkins statistics for both features are 0.767 and 0.943 respectively. This means that these features are highly clusterable. With average silhouette as the measure for determining the optimal number of clusters, either the k-means or the hierarchical clustering method with both extracted features is able to provide the same clustering results as that obtained from the true labels that we have entire knowledge about given that we are dealing with synthetic data.

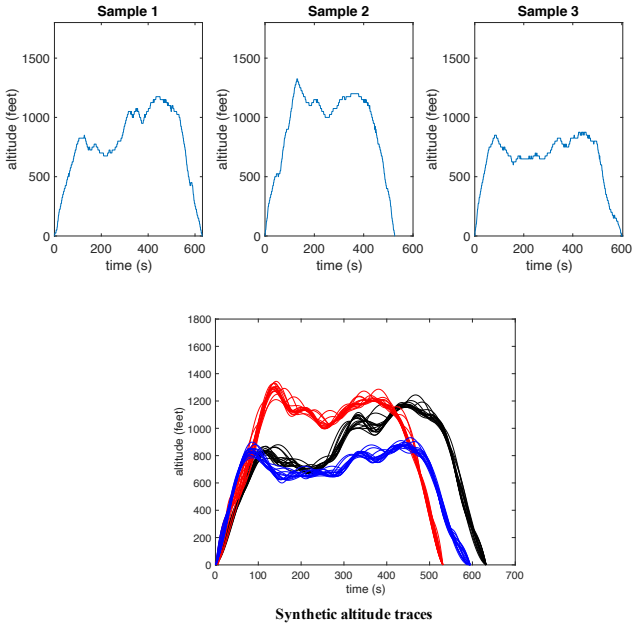


Fig. 5 Synthetic altitude data from 3 different samples

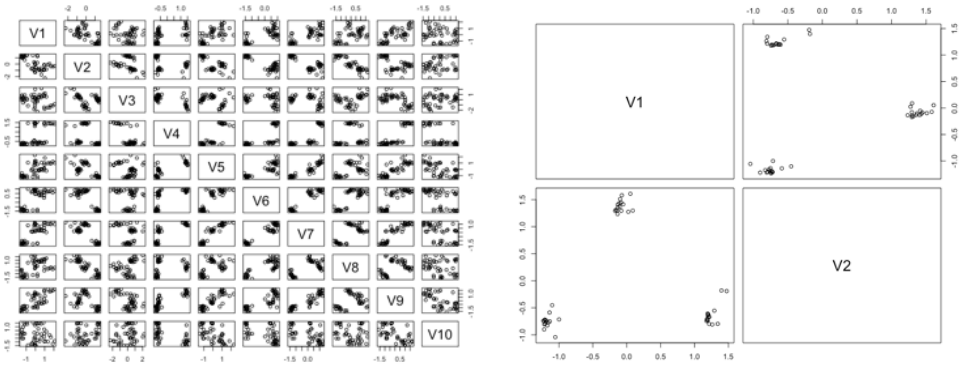


Fig. 6 Features extracted for the synthetic altitude data using B-spline (left) and FPCA (right)

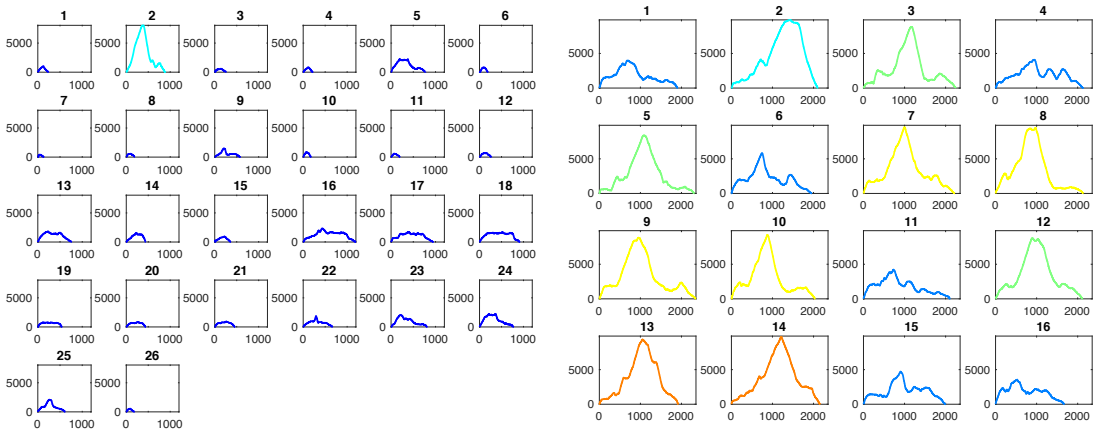
2. Real data

Unlike the synthetic altitude data, distinct clustering patterns were less frequently observed in the real altitude data from the ADS-B dataset. The Hopkins statistics for both feature extraction methods are shown in Table 1 and they are generally smaller compared to the ones computed for the synthetic data. With the data more likely to be randomly distributed in the feature space, the clustering results from the k-means method may not be stable due to the initialization

of the number of clusters. Thus, the hierarchical clustering method was chosen to find patterns in the real dataset. Besides, the FPCA feature typically has a lower dimensionality than the B-spline feature. Based on the principle of parsimony, the FPCA feature is selected to extract information from the ADS-B altitude data. In Fig. 7, two sample results, one from a group of flights with a higher Hopkins statistic and the other from a group of flights with a lower Hopkins statistic, are demonstrated separately. The results from group 6 show that the FPCA method is capable of isolating flight 2, which reached a higher altitude and has a relatively sharp profile, from the remaining flights in this group. However, the algorithm struggled to differentiate minor nuisance for those flights operated at a lower altitude. The results from group 8 show that the FPCA method can place flights into individual groups based on their profile patterns even if these flights have similar maximum altitudes and flight durations.

	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7	Group 8
Sample size	34	31	63	52	18	26	24	16
Hopkins Stat. for B-spline	0.628	0.668	0.627	0.601	0.62	0.643	0.562	0.624
Hopkins Stat. for FPCA	0.67	0.752	0.605	0.579	0.539	0.835	0.497	0.578

Table 1 Hopkins statistics for grouped altitude data



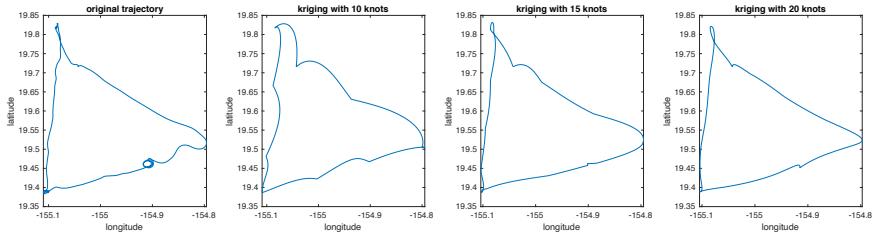
Sample results from Group 6 **Sample results from Group 8**

Fig. 7 Some sample results for the altitude analysis

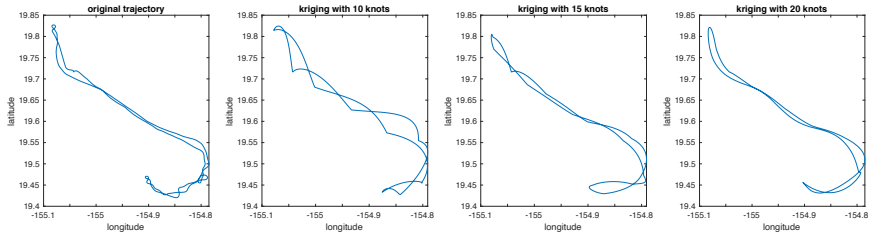
D. Trajectory data analysis

1. Synthetic data

Similarly to the altitude data analysis, our approach was first tested on synthetic data. The kriging approach was applied to both the longitude and latitude data to generate some synthetic trajectory data. The corresponding synthetic trajectories are shown in Fig. 8. The generated trajectories can capture the essence of the real trajectories while also showing some slight variations. Higher numbers of knots result in smoother synthetic trajectories. Furthermore, the kriging approach can only imitate the overall pattern of the flight but fails to catch finer details such as the minute circular loiter in the triangular trajectory. Results show that the HOG feature combined with the hierarchical clustering can accurately place flights into meaningful groups based on their trajectory patterns.



Sample 1: triangular shape



Sample 2: hook shape

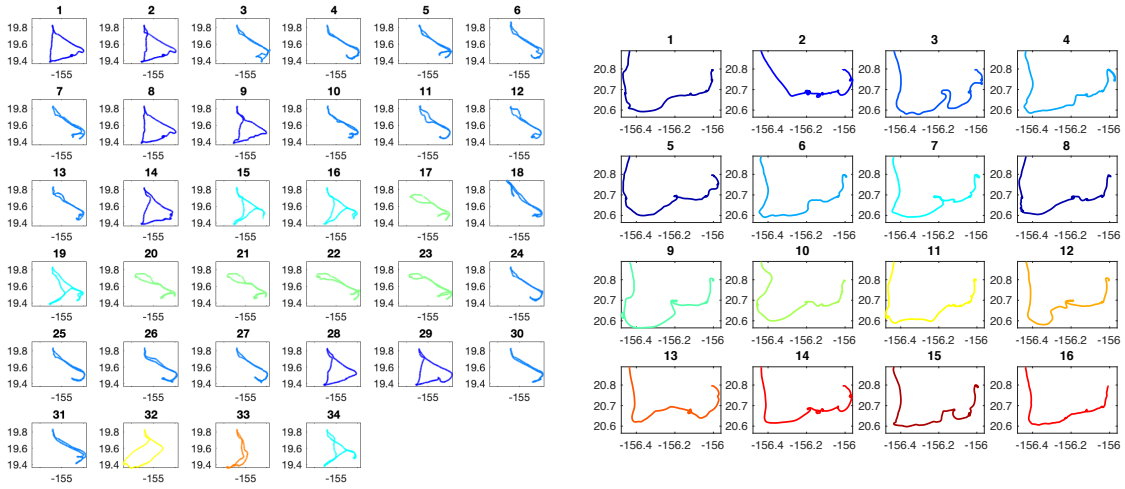
Fig. 8 Synthetic trajectory data from 2 different samples

2. Real data

The same groups of flights as in the altitude analysis were used to test our approach for the trajectory data analysis. The Hopkins statistics for each group are summarized in Table 2. Some of them are close to 0.5 which means that it would be difficult for the clustering algorithm to determine meaningful groups. Two sample results from the trajectory analysis are shown in Fig. 9. The results from group 1 (which has a higher Hopkins statistic) show that the HOG + hierarchical clustering method works well to place flights into their relevant groups based on the shape of their respective trajectories. The results from group 6 (which has a Hopkins statistic closer to 0.5) show that all the trajectories look similar at first glance. By using the HOG + hierarchical clustering method, 12 clusters are identified from a sample size of 16. The high number of clusters compared with the sample size indicates that nearly each of the flight in this group is contained within its own cluster. Two groups have more than one flight: one group with flights 1, 5, and 8, and another group with flights 14 and 16. The subtle difference in details of the trajectories might not be detectable to the human eye but the HOG + hierarchical clustering method can certainly be used to assist in discerning these patterns.

	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7	Group 8
Sample size	34	31	63	52	18	26	24	16
Number of clusters	6	12	3	4	2	2	8	12
Hopkins Stat.	0.643	0.625	0.663	0.603	0.537	0.551	0.601	0.513

Table 2 Hopkins statistics for grouped trajectory data



Sample results from Group 1 **Sample results from Group 8**

Fig. 9 Some sample results from the trajectory analysis

VI. Conclusion and Future Work

Helicopters are versatile aerial vehicles and they are ideal for a variety of operations. However, helicopter accident rates have reached a plateau and even have exhibited an increasing trend recently. To mitigate the risk of accidents, this study proposes a methodology to analyze helicopter flight data records and try to detect anomalies within a given type of operations. In this paper, we placed our focus on finding patterns within the ADS-B data from a helicopter tour operator. Several challenges associated with the analysis of this dataset were addressed and a process consisting of feature extraction and clustering analysis was proposed to fulfill this knowledge discovery (or anomaly detection) task. Two types of flight data were analyzed in this study: time-series data and trajectory data. To extract information from the time-series date into a low-dimensional feature, the B-spline and the FPCA methods were chosen as candidate methods. It is shown that the FPCA is more capable of finding relevant patterns in the time-series data compared to the B-spline. For the trajectory data, the Histogram of Oriented Gradient (HOG) feature was adopted to extract information from images of the trajectory data. With the addition of the average silhouette metric and the hierarchical clustering method respectively, patterns in the time-series and the trajectory data could be retrieved either within the synthetic data or the real data.

To bridge the gap between detected patterns and anomalies in flight data records, it is required to have SME intervention for identifying truly anomalous flights. The outliers discovered in the aforementioned analysis might correspond to rare events in a typical operation. With the feedback from SMEs, the current method can be improved to achieve anomaly detection from knowledge discovery. Furthermore, the overall process of knowledge discovery (or anomaly detection) developed in the case of a helicopter tour operator might not work well for other helicopter missions such as EMS and offshore or oil and gas. A future study will focus on extending or generalizing the proposed approach in this paper to explore different types of helicopter operations.

Acknowledgments

The work here presented is funded by the Federal Aviation Administration through PEGASAS (Partnership to Enhance General Aviation Safety, Accessibility and Sustainability), FAA Center of Excellence on General Aviation, Project No. 2: Rotorcraft Aviation Safety Information Analysis and Sharing (ASIAS). Project No. 2: Rotorcraft ASIAS is a partnership between PEGASAS researchers at the Georgia Institute of Technology, the FAA and Helicopter Association International (HAI). The information presented in this paper and contained in this research does not constitute FAA Flight Standards or FAA Aircraft Certification policy.

References

[1] U.S. Helicopter Safety Team, "U.S. Helicopter Safety Team Warns About the Next "Bump in the Road" for Fatal Accidents," http://ushst.org.dnn4less.net/Portals/0/2019-Release-Oct-and-Nov.pdf, 2019.

[2] Iseler, L., DeMaio, J., and Rutkowski, M., "An analysis of us civil rotorcraft accidents by cost and injury (1990-1996)," 2002.

[3] Fox, R. G., "The history of helicopter safety," *International helicopter safety symposium*, 2005, pp. 1–17.

[4] Gavrilovski, A., Collins, K., and Mavris, D. N., "Model-Enhanced Analysis of Flight Data for Helicopter Flight Operations Quality Assurance," *72nd Forum of the American Helicopter Society*, 2016.

[5] Payan, A. P., Lin, P.-N., Johnson, C., and Mavris, D. N., "Helicopter Approach Stability Analysis Using Flight Data Records," *17th AIAA Aviation Technology, Integration, and Operations Conference*, 2017, p. 3437.

[6] Payan, A. P., Gavrilovski, A., Jimenez, H., and Mavris, D. N., "Improvement of rotorcraft safety metrics using performance models and data integration," *Journal of Aerospace Information Systems*, 2017, pp. 26–39.

[7] Chin, H.-J., Payan, A., Johnson, C., and Mavris, D. N., "Phases of flight identification for rotorcraft operations," *AIAA Scitech 2019 Forum*, 2019, p. 0139.

[8] Zanella, P., Mavris, D. N., Collins, K., and Johnson, C., "Filter-Based Detection of the Proximity to Loss of Tail Rotor Effectiveness within Helicopter Flight Data Monitoring," *75th Forum of the American Helicopter Society*, 2019.

[9] Amidan, B. G., and Ferryman, T. A., "Atypical event and typical pattern detection within complex systems," *2005 IEEE Aerospace Conference*, IEEE, 2005, pp. 3620–3631.

[10] Budalakoti, S., Srivastava, A. N., and Otey, M. E., "Anomaly detection and diagnosis algorithms for discrete symbol sequences with applications to airline safety," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, Vol. 39, No. 1, 2008, pp. 101–113.

[11] Das, S., Sarkar, S., Ray, A., Srivastava, A., and Simon, D. L., "Anomaly detection in flight recorder data: A dynamic data-driven approach," *2013 American Control Conference*, IEEE, 2013, pp. 2668–2673.

[12] Iverson, D. L., "Inductive system health monitoring with statistical metrics," 2005.

[13] Das, S., Matthews, B. L., Srivastava, A. N., and Oza, N. C., "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*, 2010, pp. 47–56.

[14] Li, L., Das, S., John Hansman, R., Palacios, R., and Srivastava, A. N., "Analysis of flight data using clustering techniques for detecting abnormal operations," *Journal of Aerospace information systems*, Vol. 12, No. 9, 2015, pp. 587–598.

[15] Puranik, T., Jimenez, H., and Mavris, D., "Energy-based metrics for safety analysis of general aviation operations," *Journal of Aircraft*, Vol. 54, No. 6, 2017, pp. 2285–2297.

[16] Chu, E., Gorinevsky, D., and Boyd, S., "Detecting aircraft performance anomalies from cruise flight data," *AIAA Infotech@ Aerospace 2010*, 2010, p. 3307.

[17] Liao, T. W., "Clustering of time series data—a survey," *Pattern recognition*, Vol. 38, No. 11, 2005, pp. 1857–1874.

[18] Aghabozorgi, S., Shirkhorshidi, A. S., and Wah, T. Y., "Time-series clustering–a decade review," *Information Systems*, Vol. 53, 2015, pp. 16–38.

[19] Dalal, N., and Triggs, B., "Histograms of oriented gradients for human detection," *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, Vol. 1, IEEE, 2005, pp. 886–893.

[20] Hamid, N. A., and Sjarif, N. N. A., "Handwritten recognition using SVM, KNN and neural network," *arXiv preprint arXiv:1702.00723*, 2017.

[21] Lawgali, A., "Recognition of Handwritten Digits using Histogram of Oriented Gradients," 2016.

[22] Kassambara, A., *Practical guide to cluster analysis in R: Unsupervised machine learning*, Vol. 1, STHDA, 2017.

[23] Hopkins, B., and Skellam, J. G., "A new method for determining the type of distribution of plant individuals," *Annals of Botany*, Vol. 18, No. 2, 1954, pp. 213–227.

[24] Tibshirani, R., Walther, G., and Hastie, T., "Estimating the number of clusters in a data set via the gap statistic," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Vol. 63, No. 2, 2001, pp. 411–423.

[25] Rousseeuw, P. J., "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, Vol. 20, 1987, pp. 53–65.