An Unsupervised Approach to Anomaly Detection from Aircraft Flight Data

Daniel LC Mack[1] and **Gautam Biswas**[2] and **Dinkar Mylaraswamy**[3]and **Raj Bharadwaj**[3]
[1]Kansas City Royals, Kansas City, MO, USA.
email: daniel.mack@gmail.com
[2]Vanderbilt University, Nashville, TN. USA.
email: gautam.biswas@vanderbil.edu
[3]Honeywell Aerospace, Golden Valley, MN, USA.
email: {dinkar.mylaraswamy,raj.bharadwaj}@honeywell.com

Abstract

Fault detection and isolation schemes are designed to detect the onset of adverse events during operations of complex systems, such as aircraft and industrial processes. In this paper, we discuss an anomaly detection method to find previously undetected faults in aircraft systems. We discover these flight anomalies by combining an unsupervised learning technique applied to search a large database of flight operations data. The unsupervised learning technique combined with a feature extraction scheme applied to the anomalous clusters, facilitates expert analysis in characterizing relevant anomalies and faults in flight operations. We present a case study using a large flight operations data set, and discuss results to demonstrate the effectiveness of our approach.

1 Introduction

Fault detection and isolation schemes play an important role in detecting the onset of adverse events during operations of complex systems, such as aircraft and industrial processes. In previous work on aviation safety, we have developed an approach that combines classifier techniques from machine learning and expert input to better characterize faulty behaviors using available aircraft flight data [Mack, et al., 2016, in press]. Using a set of case studies, we demonstrated that the diagnostic relations derived from the data led to improved accuracy and faster detection times for the online reasoner on the aircraft.

In this paper, we discuss an anomaly detection method that uses exploratory methods to find previously undetected anomalies in aircraft flight operations [Mack, 2013]. We discover these flight anomalies using *unsupervised learning* techniques that are applied to a large database of flight operations data. We assume that most of the flight instances in this database represent normal operations, but a small subset of flights show sufficiently large deviations, and form groups or clusters that are different from the much larger groups that include mostly nominal flights. Since flight operations generate a very large amount of data, for example, our data set includes flight data collected from 37 regional jets, which flew, on the average, five flights a day. Each flight was between 30 minutes and 3 hours long, and about 182 continuous variables were recorded at rates that varied from 1 to 16 Hz. We worked with five years of flight data from these aircraft, therefore, the search for anomalies may be equated to a "looking for needles in a haystack" problem.

In our approach, we begin by transforming curated flight data into a data cube representation that we describe in the next section. We consider individual flights to be the data objects, the sensors to be the features, and each sensor's measurements over time during the flight to be the signals. In the case study described in this paper, our data cube is contextualized to capture the takeoff phase of flight so we may focus on specific types of anomalies. Before we can apply unsupervised learning or clustering algorithms to this data, we first reduce the dimensionality of the data cube into a two dimensional data set of dissimilarity or distance values between each pair of flights in our data. We apply a method based on the wavelet transform [Struzik & Siebes, 1999] as a mechanism for reducing a continuous time series signal to a set of features that retain the important characteristics of the signal. As a next step, we apply a hierarchical clustering algorithm [Johnson, 1967] to the reduced data. The clustering algorithm produces a dendrogram structure from which we extract a set of clusters. The larger clusters, i.e., the clusters that contain a large number of flights are labeled as nominal, and the remaining smaller groups are labeled as anomalies for further study and analysis.

In more detail, we extract the *significant features* that differentiate the anomalous clusters from the nominal ones. We define a significant feature as one that best differentiates the examined instance from the selected nominal set. When a cluster contains many more instances than can be examined manually, we produce a group characterization method using interactive techniques to identify the relevant features that differentiate the group from the nominal. These significant features are ranked by importance, i.e., their contribution to the distance from the nominal cluster. We present the ranked features for a cluster to the expert using plots of the anomalous signal against a nominal sample. The domain experts then analyze these deviant features to understand the nature of the anomalous group. If the experts considers the differences to be significant and have implications on the safety of the aircraft, then they label the anomalous groups as faults, and use the significant features as identifiers of the faults.

The rest of this paper is organized as follows. Section 2 provides a background on anomaly detection and anomaly detection techniques that have been employed in the flight data domain. Section 3 presents a formal definition of the problem, and our approach to solving the anomaly detec-

tion problem. Section 4 presents case studies on results we obtained by applying this methodology to flight data from 60,000 flights collected from a fleet of 37 identical aircraft belonging to a regional airline. Section 5 discusses our results, and presents our conclusions and directions for future work,

2 Background

Anomaly detection refers to the problem of finding patterns in data that do not conform to expected behavior [Chandola, et al., 2009]. These patterns typically called anomalies and sometimes called *outliers*, correspond to single data points or a small group of data points that appear to be sufficiently different from the rest of the data points in the set. Traditional approaches to anomaly or fault detection rely on a model that defines nominal behavior of a system, or on human expertise that characterizes the parameters or thresholds that separate nominal from anomalous behavior [Venkatasubramanian, et al., 2003]. However, in many situations, as discussed earlier, sufficiently accurate and complete models of the system may not be available, leading to misclassification or missing faulty behavior [Venkatasubramanian, et al., 2003]. In some situations, faulty and anomalous situations may be unknown because of a lack of sufficient experience in operating the system [Mack, 2013]. In such situations, data-driven approaches that lead to discovery of anomalous situations become valuable to protect system safety and integrity [Mack, et al., 2016].

In a simplistic sense, anomaly detection approaches define regions of space that describe nominal behavior, and then develop similarity-based measures to label data points that do not belong to the nominal regions as anomalies [Chandola, et al., 2009]. Several factors can make this task challenging. For example, defining every possible region of nominal behavior for a complex system may be difficult, and the similarity measure may be designed to accentuate certain feature differences more than others. Furthermore, as environmental conditions under which a system operates change, its nominal behaviors keep evolving, and current nominal behaviors may not be indicative of future nominal behaviors. Last, training data labeled as nominal and anomalous may be hard to come by. Further, even for labeled data, noise and corruption could distort the differences between nominal and anomalous behavior.

General approaches to exploring a feature space for identifying anomalous instances have employed a number of different learning algorithms, requiring highly tunable global models and error minimization procedures. Such methods include least-squares regression [Bishop, 2006] to derive discriminative models from data. This leads to robust algorithms that can detect a number of additive faults using receiver operating characteristic curves plotted to tune the detection algorithm and set the false alarm rates [Chu, et al., 2010]. Such approaches require large amounts of real and simulated data to derive general and robust solutions. Further, these methods are supervised approaches, and practitioners must understand the data and the results of experiments on the model (accuracy and false positives) in order to tune the system.

The domain of aviation flight data has produced a number of techniques for discovering anomalies, such as SequenceMiner [Budalakoti, et al.,2009], Orca [Bay & Scwabacher, 2003], Inductive Monitoring System [Iverson, 2004], and Morning Report [Chidester, 2003]. These methods rely on varying amounts of data, and are computationally expensive. For example, Morning Report, developed to run overnight on the previous day's flight data to generate a report that was examined in the morning. SequenceMiner focuses on clustering methods for exploring a set of instances. It reduces the dimensionality of the feature signals using a metric for measuring common sequences known as the *normalized longest common subsequence* [Budalakoti, et al., 2009]. This metric, computed for each feature, retains the original feature semantics, allowing an anomaly to be characterized in terms of the original feature signals.

Orca uses a scalable K-nearest neighbor approach to detect anomalies in data with continuous and discrete features. Each data point is treated as an independent sample in time by the algorithm, and, as a result, Orca struggles to detect anomalies that are defined by temporal signatures. On the other hand, the Inductive Monitoring System is a distance based anomaly detection method that operates on continuous parameters. The method uses incremental cluster analysis to build models of expected operation of the system, but also does not consider the temporal patterns in the data. The Euclidean distance from an anomalous data point to the nearest cluster center is reported as the anomaly score for that data point.

Morning Report builds a statistical signature across each feature of a sample to reduce it to a smaller dimension. Distance metrics, such as the Mahalanobis distance are used to find flights that are outliers from the majority of the data points. SequenceMiner and Morning Report analyze temporal signals in the data. These methods make assumptions; for example, SequenceMiner applies a symbolic transformation tp the data, and Morning Report requires a pass from another algorithm through the original data to help an expert characterize found anomalies.

More recent approaches that have produced good results in anomaly detection include an algorithm that combines Principal Component Analysis (PCA) and density-based clustering, DBSCAN [Li, et al., 2011] The approach uses PCA to project features in higher dimensional space to a lower dimensional space, and then applies DBSCAN to cluster the data in the lower dimensional space. The advantage of density-based clustering is that it is robust to noise in the data, and it requires little domain knowledge to define the size of a cluster. Also, clusters can be of arbitrary shape, and sufficiently different data points, labeled as outliers, are easy to find. The problem with this method is temporal information is lost in the unrolling process, and the transformed space makes interpretation of the anomalies a much harder task.

Another recent approach is the Multiple Kernel Anomaly Detection (MKAD) method [Das, et al., 2011]. This semi-supervised method can operate on large data spaces. Like SeqenceMiner, it first preprocesses all continuous time series data into symbolic sequences, and then applies

a similarity measure between pairs of samples or data points. The pairwise comparisons are organized for learning by building two kernels that combine the feature streams for either continuous or discrete values. The kernel method for both types is based on the normalized Longest Common Subsequence [Budalakoti, et al.,2009], the same metric used in SequenceMiner for measuring common sequences. The kernel is built for a one-class SVM classifier [Rätsch, et al., 2000]. The method for isolating anomalies attempts to exploit common sequential information for two samples represented as a single value. When built over the entire set of samples, this technique can construct the model of nominal behavior. Analysis of flagged anomalies is examined post-SVM, since the SVM model based on kernel methods is difficult to interpret.

3 Approach

Our approach is designed to address a number of the issues with both PCA-DBSCAN and MKAD, namely the temporal dependency that is ignored in PCA-DBSCAN, the lack of knowledge about nominal instances in the data for MKAD, and the ability to use the same data and clusters to identify and characterize anomalies in the data. Our approach to this exploratory anomaly detection for complex systems is broken into a series of steps.

First, we transform and contextualize the curated data to produce an initial data cube for exploration. We then apply dimensionality reduction to the cube to produce a transformed two dimensional dissimilarity matrix, which is used with a hierarchical clustering approach to generate clusters from the data. We assume the results of the clustering produce some large clusters where significant number of the flight instances reside. These are labeled as nominal data.

We divide the rest of the instances into two groups: (i) a very small number of outlier data points and (ii) some small clusters that are distinct from the large nominal clusters. As a first step, we study the outliers by *selecting significant features*, as defined earlier, to distinguish the outliers from the nominal clusters. Further analysis of the significant features, primarily done by the domain experts, enables us to label and characterize the anomalous data points and groups. We illustrate the stages of this approach in Figures 1 and 2. These two figures break up the work into two general procedures. Figure 1 represents the transformation to clustering stages of the data. The input to begin this process is a curated data set, which is then contextualized and transformed into the standard data cube.

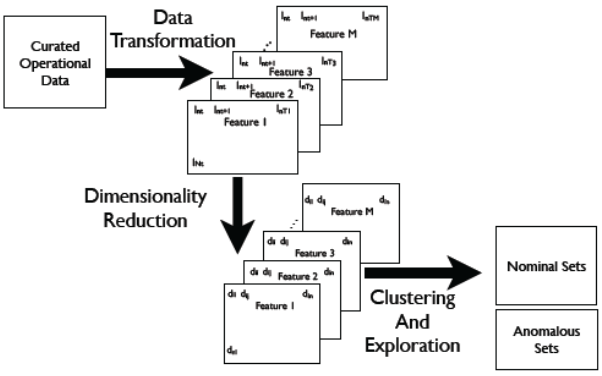

Figure 1: Part 1 of Anomaly Detection Process – Curated data to Anomalous Sets

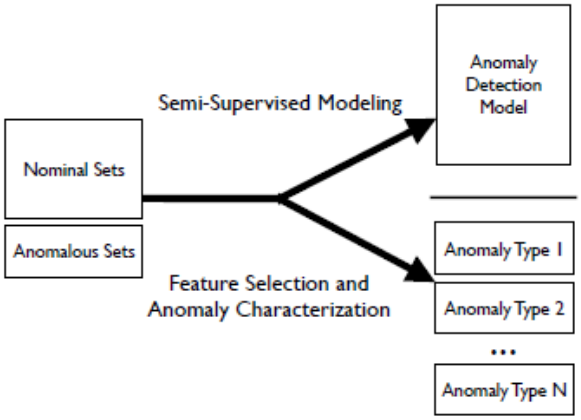

Figure 2: Part 2 of Anomaly Detection Process – Characterizing Anomalies

For computational purposes, and to make the task of finding the *significant features* easier, we apply dimensionality reduction to the data cube. This produces a two dimensional matrix of pairwise distances between the objects (flights). We employ a weighted Euclidean distance metric, and run a clustering algorithm using the distance matrix to identify two sets of clusters: (1) those that are in the nominal range; and (2) those that seem to imply *anomalies*. This takes us to the next step (see Figure 2), where we work with domain experts to identify and characterize truly anomalous situations.

These partitioned instances labeled as nominal or anomalous are the input for the procedures shown in Figure 2. Feature selection methods are employed to identify the relevant features that differentiate each anomalous group. An expert uses these features to further characterize the anomalies by their type and severity of failure. Coupled with the nominal sets, these groups provide the basis for designing models and detectors for automated anomaly and fault detection in future flights.

In this work, we have used the Haar Wavelet [Struzik & Siebes, 1999] for dimensionality reduction of the time series data. In preliminary experiments, when comparing against other data compression schemes, we found that this wavelet transform produced the best results against a variety of signal types, while also capturing the frequency and position characteristics of temporal changes that occurred within individual waveforms.

We convert each time series signal to a set of feature values using a wavelet transform, which, in addition to generating numeric feature values that capture the space-time characteristics of each signal waveform, also serves to compress and smooth the time series signals. In this work, we have applied the discrete Haar wavelet transform (DWT) (Strang, 1993), that uses translations and dilations of a square wave function. Computing the Haar wavelet coefficients is equivalent to passing the time series signal through a series of shifted and cascaded low- and high-pass filters that decomposes the signal into high and low frequency bands, which are then down-sampled to capture the local time-frequency characteristics of the signal.

Given the nature of the computation the number of discrete samples in the signal must be a power of 2. The set of features extracted for each object, O_i, is represented as a vector, $f_i \in R^{m \times l}$, where m represents the number of time series signals, and l is the number of levels extracted for each signal by the Haar transformation. As a result, each data object is represented by a set of $m \times l$ features. We found that the Haar wavelet decomposition applied to our data without the need for padding or removal, since each feature signal is a power of two. Since each aircraft sensor samples at a rate that is a power of two, a 32 second signal at full resolution remains a power of two with 32, 64, 128, 256 or 512 samples.

A hierarchical clustering algorithm using the complete link method [Murtagh, 1983] was used to create the dendrogram from the dissimilarity matrix of pairwise object distances. Complete link clustering only joins two clusters together if the furthest distance between any two points in the clusters is the smallest distance value remaining in the adjacency matrix. This information, known as the *linkage*, is stored for the links between single instances, as well as between the clustering of multiple instances. Complete link clustering yields clusters that are well separated and compact. Since we want to separate out the objects that are different from the majority of the objects in the data (considered nominal), this methodology should help produce clusters that are quite different from the nominal, and, therefore, candidates for anomalies..

The structure of the generated dendrogram can be used as a mechanism for visualizing the structure in the unlabeled data. We use the standard approach of looking for clusters that break the data into groups by applying a cutoff to the dendrogram at a chosen distance. Beyond this visualization, the dendrogram serves the operator and expert as a marker for identifying the cutoff in the hierarchy where clusters should form for this data [Jain & Dubes, 1988].

Locating the cutoff can be performed manually by eyeballing the dendrogram structure, or it can be performed by searching through a number of cutoffs with a goodness of fit calculation, such as the Cophenetic correlation coefficient [Farris, 1969] or the inconsistency coefficient [Jain & Dubes, 1988]. The overall purpose of these utility measures are to provide measures that help the researcher determine whether the clusters generated are too large and general, or too small and specific. The approaches compare the within cluster distances against the between cluster distances. Our approach in this paper used a search method, where we interleaved the use of the linkage criterion, with expert human feedback to identify useful anomalous clusters.

Given a possible cluster that represents an anomalous set of objects (flights), the question then becomes "*What features separate these anomalous flights from the set of flights that are labeled nominal by clustering*?" Feature selection for these instances or small groups is governed by the *significant features*, i.e., signals where the anomalies have the greatest differences from the nominal set. These differences can be found by examining the feature-by-feature differences between the anomalous and nominal sets of objects. After identifying the distances that produce the highest values (i.e., differences), we order the features based on the difference measure, and then pass them onto a human expert for further evaluation and characterization.

The choice of an average dissimilarity for the distances between the anomalous and nominal feature sets does not take into account noisy sensors, which can cause problems with the ordering of the significant features. We mitigate this issue by using a probability test to identify the likelihood that the anomalous distances would be drawn from the same distribution as the nominal distances.

Since we do not make any assumptions about the distribution, and the fact that the distances represent continuous values, we decided to use the non parametric Kolmogorov-Smirnov test [Conover & Conover, 1980] as the test to determine if the feature waveforms for two groups were significantly different. If the null hypothesis is rejected, we can feel confident that the distances are likely different between the two groups. If it is not, we can believe that the distances are sufficiently small, in which case the average distance would have been quite low, or the values are quite spread out among nominal and anomalous distances meaning that the corresponding sensor is noisy and unreliable for identification purposes.

Once ordered, the process of showing the significant features to the expert for characterization of the anomaly uses a tiered system. As a first step, the ten highest significant features were presented to the expert, and if the expert required more information, the next 10 were presented, and so on, till the distances threshold of normalized distances fell below 0.1. The expert preferred this tiered approach so that they could define anomalies in terms of their most significant differences from the nominal, and then provide supporting evidence as needed to fully understand the anomaly. The examples presented in the case study illustrate this process.

The significant features are displayed to the expert using temporal plots of the signal. The plots clearly mark the anomalous signal, but also plot a random sample from the nominal set for comparison purposes. For example, identifying an anomalous takeoff became easier by comparison against the plots generated by the nominal flights.

4 Case Studies

Honeywell Aerospace in Minneapolis, USA provided the flight data used for improving detection of existing faults [Mack, et al., 2016, in press] and discovering new faults and anomalies (the topic of this paper). It represented five years of flight data recorded from a former regional airline that operated a fleet of 4-engine aircraft, primarily in the Midwest region of the United States. Each aircraft in the fleet flew approximately 5 flights a day and data from about 37 aircraft was collected over the five year period. This produced over 60,000 flights. Since the airline was a regional carrier, most flight durations were between 30 and 90 minutes. For each flight, 182 features were recorded at sample rates that varied from 1Hz to 16Hz. Overall this produced about 0.7 TB of data.

For this study, we pre-processed the raw flight data to remove sensors that are not germane to the operation of the aircraft. This removal helped improve the efficiency of the approach. With the help of our experts, we reduced the number of waveforms used in our study from 182 to 87.

This data was further curated with location of each instance. The location, as a meta-data, was used later to classify anomalies that can be attributed to environmental conditions. The location, identified by mining the latitude and longitude positions at takeoff, is responsible for variance in operations associated with different altitudes and geographical elements in the data, such as typical weather conditions and length of the runway.

After applying all of the curation steps, we selected 5333 flights to include for forming our data cube (see Figure 1). This complete set covers flights of 12 different aircraft over a period of 5 years. This provided a broad enough selection to encompass a variety of flight situations that include takeoff locations, and weather, reducing the need to perform the analysis only for restrictive contexts.

We then extracted data for a chosen phase of operation for the aircraft. We focused on a specific phase of flight in order to contextualize the instances for the same period of time during flight. For this study, we focused on aircraft takeoff, a situation when the aircraft equipment and pilots are subjected to the most stress. To isolate the takeoff data, we used a "phase computer method", which relied on the computer of the aircraft to detect takeoff situations based on pre-specified conditions. This phase generally begins when the pilot applies significant thrust to the engines. We then took expert advice and extracted 90 seconds of flight data from this time point to represent the takeoff phase of the aircraft. The data was then down sampled, producing feature signals that had the same temporal length and were synchronized on the same points in time.

Our goal was to look for situations during flight operations, where the aircraft operated in previously unknown modes. These modes could be attributed to the equipment, the human operators, or environmental conditions (e.g., the weather). In such situations, the data may represent anomalies, i.e., aberrant, undesirable and faulty behavior [Chandola, et al., 2009], but there could be situations where unexpected behaviors occur, but they can be explained by circumstances (e.g., take-off from a high altitude airport) that differ from the norm.

Figure 3 shows the initial clustering results. This partitioning shows that the flight objects are subdivided into a large set which is considered nominal, and a number of smaller sets that need to be checked further to discover anomalous flights. The anomalous clusters containing about 138 flights appear on the right of the figure (they are enclosed in a rectangle).

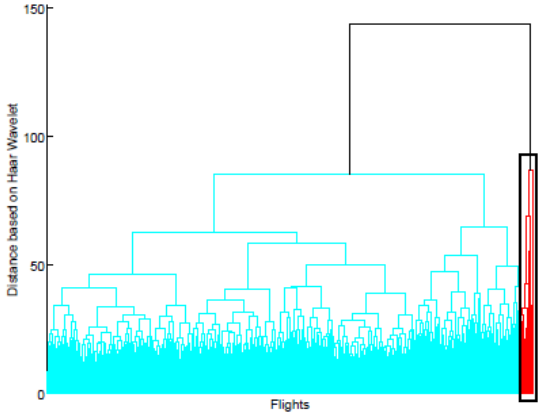

Figure 3: Dendrogram for 5333 flights generated using Complete-Link Clustering

Figure 4 shows a blowup of the dendrogram on the right, and three clusters (color-coded) are now clearly visible. From right to left, anomalous cluster 1 contains 9 flights, anomalous cluster 2 contains 39 flights, and anomaly cluster 3 contains 90 flights, respectively. All total, these 138 flights make up just 2.5% of the total flight objects. In this work, we focus on the three groups, and show how the expert investigates these different anomalies. In the rest of this section, we primarily discuss the anomalies found in group 1. We do not discuss the rest of the anomalies due to space limitations in this paper.

Characterizing Anomalies. On further analysis of Group 1, the experts found that the flights in this group fell into three specific categories. The first, a singleton object, Flight 5186 had the engine sensor values for the second engine that significantly deviated from nominal implying

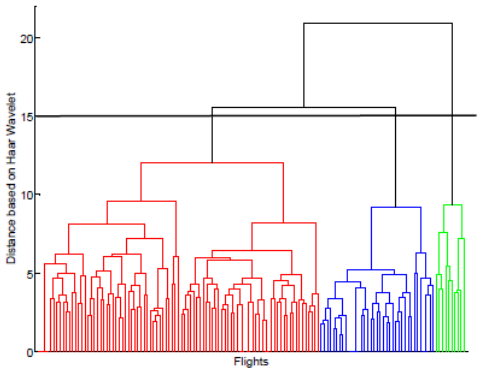

Figure 4: The three groups of potential anomalous flights

that the anomaly in flight 5186 was specifically associated with engine two. The significant feature deviants that were associated with engine two, such as Engine Temperature indicated that this engine was not producing any power. However, all of the other sensor readings indicated that this flight was in full takeoff at a normal altitude, and the other engines registered a higher than normal power readings. The experts examined all of these features and concluded that engine 2 was nonfunctional during the flight. If this were a flight with passengers, it would have represented a highly unusual situation, with strong safety implications. However, the experts concluded that this was a flight where the aircraft was being flown to an airport where maintenance work could be performed on the fault engine. The experts confirmed this by looking at the latitude and longitude of the start and destination airport.

The second group of anomalous flights was defined by a collection of environmental sensors, such as total air pressure and temperature, and the altitude as the significant deviant features. All flights in this group had 7900 feet as their takeoff altitude. Corroborating this information was the fact that the takeoff location was in a mountainous region of the United States. Since the radio altitude was not ranked in the top 10 significant features, this would appear to eliminate the fact that these takeoffs were otherwise anomalous compared to the nominal values. Instead, it was because this location was rare for this airline. The experts asked to examine an engine parameter, and then found engine temperature for the third engine as a significant

deviant feature in the second set of features. The conclusion was that the engine was running at higher power values, and it confirmed the experts' suspicions that these anomalies corresponded to high-energy takeoffs. In this case, the deviant features explaining the anomalies in the flights related to environmental parameters, but it was good to see that an anomaly caused by environmental parameters separated out from other anomalies by looking at the deviant features.

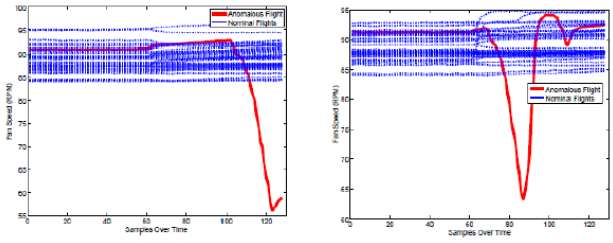

Figure 5: Fan Speed for Engine 3 for Flights 1256 and 3316

The third set within group 1 contains two flights, and a selection of the engine parameters, such as core speed and the fan speed in the engines, the fuel flow sensor values for the first and fourth engine, and the power level angle for the fourth engine significantly deviated from the nominal flights. Figure 5 shows the fan speed of the anomalous flight engines when compared against a random sampling of 50 nominal flights, and one notes that deviation in speed are quite large. However, after looking at all of the significant feature deviations for the two anomalous flights, the experts came to the conclusion that these two flights were quite different in terms of their takeoffs.

The expert used other significant features, such as flight path acceleration to understand the change in engine parameters in context for each flight. The flight path acceleration for flight 1256, illustrated in Figure 6, showed that the airplane slowed down off after takeoff. The expert postulated that this could be a part of the flight plan, since all engines were consistent in their changes. The expert decided that this type of flight takeoff was not unusual, and their conclusion was that the automatic throttle did not change mode as it should have, but that was very likely an auto pilot decision.

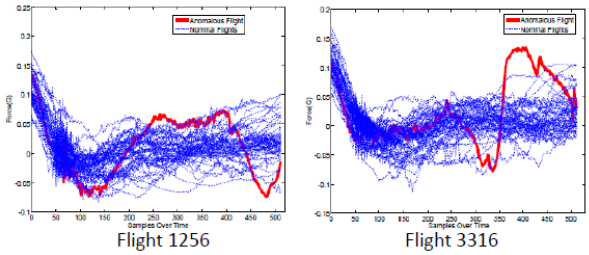

Figure 6: Flight Path Acceleration at Takeoff

Flight 3316, which initially appeared similar to flight 1256 turned out to be quite different. The experts concluded that the auto throttle disengaged in the middle of the climb. The automatic throttle is designed to maintain either constant thrust from the engines, or as a controller to maintain constant speed. The behavior of the significant deviant features in this case was unusual because the implication was that the auto thruster switched from maintaining speed during takeoff to a setting that applied constant thrust.

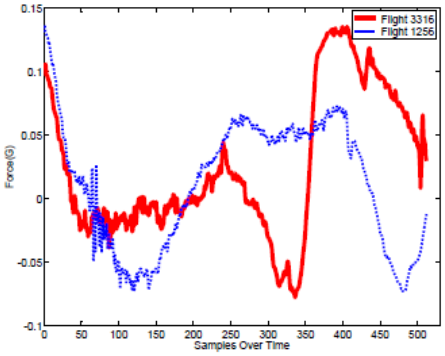

Figure 7: Comparison of Flight 1256 and 3316 Path Accelerations at Takeoff

The change in setting in the auto thruster indicated that the aircraft was on the verge of a stall. Figure 6 showing the flight path acceleration sensor confirms this. The sensor was trending high, and if the aircraft continued to operate along that trajectory, it would very likely have stalled during the takeoff. As the experts pieced together the information from the different features that showed significant differences, they explained that the automatic throttle would switch to a possibly lower thrust setting to compensate for this situation.

By examining the engine parameters, the expert verified that all the engines responded in an appropriate fashion to this throttle command. This meant that the aircraft responded and slowed down (the acceleration dropped at around 350 samples). Figure 7 shows the acceleration for the two flights plotted together. This shows that while both were clustered together, the expert could explain the difference in the two situations by examining the set of significant features. While flight 3316 certainly did not demonstrate a flaw in the aircraft or in its operation, the expert found the anomaly interesting and decided they would ponder on "*why did the aircraft accelerate in such a fashion and come so close to a stall condition*?" Since the expert could not determine the root cause, such an incident, once brought to their notice, resulted in their studying this situation in more detail to ensure safety would not be compromised in these situations. Further, the experts would also generate future guidelines so that pilots would not try to compensate for these situations manually, instead let the aircraft autopilot take over and compensate for this situation. The autopilot was better at avoiding stall conditions and more smoothly moving the aircraft trajectory away from stall conditions.

4 Discussion and Conclusions

This paper demonstrates an exploratory approach for identifying and characterizing anomalies in a large multivariate signal dataset of flight segments. This approach starts from curated data, extracts an appropriate part of the data, uses reduction techniques to convert the continuous time signals to a set of feature values, and then applies hierarchical clustering to the dataset. The derived clusters can be broken into one or more large nominal sets, and a smaller set that are hypothesized to be anomalies. Our approach also identifies significant features in the anomalous groups that

help experts analyze and explain the nature of the anomalies. The anomalous cases discussed in this paper, corresponded to small anomaly sets that the experts could analyze singly or in pairs. However, we have also developed methods, not discussed in this paper, where we can process anomalous groups that have larger number of objects using feature selection methods, such as projection pursuit [Mack, 2013].

We demonstrate how experts work with the anomalies our methods generate, and provide examples of how they may extract information that improves aviation safety. In our work, our focus has been on data related to aircraft takeoff, a phase that is known to be strenuous for the pilot and the aircraft equipment. We have shown by example that some anomalies can be related to the operating environment

Overall, the anomalies we discovered ranged from environmental causes, such as high altitude takeoffs, to ones in which the aircraft experiences changes in engine performance, to anomalies that indicate a pilot choice that would be worth investigating further from an aviation safety viewpoint. The overall case study showed that from the eight types of anomalies presented to our aircraft experts, three were flagged as interesting for further study [Mack, 2013]. This included a dead engine, a situation that led to a near stall, and a pilot choice to not switch to the autopilot soon after take-off.

In general, the expert found the method useful for identifying actionable anomalies from such a large dataset. One of the primary general contributions of this work is its applicability to analyze large volumes of unlabeled data, and to help practitioners and experts focus on interesting situations that were previously unknown. In terms of the bigger picture, it differs from traditional anomaly detection approaches [Chandola, et al., 2009] in that it starts with a completely unsupervised method, extracts information that is considered to be deviant from the nominal, and then uses expert analysis to characterize the new information. The overall methodology is quite general – in other work, it has been applied to analyzing anomalies in baseball pitching data [Mack, 2013] and more recently in finding anomalies in spacecraft missions from telemetry data [Biswas, et al., 2016, in review].

4 Acknowledgments

This work was supported by NASA NRA under Grant NNL09AA08B from the Aviation Safety Program.

References

[Bay & Schwabacher, 2003] Bay, S. D., & Schwabacher, M. Mining distance-based outliers in near linear time with randomization and a simple pruning rule. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. pp. 29-38.

[Bishop, 2006] Bishop, C.M. *Pattern Recognition and Machine Learning*, Springer.

[Biswas, et al., 2016, in review] Biswas, G., Khorasgani, H., Stanje, G., Dubey, A., Deb, S., and Ghoshal, S. An Application of Data Driven Anomaly Identification to Spacecraft Telemetry Data, 2016 *Prognostics and Health Management Conference*, Denver, CO.

[Budalakoti, et al., 2009] Budalakoti, S., Srivastava, A. N., & Otey, M. E. (2009). 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, , 39(1), 101-113.

[Chandola, et al., 2009] Chandola, V., Banerjee, A., & Kumar, V. Anomaly detection: A survey. *ACM computing surveys*, 41(3), Article 15, pp. 15:1—15:58.

[Chidester, 2003] Chidester, T. R. Understanding normal and atypical operations through analysis of flight data. In *Proceedings of the 12th International Symposium on Aviation Psychology*, Dayton, OH, pp. 239-242.

[Chu, et al., 2010] Chu, E., Gorinevsky, D., & Boyd, S. Detecting aircraft performance anomalies from cruise flight data. In *AIAA Infotech Aerospace Conference*, Atlanta, GA, April.

[Conover & Conover , 1980] Conover, W. J., & Conover, W. J. (1980). *Practical nonparametric statistics*. Wiley, NY.

[Das, et al., 2011] Das, S., Matthews, B. L., & Lawrence, R. Fleet level anomaly detection of aviation safety data. In *IEEE Conference on Prognostics and Health Management* (PHM), pp. 1-10.

[Farris, 1969] Farris, J. S. On the cophenetic correlation coefficient. *Systematic Biology*, 18(3), 279-285.

[Iverson, 2004] Iverson, D. L. Inductive system health monitoring. *Proceedings of the 2004 International Conference on Artificial Intelligence*. Las Vegas, NV, June.

[Jain & Dubes, 1988] Jain, A. K., & Dubes, R. C. *Algorithms for clustering data*. Prentice-Hall, Inc.

[Johnson, 1967] Johnson, S. C. Hierarchical clustering schemes. *Psychometrika*, 32(3), 241-254.

[Mack, et al., 2016] Mack, D. L., Biswas, G., Koutsoukos, X. D., & Mylaraswamy, D. Learning Bayesian Network Structures to Augment Aircraft Diagnostic Reference Models. *IEEE Transactions on Automation Science and Engineering*, in press.

[Mack, 2013] Mack, D. L. Anomaly Detection from Complex Temporal Sequences in Large Data. *Doctoral dissertation*, Vanderbilt University, USA.

[Murtagh, 1983] Murtagh, F. (1983). A survey of recent advances in hierarchical clustering algorithms. The *Computer Journal*, 26(4), 354-359.

[Ratsch, et al., 2000] Rätsch, G., Schölkopf, B., Mika, S., & Müller, K. R. SVM and boosting: One class. *GMD-Forschungszentrum Informationstechnik*.

[Strang, 1993] Strang, G. (1993). Wavelet transforms versus Fourier transforms. *Bulletin of the American Mathematical Society*, 28(2), 288–305.

[Struzik & Siebes, 1999] Struzik, Z. R., & Siebes, A. The Haar wavelet transform in the time series similarity paradigm. In *Principles of Data Mining and*

Knowledge Discovery (pp. 12-22). Springer Berlin Heidelberg.

[Venkatasubramanian et al., 2003] Venkatasubramanian, V., Rengaswamy, R., Yin, K., & Kavuri, S. N. A review of process fault detection and diagnosis: Part III: Process history based methods. *Computers & chemical engineering*, 27(3), 327-346.