Article can be accessed online at http://www.publishingindia.com

Delay Prediction of Aircrafts Based on Health Monitoring Data

B. A. Dattaram*, N. Madhusudanan,*

Abstract

Flight delay is a major issue faced by airline companies. Delay in the aircraft take off can lead to penalty and extra payment to airport authorities leading to revenue loss. The causes for delays can be weather, traffic queues or component issues. In this paper, we focus on the problem of delays due to component issues in the aircraft. In particular, this paper explores the analysis of aircraft delays based on health monitoring data from the aircraft. This paper analyzes and establishes the relationship between health monitoring data and the delay of the aircrafts using exploratory analytics, stochastic approaches and machine learning techniques.

Keywords: Aircraft Delay, Faults and Alerts, Markov Chains, Time before Failure, Stochastic Ensemble

Introduction

The failure of an aircraft to take off due to component faults and unplanned maintenance leads to revenue loss, affecting the core business of the aircraft owning companies. The problem with component faults is significantly observed in the ageing aircrafts. It is, therefore, necessary to anticipate delays so that proper maintenance processes can be initiated before an actual delay occurs. The health of the aircraft is monitored through fault and alert messages which are relayed from the different subsystems, during its journey. These faults and alerts are leading indicators of the health of the aircraft. We, in this paper explore the relationship between aircraft delays and the fault and alert messages from the aircraft.

Aircraft Sub-systems Description and ATA Codes

An aircraft can be considered as an assembly of different sub-systems like Engine, Hydraulics, Cockpit, Landing Gear, Electrical Components etc. which work cohesively. The industry standard body has established codes for each of the sub-systems. Table-1 describes Air Transport Association of America (ATA) chapter numbers for some of the important sub-systems of the aircraft.

In this paper terminologies like Sub-Systems, Airframe Systems and ATA Chapters are interchangeably used. The complete list of ATA Chapters and corresponding codes is provided in.

Table 1: List of Airframe Systems and Corresponding ATA Numbers

ATA Number	*ATA Chapter Name*
ATA 23	Communications
ATA 22	Auto Flight
ATA 24	Electrical Power
ATA 28	Fuel
ATA 72	Engine – Reciprocating

Data Description

In this section, we will describe the data that we have used for Exploratory Analysis and Time-to-Failure Prediction Model. The data consists of two separate sets: the first set contains the alerts and faults obtained from all aircrafts during their flight journey and the second set contains the delay dates of all aircrafts.

Faults and Alerts data consists of the Aircraft Number, ATA Chapter code associated with the fault or alert message and also the Timestamp. Table-2 provides a snapshot of Alerts

* IBM India Pvt Ltd, Bangalore, Karnataka, India. E-mail:dattarao@in.ibm.com, mnaraya7@in.ibm.com

and Faults. These are the health monitoring messages relayed by different aircrafts during the journey and may not always indicate an imminent failure.

The Delay Table consists of delay or cancellation events along with the Aircraft Number, Timestamp and associated ATA Chapter code. Delay events represent situations when an aircraft failed to take off. In this paper, we establish the relationship between the Faults and Alerts data and delays or cancellations (hereafter also referred to as Failures) using exploratory and machine learning techniques. In this paper, alerts and faults have been treated homogenously. Delays and cancellations have been treated together as failure events. Data for a total of 63 aircrafts all belonging to the same series with same engine was considered. The age of the aircraft is in the range of 15-20 years. A total of one year of data was provided.The total number of delay events were approximately 3.5% of the total take-off events.

Table 2: Format of Alerts and Faults Data Obtained from Different Aircrafts

Aircraft Number	*ATA Chapter Code*	*Message Type*	*Timestamp*
A1	32	Alert	2012/04/05 14:11:05
A2	73	Fault	2012/04/05 14:30:25
A23	73	Alert	2012/04/05 13:11:05
Event Type	Aircraft Number		Timestamp
Delay	A1		2012/04/05
Cancellation	A1		2012/04/12
Delay	A2		2012/05/05

Related Work

In the field of aerospace, a lot of work depends on reliability theory and preventive maintenance. In particular, there is lot of effort towards data mining based prediction for component replacements (Létourneau, Fazel, & Matwin, S. 1999). These approaches also make use of the sensor data obtained from the aircraft along with the text comments from maintenance crew to predict the most optimal time to replace a component. Data mining methods like Naïve Bayes, decision trees or even hybrid models have been used earlier to predict the most optimal time for component replacement.

Delay Prediction: Our Approach

In this section, we will describe our approach on data analysis and delay prediction using machine learning techniques. In the next few sections we describe our insights and results from exploratory analysis, Hidden Markov Models and prediction of Time Before Failure (TBF) using a combination of regression trees and stochastic ensemble modeling.

Exploratory Analysis

The objective of exploratory analysis is to establish that there is information contained within the ATA Fault and Alert messages for the prediction of the next failure event and the consolidation of additional information which can prove useful for understanding underlying patterns in the occurrence of messages or for modeling TBF. In this method, we have analyzed the delay events and the

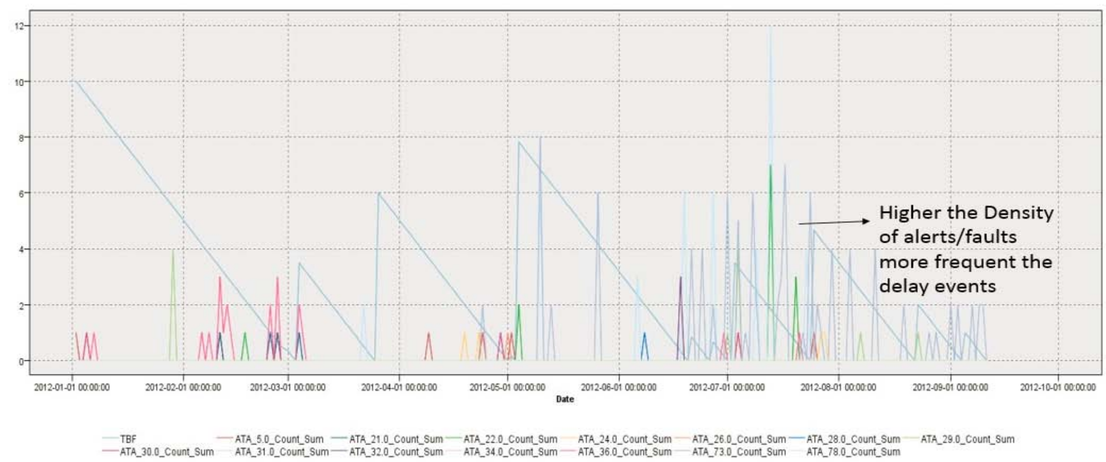

Figure 1: Plot of ATA Count vs. Delay Events for One of the Aircraft

ATA code occurrences based on the count of each ATA aggregated on daily basis.

Figure 1 plots the aggregated count of each ATA message per day against the delay events and TBF. This plot allows us to make observations which reveal the relationship between alert and fault messages of different ATA codes and the failure events.

It can be observed that the density of ATA alerts or faults is closely related to the delay events represented by the saw-tooth waveform. A higher amplitude of the saw-tooth represents a longer time before another failure event occurred. At the beginning of the year, the alert and fault message density is low and so are the delay events. When the density of alerts and faults increases, the TBF value also drops leading to a delay event. This visualization allows us to establish a causal relationship between alerts and faults messages and delay events due to different sub-system issues.

Based on insights revealed by the above trend chart, further analyses of daily message count and their relationship to TBF was performed for each ATA message, across all aircrafts. Figure 2 shows the results for one of the ATA codes.

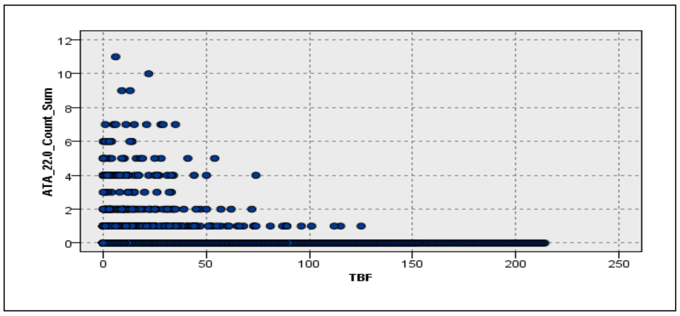

Figure 2: Per day Count of ATA 22 Messages vs. TBF

The interpretation of the above graph is as follows - as the daily count of ATA 22 messages increases, the distribution of the TBF shifts closer to zero. In other words, a higher frequency of ATA 22 messages indicates a higher temporal proximity to the next failure event. It was found that the pattern is similar across all faults and alert messages of each ATA - a higher frequency indicates a closer proximity to failure.

Hidden Markov Model

HMMs are two layered stochastic models with one hidden layer which cannot be observed. The hidden layer Markov process can be observed through the observations or symbols which hidden layers emit in agreement to the laws of probability and Markov chains. HMMs are widely used in gene sequencing and speech recognition.

Why HMMs for Aircraft Delay Prediction?

In this paper, we have made an attempt to model the aircraft as a two layered Markov chain. The hidden layer represents the interaction of different aircraft sub-systems and is modelled as a first order Markov chain. This interaction among different sub-systems is complex to specify and model and hence is considered as a hidden layer. The top layer which can be observed in the form of alerts and faults messages with specific ATA codes is also treated as a first order Markov chain. It can be seen that the interaction of the sub-systems in the hidden layer emits the observations or the messages with ATA codes. This is depicted as a schematic diagram in Figure 3.

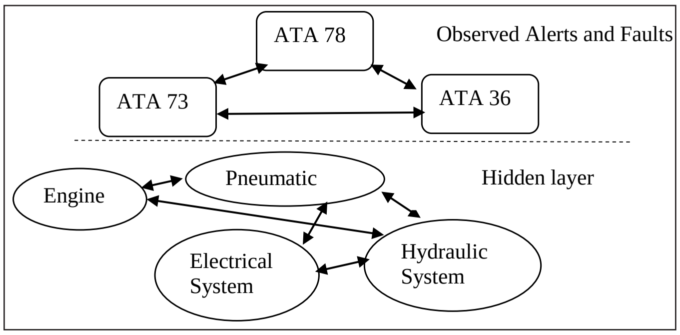

Figure 3: Schematic Diagram Showing Hidden and Observed Layers

HMM is proposed as a model with the following objectives:

1. HMM Model on aircraft delay data should help in understanding the interactions among the different sub-systems. Like, how tightly coupled are certain sub-systems or are there sub-systems which do not influence others
2. HMM based Delay Prediction Model

HMM Specification

A discrete HMM model is specified using emission and transition probabilities. Emission probabilities is defined as the probability of an observation or a symbol being emitted from a hidden state. Transition probabilities include the probability of transition from one hidden state to next hidden state.

In order to achieve the above objectives, we define the HMM with 4 hidden states and 16 symbols which compromise all the ATA codes and the delay event represented as symbol "F" (Salfner, 2005).

Observations={"5","21","22","24","26","28","29","30" ,"31","32","34","36","45","73","78","D"}

Hidden States = {M1, M2, M3, DELAY}

Training the HMM

Once the model is initialized with emission and transition probabilities, the model is trained using Baum-Welch algorithm to derive the most likely set of transition and emission probabilities. 80% of the data is used as input to Baum-Welch algorithm to derive the emission and transition probabilities.It is ensured during initialization that only the Failure State emits the Delay or "D" symbol. By initializing only the Failure State to emit the Delay symbol, it can be guaranteed that failure symbol in training sequence will change the transition to Failure State (Salfner, 2005). For details of theory and background on Viterbi and Baum-Welch algorithms, refer to (Salfner, 2005; Rabiner & Juang, 1986)

Results from HMM model

As mentioned earlier, one of the objectives of the HMM model is to understand the interactions among the different sub-systems of an aircraft. In this regard, the Table 4 lists the emission probability from each of the hidden states and Table 5 lists the transition probabilities.

From the emission probabilities, it can be observed that hidden state M1 has maximum emission probability only for Indication and Recording sub-systems while M2 has the most number of Protection sub-systems along with Landing, Pneumatic and Navigation. The maximum emission probability for each of the ATA codes and the hidden state reveals how some sub-systems in the aircraft are more tightly coupled. It also explains the sequence of ATA codes that are observed in the alerts or fault messages.

Table 4: Hidden State and Emission Probabilities

Hidden State	*Observations with maximum Emission Probability for the Hidden states*
M2	Maintenance Checks(5) , AC and Pressure (21), Electrical Power(24), Fuel(28), Auto Flight (22), Rain protection(30), Fire protection(26), Landing(32), Pneumatic(36), Navigation(34), Diagnostic(45), Exhaust(78), Hydraulic(29)
M3	Engine (73)
M1	Indication or Recording (31)
DELAY	Delay (D)

Table 5: Transition Probabilities Filled up to 3 Digits

	M1	*M2*	*FAILURE*	*M3*
M1	0.77	0.15	0.037	0.036
M2	0.015	0.85	0.103	0.027
FAILURE	0.035	0.88	0	0.079
M3	0.009	0.102	0.038	0.84

The transition probabilities reveal that M2 hidden state has transition probability of 0.103 to Failure state while other states have very low probability of transition to Failure state.

Predicting the Probability of the Failure State

Given an observation at time t, the next hidden state at time t+1 can be estimated by multiplying the posterior probabilities at time t and the transition probabilities obtained from the HMM. Posterior probability represents the marginal distribution of a state given by the observations. Results from running the test data of different sequence of ATA codes reveal that HMM predicts the Failure state with high accuracy (80%) when the underlying hidden state sequence is dominated by state "M2". But HMM fails to predict the Failure state when the hidden state sequence is dominated by "M3" i.e. the observed ATA codes are from Engine sub-system.

Failure Prediction Model Using Stochastic Ensembles

This section outlines the methodology underlying the stochastic ensemble model for predicting the Time Before Failure (TBF). This method is predicated on an approach where each ATA message occurrence / event is regarded to contain information which allows us to estimate TBF with some level of certainty. However, a combination of such estimations from multiple events can also be used to augment the final estimate. In other words, the frequency of each ATA message and equivalent metrics can be used as a stochastic predictor of TBF and these probability distributions can be progressively combined to arrive at a more deterministic estimation of TBF.

The modelling methodology adopts three stages –

1. Regression Tree Modelling – Uses clustering to establish time boundaries and derive frequency metrics derived for each ATA message event and build multiple regression tree models to predict TBF
2. Distribution Fitment – Fits an appropriate distribution to describe the spread of TBF in each leaf of the regression tree
3. Distribution Readjustment and Stochastic Ensemble – Readjusts the fitted distributions for earlier fault/alert message events to account for the time elapsed since those events occurred. Combines all readjusted TBF distributions observed since the last failure event to obtain an accurate estimate of TBF

The following subsections outline each of these stages in greater detail, along with results obtained for the same. This paper restricts itself to a description of the core methodology and does not detail out programming intricacies such as techniques used for handling NULL or NA values generated during data processing.

Regression Tree Modeling

Regression trees are decision trees built for the prediction of continuous variables. The basis of establishing a statistically significant variation in the distribution of TBFs across node splits is determined using the ANOVA method.

Based on the insights from exploratory analysis, the predictors used for TBF prediction have to be related to the frequency of messages which is calculated as F = N/T, where N is the number of messages over a time span T. Daily message count aggregates used for exploratory analysis are not ideal because daily boundaries were defined for convenience and do not reflect inherent time boundaries present in the data. Therefore, new time boundaries were derived from the data using a simple heuristics based univariate clustering method.

Messages for each ATA for each Aircraft were processed in chronological sequence. A new data point is retained in the same cluster if it falls within 1.96σ deviations from the mean of the cluster. If the aforementioned criterion fails, the data point is classified into a new cluster. Sample results for ATA message73 for a specific aircraft are shown in Figure 4. The diagram represents each occurrence of a message with a dot, and the vertical lines represent the beginning of a new cluster.

For any message event, let $(N_{ATAj})_{Ti\text{-}Ti^c}$ represent the number of messages of ATA code ATAj occurring between Ti^c - the time which represents the beginning of the current cluster to which the event belongs, and Ti - the time of the message event. Then the frequency computed for the event is given by –

The frequency of messages is calculated for two scenarios - (a) by treating each ATA message as a separate event stream (b) by treating all messages as a single message

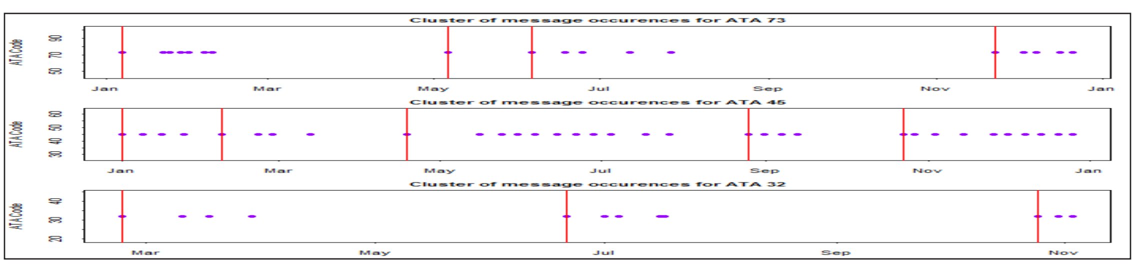

Figure 4: A Sample of Cluster Boundaries which were Extracted - ATA message 73 for one specific Aircraft

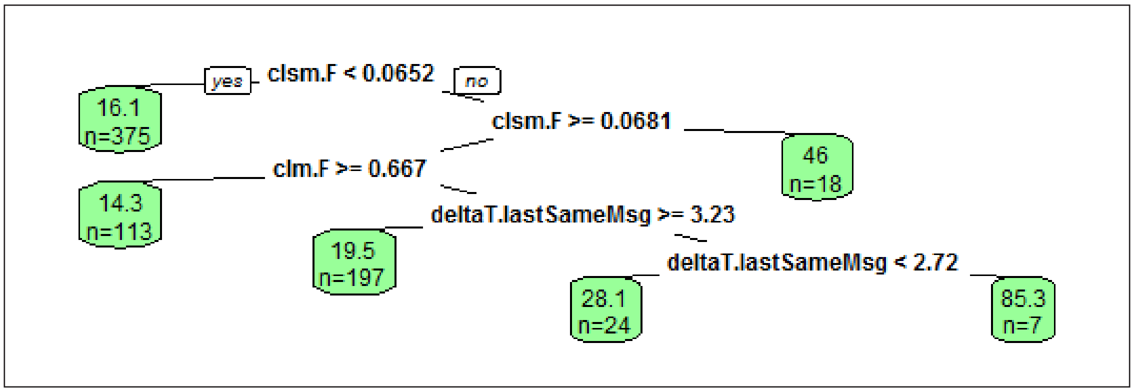

Figure 5: Regression Tree Built for ATA 5 Message Events

stream irrespective of the ATA code. This yields two different frequency metrics.

The regression tree models are built for each ATA code, across aircrafts, resulting in 15 regression tree models. One such regression tree built for ATA 5 is shown in Figure 5. The following predictors were computed for all aircrafts and considered for the regression tree model –

(i) Cluster frequency computed for each ATA message stream for every event (clsm.F)

(ii) Cluster frequency computed for all message events disregarding ATA codes (clm.F)

(iii) Time elapsed since the last message of the same ATA code (deltaT.lastSameMsg)

(iv) Time elapsed since the last message of any ATA code (deltaT.lastMsg)

Distribution Fitment

Each leaf of the regression tree model in Figure 5 displays the average Time Before Failure (TBF) of values in that leaf. However, each leaf (hereafter referred to as Class) contains a spread of TBF values which represent a probability distribution for TBF. This section discusses the fitment of an appropriate distribution to model the TBF for each leaf (Class).

From an examination of Cullen and Frey Graphs plotted for the TBF distributions, we can conclude that a gamma function is an appropriate choice for modelling the TBF distribution across all ATAs and Classes. Figure 6 shows a sample set of Cullen Frey Graphs plotted for ATA 45 for Classes 1 and 2. The graphs show that most of the observations fall close to the dashed line characterized by a gamma distribution.

Figure 7 shows the results of a sample fitment for ATA 73 Class 2. The results of each fitment yield two parameters for the fitted gamma distribution – shape and rate.

The final result of distribution fitment is a lookup table of values for each ATA code and Regression Tree Class

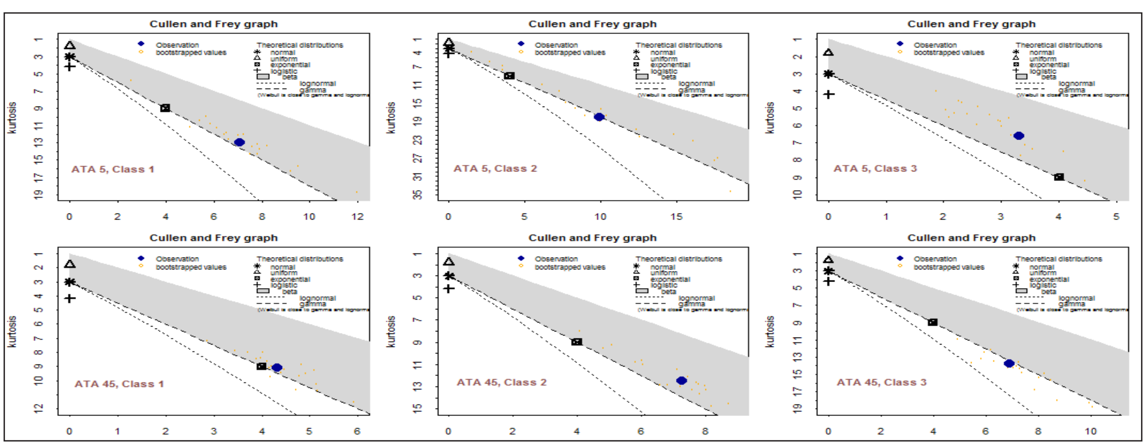

Figure 6: Cullen Frey Graphs for ATA 45 Classes 1 & 2

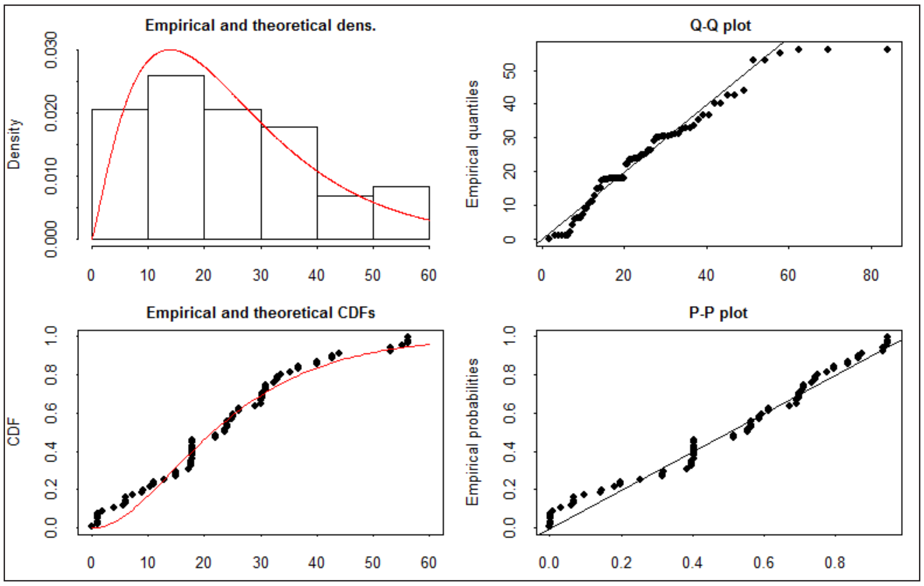

Figure 7: Gamma Distribution fitted for ATA 73 Class 2

for describing the distribution of TBF. A snapshot of this table is shown in Table 6.

Table 6: Snapshot of Distribution Fitment Lookup Table

ATA Code	*Class*	*Shape*	*Rate*
73	1	1.161687	0.08051
73	2	2.24427	0.089735
73	3	0.993364	0.065011
73	4	0.887306	0.057574
45	1	1.038601	0.071375
45	2	0.76693	0.029999
45	3	0.973843	0.055013
36	1	0.904286	0.04644
36	2	1.005415	0.061267
36	3	0.724237	0.043222
…	…	…	…

Distribution Adjustment and Stochastic Ensemble

This section describes the technique behind aggregating the stochastic predictions made by each ATA message event to provide an integrated prediction of Time Before Failure (TBF).

Integral to the technique of Stochastic Ensemble of predictions from occurrence of events is the readjustment of probability distributions. In this application, since the prediction made is the time to failure, the elapse of time allows us to invalidate a subset of the TBF distribution obtained from an event in the past. For example, if 6 days have elapsed since the last ATA 73 message which fell under Class 2, we know that the $p(TBF < 6) = 0$.This information can be used to progressively readjust the distribution through the passage of time.

Figure 8 illustrates the process of distribution readjustment of the probability distribution function and the cumulative distribution function after an elapsed time period of 6 days post the occurrence of the original event.

Stochastic ensembles are generated using the following process steps.

(i) The frequency estimators for each ATA event are used to predict the Class based on the Regression Tree Model

(ii) TheTBF distributions for each ATA and Class since the last failure event are obtained from the lookup table and adjusted for time elapsed

(iii) The readjusted distributions are combined using the stochastic ensemble method

The methodology adopted for combining the readjusted distributions has to satisfy the preconditions of being able to mimic the behavior of a deteriorating system and introducing greater certainty in the prediction of TBF. In addition, the information contained in distributions which predict a lower TBF need to be retained in the ensemble process. These properties are satisfied by considering a

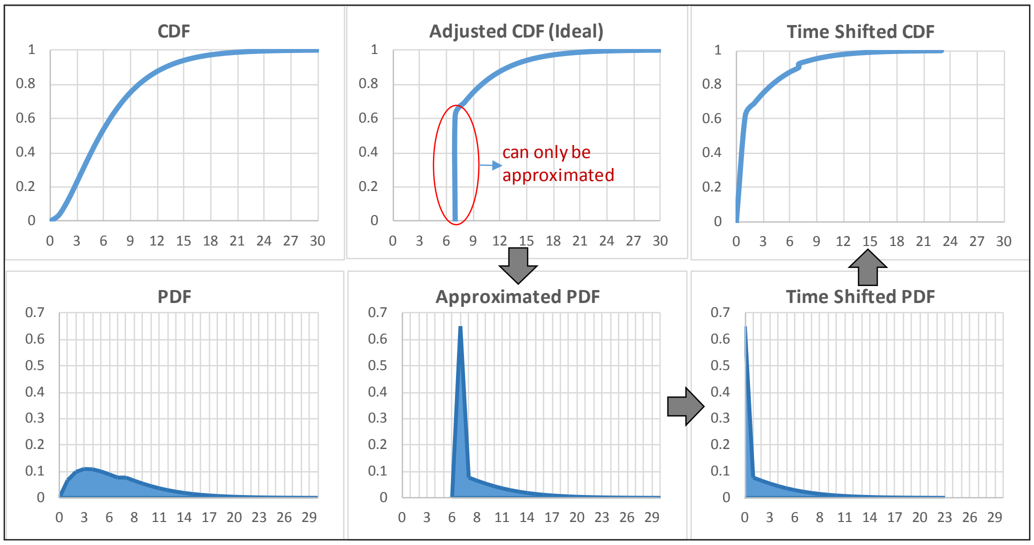

Figure 8: Distribution Readjustment Example (Left) Original Distribution at T=0 (Middle) Redistribution After T+6 days (Right) Final Distribution Time Shifted for 6 Days

product of the inverse cumulative distribution functions. If x_i is a random variable that denotes the uncertainty in TBFs predicted by past ATA message events, where $i \in \{1,2,3,\ldots n\}$, then the CDF of the combined distribution x_c is given by the below equation.

The generation of adjusted redistributions and their combination was achieved by employing Monte Carlo methods to generate a sufficiently large random sample based on the parameters of the gamma distribution, readjustment of the sample, calculation of cumulative probabilities and an interval based calculation of geometric means across transformed CDFs. The results of the TBF prediction before and after the ensemble process are shown in Figure 9.

Since the current model updates the TBF prediction only when new information in the form of an ATA message event occurs, sparse event counts in the latter half of the dataset are seen to affect model performance. However,

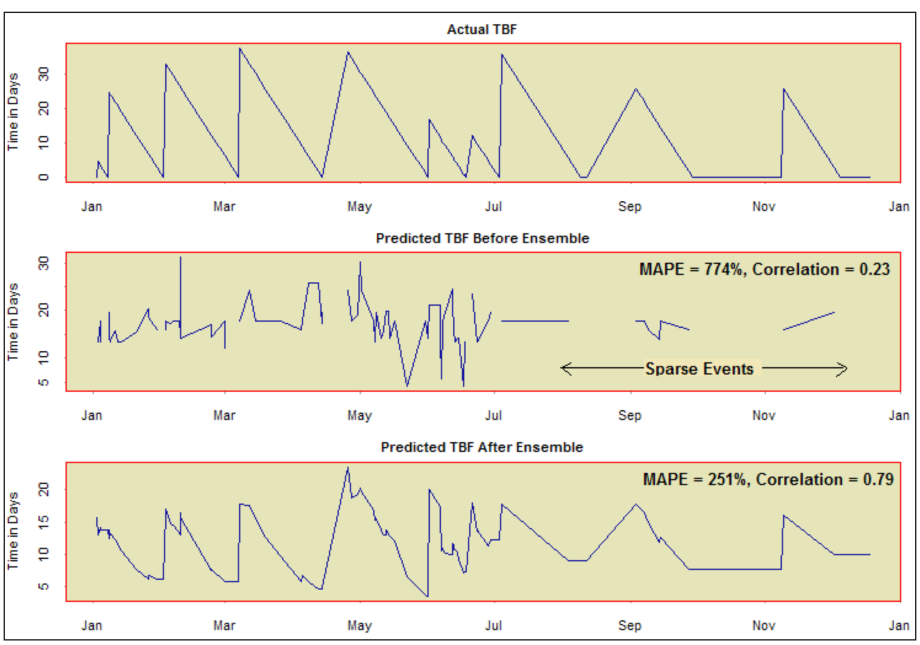

Figure 9: Visualization of Results for a specific Aircraft (Top) Actual TBF (Middle) TBF from Regression Tree Model (Bottom) TBF from Stochastic Ensemble

we observe that the Stochastic Ensemble of TBF distributions not only models the progressive deterioration and approximates the saw-tooth waveform observed in the actual TBF, it is also able to reduce the Mean Average Percentage Error (MAPE) by nearly 70% and improves the correlation by more than 300%, as compared to a standard regression tree model.

CONCLUSION

In this paper we examined the results provided by exploratory data analysis of aircraft health and failure data and used the insights for developing a predictive model for aircraft time to failure. We also explored the use of Hidden Markov Models for delay prediction and conclude that more appropriate training sequences can improve the HMM accuracy of predicting the delay state.

Finally, in this paper, we demonstrated how time adjusted probability distributions can be used with ensemble methods to effectively leverage stochastic information and model progressive deterioration of aircraft health. The approach also significantly improves prediction accuracy and correlation by combining predictions from several models.

REFERENCES

https://en.wikipedia.org/wiki/ATA_100 (Wikipedia ATA list)

Létourneau, S., Fazel, F., & Matwin, S. (1999). *Data Mining Based Component replacement*. National Research Council Cananda.

Salfner, F. (2005). Predicting failures with hidden Markov models. In *Proceedings of 5th European Dependable Computing Conference (EDCC-5)*, pages 41–46, Budapest, Hungary, Apr. 2005. Student forum volume.

Rabiner, L. R., & Juang, B. H. (1986). *An Introduction to Hidden Markov Models.* IEEE ASSP Magazine.