A Comparison of Data-driven Techniques for Engine Bleed Valve Prognostics using Aircraft-derived Fault Messages

Marcia Baptista[1], Ivo P. de Medeiros[2], Joao P. Malere[3], Helmut Prendinger[4], Cairo L. Nascimento Jr.[5], and Elsa Henriques[6]

[1,6] *Instituto Superior Tecnico, Universidade de Lisboa, Lisbon, 1049-001, Portugal*
marcia.baptista@ist.utl.pt, elsa.h@ist.utl.pt

[2,3] *Technol. Dev. Dept., Embraer SA, Sao Jose dos Campos, Brazil*
ivo.medeiros@embraer.com.br, joao.malere@embraer.com.br

[5] *Instituto Tecnologico de Aeronautica (ITA), 12228-900, So Jose dos Campos-SP, Brazil*
cairo@ita.br

[4] *National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan*
helmut@nii.ac.jp

Abstract

Prognostics plays an increasingly important role in preventive maintenance and aircraft safety. An approach that has recently become popular in this field is the data-driven technique. This approach consists in the use of past data and advanced statistics to derive estimates for the reliability of an equipment without relying on any physics or engineering principle. Data-driven models have been based on two types of historical data: past failure times and health monitoring data. A kind of health monitoring data rarely used in data-driven models are aircraft-derived maintenance messages. These data consist of fault messages derived from the aircraft onboard systems to notify any unexpected events or abnormal behavior as well as to send warning signals of equipment degradation. Fault messages have not received much attention in aircraft prognostics mostly due to its asynchronous and qualitative nature that often causes difficulties of interpretation. The main goal of this paper is to show that data-driven models based on fault messages can provide better prognostics than traditional prognostics based on past failure times. We illustrate this comparison in an industrial case study, involving a critical component of the engine bleed system. The novelty of our work is the combination of new predictors related to fault messages, and the comparison of data-driven methods such as neural networks and decision trees. Our experimental results show significant performance gain compared to the baseline approach.

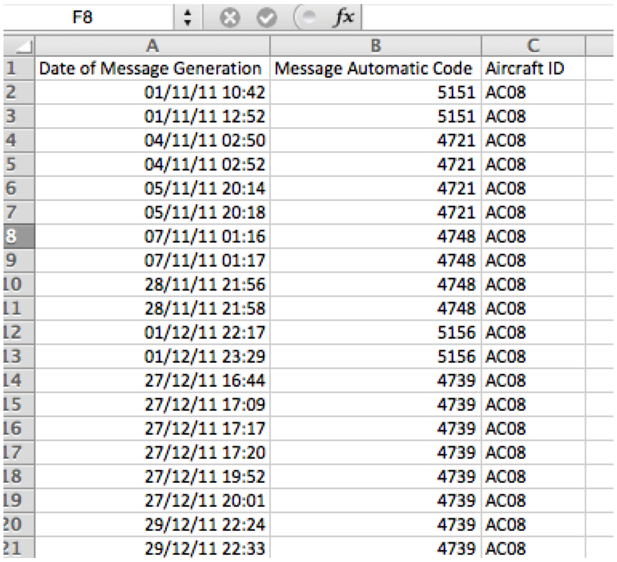
1. Introduction

Aircraft maintenance and repair operations, including unscheduled maintenance, account for 10-20% of the direct operating costs of an airline (Knotts, 1999). As a business with extremely thin margins (recorded in 2014 at 2.2%, IATA 2015 Annual Review), pressure to reduce costs while increasing service quality drives the need for better maintenance planning in the airline industry.

An engineering discipline claimed to be able to reduce maintenance costs by 25% (Camci, 2005) is failure prognostics. This discipline attempts to identify the best timing to conduct a maintenance action by predicting when the health condition of an equipment evolves beyond an acceptable threshold (Coble & Hines, 2011). Due to this active role in preventing system failure, prognostics can play a decisive role to improve airplane reliability, and lengthen maintenance check intervals in aviation.

A technique that has recently become popular in prognostics is the data-driven technique (Si, Wang, Hu, & Zhou, 2011; Schwabacher & Goebel, 2007). This approach is based on the assumption that the behaviors of a complex system cannot be fully grasped by a physically based model. Instead, advanced statistics and machine learning methods are used to learn a model directly from a set of data that is representative of all the behaviors found in the system (Schwabacher & Goebel, 2007).

Data-driven prognostics has been based on two types of data: survival (time to event) data and health monitoring data (Si et al., 2011). Survival data usually consist of failure or re-



	A	B	C
1	Date of Message Generation	Message Automatic Code	Aircraft ID
2	01/11/11 10:42	5151	AC08
3	01/11/11 12:52	5151	AC08
4	04/11/11 02:50	4721	AC08
5	04/11/11 02:52	4721	AC08
6	05/11/11 20:14	4721	AC08
7	05/11/11 20:18	4721	AC08
8	07/11/11 01:16	4748	AC08
9	07/11/11 01:17	4748	AC08
10	28/11/11 21:56	4748	AC08
11	28/11/11 21:58	4748	AC08
12	01/12/11 22:17	5156	AC08
13	01/12/11 23:29	5156	AC08
14	27/12/11 16:44	4739	AC08
15	27/12/11 17:09	4739	AC08
16	27/12/11 17:17	4739	AC08
17	27/12/11 17:20	4739	AC08
18	27/12/11 19:52	4739	AC08
19	27/12/11 20:01	4739	AC08
20	29/12/11 22:24	4739	AC08
21	29/12/11 22:33	4739	AC08

Figure 1. Fault messages registered between 01/01/11 10:42 and 29/12/11 22:33 for a commercial aircraft.

placement times (Moreira & Nascimento Jr., 2012). Health monitoring (HM) data consist of any data related to the estimation of the equipment current degradation. This definition includes aircraft sensory signals as well as inferred degradation signals.

A kind of health monitoring data rarely studied in aircraft prognostics are aircraft-derived fault messages. These messages consist of early warnings derived asynchronously by the aircraft onboard systems whose primary function is to signal a *fault*, that is, a deviation from standard operation (Isermann & Balle, 1997). Fault messages are derived from processing equations that range from simple constructions, such as when a sensory signal (e.g temperature) registers a sudden spike or exceeds a predetermined threshold, to more elaborate combinations of sensory signals.

Figure 1 shows an example file describing the fault messages of a commercial jet. In the file, each row represents a fault message that is characterized by a processing date and a fault class description (numeric code).

Research on the general field of fault-based prognostics is incipient for two main reasons. First, fault messages have been used mostly for diagnostics and fault analysis purposes (Strong, 2014). Second, these data are qualitative by nature and thus harder to analyze than quantitative data. The interpretation of a fault message is usually a nontrivial process, influenced by all the elements involved at its creation.

In this paper we investigate how fault messages can enhance prognostics in aviation. In particular, we aim to show that fault-based data-driven models can provide better performance than the simplest kind of prognostics, i.e prognostics based on survival data. This comparison is illustrated in an industrial case study involving the maintenance life cycle of a critical component of the jet engine.

The novelty of our work is the use of new predictors related to the fault messages of the aircraft maintenance system. Our proposed data-driven models combine fault data and a wide range of machine learning techniques. We study linear models, support vector machines, Bayesian models, instance-based learning, and decision trees. Please note that most of these techniques have already been studied in prognostics (Si et al., 2011) but not with fault data.

This paper is organized as follows. Section 2 provides an overview of related work. Section 3 describes our data set and methodology. Topics such as algorithms used and data processing steps are discussed. Section 4 presents our research hypothesis and results. A discussion of the results is presented in Section 5. Finally, Section 6 summarizes the major findings of this study and discusses future research work.

2. RELATED WORK

The first developments in aviation of prognostics and health management (PHM) technologies date back to to the Joint Strike Fighter (JSF) program launched in 1993. This program, costing nearly $1 trillion over the course of its lifetime, was able to design fighter jets with unprecedented processing and reasoning capabilities that relied on advanced sensor technology (Steidle, 1997).

Despite the initial efforts of the JSF program, only in recent years have the most advanced prognostics tools and methods been developed and implemented for the purposes of commercial aircraft maintenance (Heng, Zhang, Tan, & Mathew, 2009). Several approaches to prognostics have been proposed, ranging from low fidelity models, such as historical failure rate models, to high-fidelity physics-based models (Byington & Roemer, 2002).

Life usage models, or failure rate models, are often considered the simplest and most widely used form of prognostics (Schwabacher & Goebel, 2007). This kind of model is useful when a physical model of the component is unfeasible and there is insufficient sensor data to assess the equipment condition.

Life usage models rely on the survival times (time to event) of a large sample of components to predict the remaining time to a repair or failure of an individual component. Here, the variable of interest is the equipment operational life (Frangopol, Kallen, & Van Noortwijk, 2004) – predictions are based on the passage of time and/or measures of usage such as an airplane number of flights (i.e. cycles).

Rausand and Høyland (2004) propose different statistical distributions to model life usage, such as exponential, Weibull, normal, log-normal, logistic, log-logistic and Gamma distributions. In prognostics, the exponential and Weibull distributions are the most commonly used methods (Abernethy, 1996). The exponential distribution is simple and easy to apply and the Weibull has the ability to adjust to different reliability stages namely, infant, mature and wear out phases.

Another approach to prognostics is model-based algorithms. Here, "model-based" means a (more or less) hand-coded representation of human knowledge about the system (Schwabacher & Goebel, 2007). Traditional model-based techniques include state-space models (Isermann, 2006) and dynamic ordinary or partial differential equations (Vachtsevanos, Lewis, Roemer, Hess, & Wu, 2006).

The most sophisticated model-based techniques include rule-based expert systems such as SHINE (James & Atkinson, 1990) and Gensym G2 (GensymWebsite, 2007). Other examples of model-based techniques are finite-state machines (Williams & Nayak, 1996; Kurien & Nayak, 2000), and qualitative reasoning (Weld & Kleer, 1989).

Unlike electronic or electrical systems, mechanical equipment typically fails slowly as structural failures progress to a critical level. Monitoring these trends provides an opportunity to assess degradation and better estimate the remaining useful life of the equipment over a period of time. Here, the typical scenario is a slow but progressive change to a major structural failure mostly due to fatigue effects. Repetitive stresses induced by factors such as vibration loads (high cycle fatigue) or temperature cycles (low cycle fatigue) form frequently the basis for the equipment damage accumulation. This complex combination of variables often makes it difficult to develop a successful physics-based failure model (Mathur, Cavanaugh, Pattipati, Willett, & Galie, 2001). In such cases, empirical forecasting models, that is, data-driven approaches, are often preferred (Sankavaram et al., 2009).

Many data-driven techniques have been used in structural prognostics, from multivariate statistical methods, such as dynamic principal components (PCA), linear and quadratic discriminants and partial least squares (PLS), to black-box methods, such as neural networks, support vector machines, decision tree classifiers, graphical models and fuzzy logic (Pecht, 2008).

One of the most popular data-driven approach to prognostics is to use artificial neural networks to model the system (Goebel, Saha, & Saxena, 2008). The statistical performance of this technique has been shown in a number of real-life applications (Brotherton, Jahns, Jacobs, & Wroblewski, 2000; Wang & Vachtsevanos, 2001). Furthermore, neural nets, such as radial basis functions (RBFs), have been shown to be capable of "novelty detection" (identifying unexpected events in face of past history) (Brotherton & Johnson, 2001).

Despite the considerable number of works proposing new data-driven approaches to fault prognostics, comprehensive studies comparing the different approaches are rare. The few studies put forth seem to indicate that performance depends highly of the application domain. For instance, the work of He and Shi (2002) appears to contradict the established idea that neural networks produce the most accurate results. When studying valves in reciprocating pumps, the authors found that support vector machines (SVMs) yielded better accuracy than traditional neural nets in fault detection. These findings however, were only relative to their specific diagnostics field. Further research could determine whether and when approaches such as SVMs are able to produce more accurate prognostics results than neural nets.

Perhaps one of the most comprehensive comparison study on data-driven techniques is the work of Loyer, Henriques, and Wiseall (2014). Here, the authors compare a wide range of binary classifiers from linear regression to ensemble models to predict the probability of servicing a jet engine component at a major shop visit. Distinct approaches are discussed according to their ability to capture better or worse different perspectives of the data. The authors consider that ensemble models based on trees (random forests and boosted trees) present a good compromise between performance and interpretability while neural nets offer the best absolute performance.

In most data-driven models, input comes directly from routinely monitored sensory signals such as calorimetric, spectrometric and calibration data or power, vibration, acoustics, temperature, pressure, oil debris, currents and voltage data. Despite the importance of these data sources, some authors (Galar, Palo, Van Horenbeek, & Pintelon, 2012; Strong, 2014) have advocated the use of alternative data sources, such as the aircraft onboard control systems.

Data derived from onboard control systems consist in fault messages that aim to detect damaged or faulty equipment. These messages are widely used to define maintenance and repair procedures but have been seldom used in aircraft prognostics.

Please note that the term fault, as it is used here, means an unexpected deviation of at least one characteristic property or parameter of the system from the acceptable, or standard condition (Isermann & Balle, 1997). This notion differs from that of failure, which is a permanent interruption of the equipment ability to perform its function under its regular set of operating conditions (Isermann & Balle, 1997).

To isolate the failure of an aircraft equipment and eliminate as much cascade effects, onboard systems use fault processing equations to combine and elaborate on the various sources of flight data. As a result, fault data is different from sensor data in the sense that they are categorical and more prone to a subjective interpretation.

One of the few attempts to explore fault data in prognostics is the work of Strong (2014). Here, the author used fault messages to estimate the remaining time to failure of two industrial devices in a nuclear plant: an actuator and a motor. To construct its prognostics parameters Strong used simple methods based on averages and counts of the number of fault messages. A merging procedure was used to fuse the fault-based prognostic parameters to sensor data in a general path model (GPM) (Lu & Meeker, 1993). The main result of the study was that integrated prognostic parameters had significantly higher accuracy than parameters based solely on the information of fault messages.

Despite the importance of Strong's (2014) work, research on this topic could be improved in several ways. First, other modeling techniques besides GPM could be investigated. The GPM technique has been subject to some discussion (Garvey & Hines, 2007; Coble & Hines, 2011) due to use of a single degradation signal and the assumption of a failure threshold. A more complex combination of predictors could also provide more accurate results. It is along these lines that we present this work.

3. MATERIALS AND METHODS

In this section, we describe our data set and methodology. Section 3.1 presents an exploratory analysis of our data using statistical methods and unsupervised techniques such as K-means clustering. In section 3.2 we briefly discuss some of the methods and techniques used in our experiments.

3.1. Dataset

The experiments in this study are based on a real-world data set from a major aircraft producer. The data set reports on the 588 removals of a system of two identical bleed valves and on the 700 000 fault messages recorded from 39 commercial jets (two airline companies). The removals were recorded between January 2010 and June 2015 while the messages were collected between October 2011 and November 2014. The two-valve system studied here is considered a single system prone to failure, that is, as a multi-component system.

The two-valve system analyzed in this study is a critical element of the engine bleed air system – it allows the selection of either left, right or both engines as bleed air sources. Figure 2 illustrates in a simplified way the studied bleed system. The valves of interest, the engine bleed valves, are located between the compressor section of the engine(s) and the heat exchanger of the bleed system.

Engine bleed valves are *line-replaceable units* (LRUs) requiring frequent removal – they are designed to be removed and replaced quickly at the operating level in order to restore the engine bleed system to an operational ready condition. These valves are also *rotables* in the sense that they can be repeatedly and economically restored to a fully serviceable condition. Most often, these valves are replaced by new or repaired inventory items. Less frequently, engine bleed valves are removed, repaired on the fly and put back on the aircraft. This type of replacement accounts for around 20%-30% of all repairs in our data set.

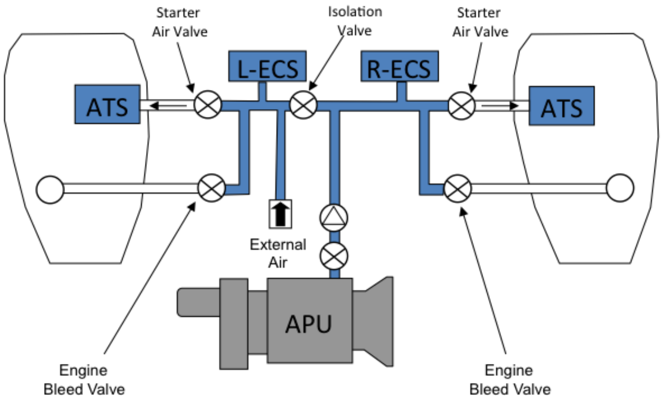


Figure 2. General schematics of an aircraft bleed air system.

The recording of removal dates is sometimes vulnerable to human mistakes by maintenance staff. Accordingly, we used a cleaning procedure to identify missing removals and other recording errors. In particular, the medcouple outlier method (Hubert & Vandervieren, 2008) was used to identify abnormal long/short time to removals. The results of this method were considered more plausible than results of the traditional box plot method (Tukey, 1977). While the latter approach detected 49 outliers (time to removal above 277 days), the medcouple detected only 4 outliers (time to removal above 653 days), a more reasonable proportion of 0.68% outliers in the overall set of 588 removals.

In our data set, time to (next) removal is a random variable with a probability density that resembles a Weibull distribution as shown in Figure 3. The cumulative distribution chart of Figure 4 illustrates how our empirical data sample is well fit to the theoretical Weibull model.

In addition to removal events, our data set also comprises fault data for the 39 jets. These data consists of all the automatic fault messages exchanged between the aircraft central maintenance computer (CMC) and ground facilities between October 2011 and November 2014. For each fault message, we have the following information: (1) date of message transmission and (2) processing code. These processing codes consist of 92 distinct numeric characters. Please review Figure 1 for an example of fault messages.

It is important to note that the analyzed fault messages do not provide direct information on the condition of the engine bleed valves. Instead, the messages convey information about the overall health of the bleed air system, such as when the system overall temperature goes beyond a given limit.

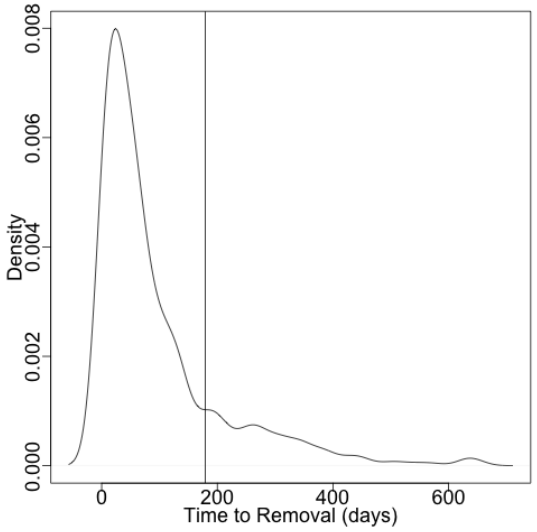


Figure 3. Probability Density Function (PDF) of time to next removal.

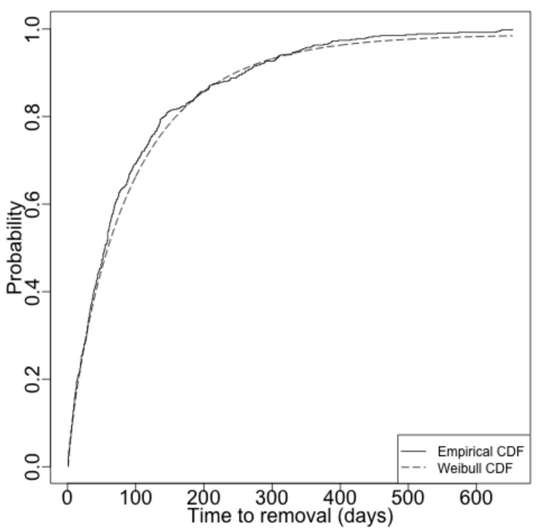


Figure 4. Cumulative Distribution Function (CDF) of time to next removal.

Since fault messages do not provide explicit information on the particular degradation of each unit in the bleed system, detection of whether and which assets require repair is a complex process. For instance, note Figure 5 where we show the arrival of 7 different types of messages for aircraft 24. Here, it is shown that the rate of messages (for instance messages 5290 and 5410) is often a trigger for a removal. However, it is not clear how the total number of a specific kind of messages triggers a removal – take the example of message 5290: sometimes it triggers a removal around the 20 messages while at other times this happens at 60 messages. This difference is often related to minimum equipment list (MEL) requirements.

Overall, our data set comprises around 5 and 2 hundred thousand messages for airline 1 and airline 2, respectively. From this set we ignored messages for which there was no previous or next removal as it was not possible to calculate accurate cumulative statistics for these messages. Overall, almost 150 thousand messages (19%) were disregarded in this process. For a graphical representation of this data cleaning process please see the event plot of Figure 6a. The plot illustrates the timeline of each of the 39 jets along the x-axis. Here, removals are marked as black circles, messages as black lines, and ignored removals and messages are marked in red.

Package NbClust (Charrad, Ghazzali, Boiteau, & Niknafs, 2012) combined with K-means detected two clusters of messages according to time to next message as shown in Fig. 6b. In the first group are messages spaced less than 1.4 hours from the next message. These messages are exchanged during flight. Messages spaced more than 1.4 hours from their next message are rare (15%) and may be evidence of longer aircraft stops or absence of warnings.

Regarding the cumulative sum of all messages, Fig. 6c shows that it is not clear that a threshold for a removal exists. The high dispersion (standard deviation of 2198 days) around the expected value of the number of messages at removal (2016 days) seem to indicate that this factor may not be a good sole predictor of a removal. The same seems to be true for the number of messages of a given code at removal.

3.2. Methodology

In this section we describe the methodology followed to investigate the hypothesis of our study:

> Predictive models based on fault data outperform traditional prognostics models based on survival analysis.

Our methodology consisted in comparing two prognostics approaches: (a) life usage models based on time to removal data and (b) data-driven models based on fault messages. The first approach is the traditional approach used in maintenance, where maintenance decisions (e.g., preventive hard time intervals) are determined based on statistical failure time analyses (Ahmad & Kamaruddin, 2012). The second (data-driven) approach combines a sophisticated type of health monitoring data – the fault messages derived in real-time from the aircraft onboard systems – with advanced techniques from machine learning and artificial intelligence. The main goal here is to improve the accuracy and online predictive power of condition-based prognostics models in aeronautics.

In this study the target variable was the remaining time to a removal. Model accuracy was evaluated and compared in terms of mean absolute error (MAE), root-mean-square (RMSE),

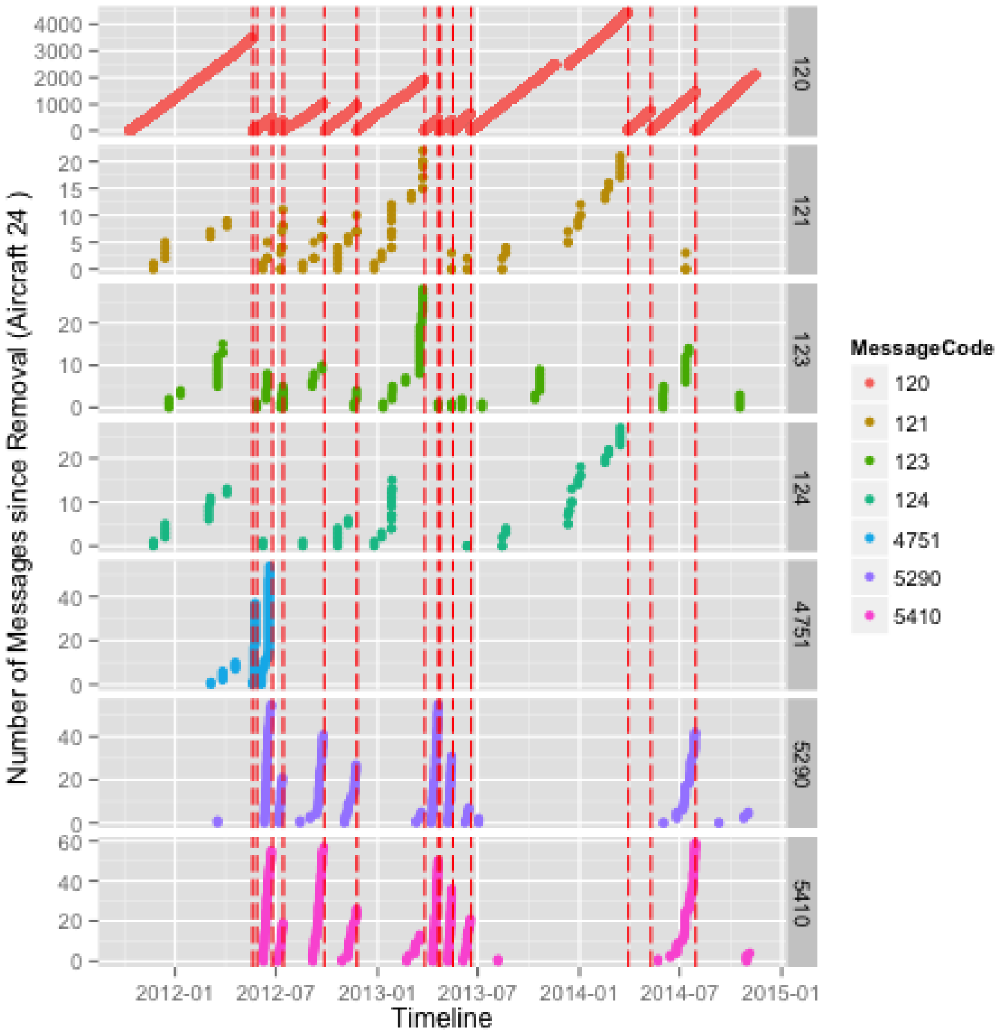


Figure 5. Timeline of aircraft 24.

Table 1. Performance metrics.

Metric	Abbr	Formula
Mean error	ME	$\frac{1}{N}\sum_{i=1}^{N}(\hat{T}_i - T_i)$
Root mean squared error	RMSE	$\frac{1}{N}\sum_{i=1}^{N}\sqrt{\left(\hat{T}_i - T_i\right)^2}$
Mean absolute error	MAE	$\frac{1}{N}\sum_{i=1}^{N}\left\lvert\hat{T}_i - T_i\right\rvert$

Note: N stands for number of observations. For each observation i, the model (either life usage or data-driven) predicts $\hat{T}_i$ for the T_i observed value. Here, variable T means remaining time to a valve removal.

and mean bias errors (average residuals ME). Table 1 details the metrics evaluated for both models.

The life usage approach consisted in applying the Weibull-Pareto distribution to our data set of removal times. To evaluate our Weibull analysis, 10-fold cross-validation was performed. In this procedure, each test fold was compared against a set of removal events generated from a Weibull model fit to the training data.

The data-driven approach was based on removal times and fault messages. Here, we developed distinct models using state-of-the-art data-driven techniques. In particular, we applied five of what have been considered the top 10 algorithms in data mining (Wu et al., 2008): k-nearest neighbor regression (KNN), regression trees (RPART R package), linear support vector machines (SVM), Bayesian generalized linear models (Bayes), and Gradient Boosting with Regression Trees (Boost trees). We also applied linear regression (LM) and neural networks (NN).

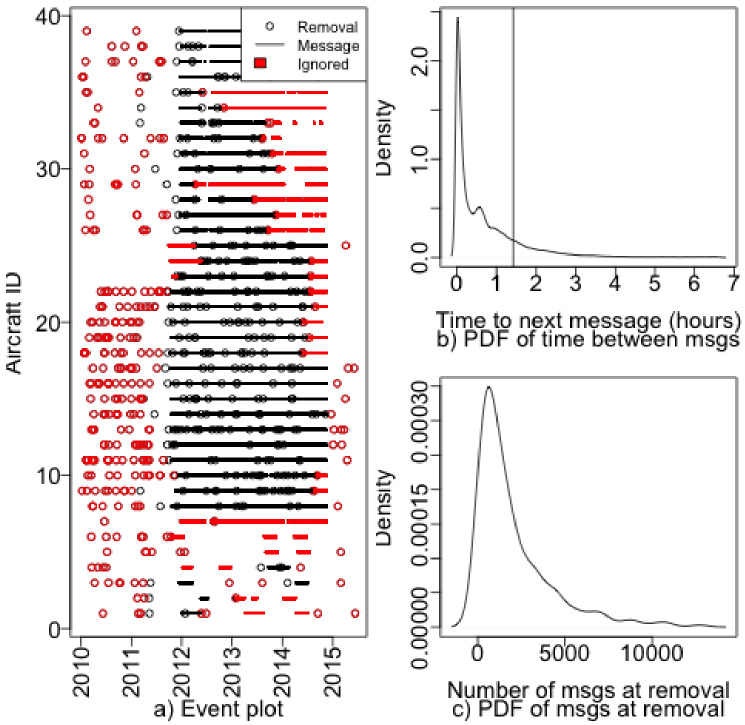


Figure 6. Event plot of data set (ignored messages and removals marked in red) and kernel density plots of time between messages and number of messages at removal.

The data frame of the data-driven models consisted of 520 thousand observations. Each observation described the arrival of a maintenance message and was characterized by 8 attributes: (1) past mean time to removal, (2) past variation of time to removal, (3) time since message, (4) time since message of same code, (5) number of messages since removal, (6) number of messages of same code since removal, (7) previous time to removal, and (8) variation between last two times to removals (drift). We found that two of these attributes, attribute 7 and 8, were highly correlated with each other with an absolute correlation higher than 0.75. Accordingly, we considered as our predictors all features except attribute 7. For a more detailed description of our correlation analysis please refer to Tab. 2, where the matrix of Pearson's r rank correlation coefficients for all possible pairs of attributes is presented.

Since the maintenance message data consisted of time-wise dependent data, we could not use the classical 10-fold cross-validation scheme to evaluate the data-driven models (Arlot, Celisse, et al., 2010, p. 65-66). Instead, a stratified cross-validation scheme was used. In the devised method, all the observations corresponding to messages within the same removal interval were in a single fold. This way, it was ensured that each training set contained information that occurred only after the testing sets.

4. RESULTS

In this section, numerical results are presented to illustrate the superiority of data-driven models based on fault messages over traditional life usage prognostics (Weibull analysis). Table 3 presents a comparison of all tested approaches concerning Mean Absolute Error (MAE), Mean Error (ME), Root Mean Squared Error (RMSE) and computational performance.

As illustrated in Table 3, our baseline, the Weibull model had, as expected, the worst predictive result of all approaches, both in terms of MAE and RMSE. The best data-driven approaches, Support Vector Machines (SVM) and Boosting trees (Boost), represented an increase of performance of 41.16% and 47.51% in terms of MAE and RMSE respectively.

Concretely, and as shown in Figure 7, the MAE of 110.04 days of the Weibull model was significantly higher than the MAE of the remaining approaches, which ranged from 83.70 days for the K-Nearest Neighbors (worst data-driven MAE result) to 64.75 days for the Support Vector Machines (best data-driven MAE result).

The RMSE results also favored the data-driven models. As shown in Figure 8, the RMSE of the Weibull model, 162.17, was considerably higher than the RMSE errors of the remaining approaches, which ranged from 117.88 for the Neural

Table 2. Correlation table of attributes.

Attribute	1	2	3	4	5	6	7	8
1 Past mean time to removal								
2 Past variation (std) of time to removal	0.70***							
3 Time since last message	0.01*	0.00						
4 Time since message of same code	0.02***	0.01***	0.05***					
5 Number of messages since removal	-0.07***	0.07***	-0.01***	0.04***				
6. Number of messages of same code since removal	-0.13***	0.07***	-0.02***	-0.07***	0.57***			
7. Previous time to removal	0.27***	0.29***	0.00	-0.02***	-0.13***	-0.05***		
8. Previous variation of time to removals	0.08***	0.26***	0.00	-0.01***	-0.09***	0.01**	0.92***	
9. Time to removal	0.13***	0.21***	0.01*	0.02***	-0.06***	-0.01*	-0.11***	-0.09***

Note that * means p<.05, ** means p<.01 and *** means p<.001

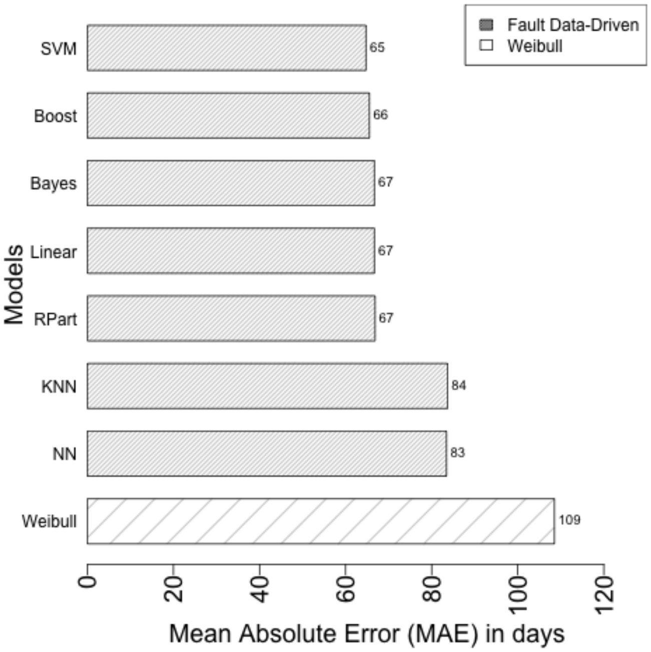


Figure 7. Performance based on Mean Absolute Error (MAE) for Weibull and regression models.

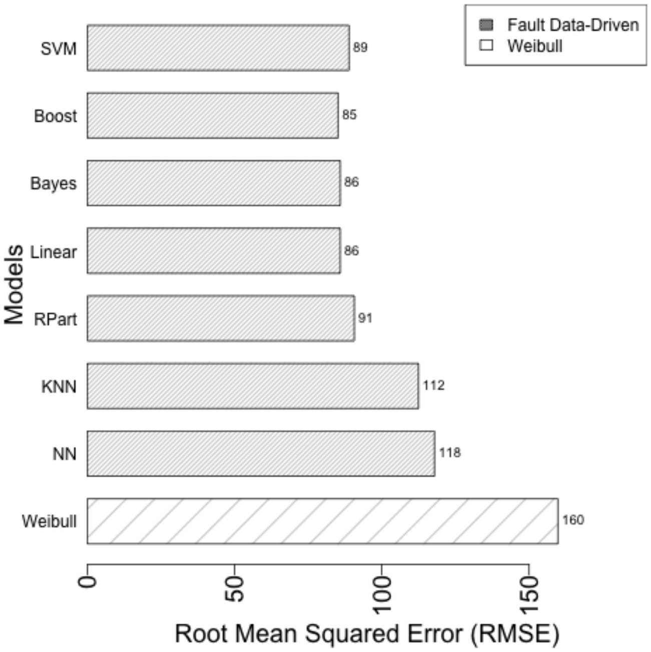


Figure 8. Performance based on Root Mean Squared Error (RMSE) for Weibull and regression models.

Nets (worst data-driven RMSE result) to 85.13 for the Boosting Trees (best data-driven RMSE result).

The graphical residual analysis shown in Fig. 9 reinforced the claim that data-driven models based on fault messages are superior to life usage models. Please note the presented residual analysis is based on standardized residuals (Osborne & Waters, 2002) instead of regular ones. Standardized residuals consist of mean errors (ME) divided by their standard error:

$$e_i^* = \frac{\hat{T}_i - T_i}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(T_i - \overline{T})^2}} \qquad (1)$$

Standardizing is a method for transforming residual data so that its mean is zero and standard deviation is one. One of the advantages of standardized residuals is that they quantify how large the residuals are in standard deviation units which allows for a more straightforward comparison of model results across distinct domains and applications. If the distribution of the residuals is approximately normal, then 95% of the standardized residuals should fall between -2 and +2. The residuals that fall outside of + or 2, then should be considered unusual.

As depicted by the regression (red) and smoothing spline (blue) lines in Figure 9a, the average standardized residuals of the Weibull model were $+2/-6$ which means that the model had acceptable underestimation errors ($+2$) but large overestimation errors (-6). In turn, the average standardized residuals of the data-driven models were $+2/-4$ and

Table 3. Comparison of model performance.

Metric	Weibull	NN	KNN	RPart	Linear	Bayes	Boost	SVM
Time	0.02	0.14	1.09	0.19	0.02	0.07	1.90	33.83
MAE	110.04	83.43 (24.18%)	83.70 (23.93%)	66.84 (39.26%)	66.70 (39.38%)	66.70 (39.38%)	65.52 (40.45%)	64.75 (41.16%)
RMSE	162.17	117.88 (27.31%)	112.43 (30.67%)	90.65 (44.10%)	85.86 (47.06%)	85.86 (47.06%)	85.13 (47.51%)	88.88 (45.20%)
ME	-5.06	-83.40	-0.15	0.01	0.75	0.75	-11.13	-21.78

* Time stands for mean processing time (seconds), ME for Mean Error (days) where ME = mean(simulated - observed), MAE for Mean Absolute Error (days) and RMSE for Root Squared Mean Error. The mean processing time is taken as an average over 5 experiments. Value in brackets in Time, MAE, and RMSE column indicate performance improvement (%) in regards to baseline, the Weibull model.

$+2/-2$. This means that these models were less prone to overestimate the remaining time to removal since their residuals never went beyond 2 or 4 standard deviations from the mean while the residuals of the Weibull model reached the 6 standard deviations less from the mean.

The Weibull model also showed a higher degree of residual heteroscedasticity than the data-driven models, that is, the size of the mean error (ME) differed more across values of the independent variable (remaining time to failure). Also, in this model residuals were becoming larger as the prediction of remaining time to a removal moved from small to large. In contrast, data-driven methods exhibited smaller error margins for the same predictions. Support Vector Machines (SVM) and Boosting Trees (Boost) were particularly efficient in minimizing errors – residuals were always less than -2, meaning that the majority of the prediction residuals were within less than 2 standard deviations from the mean.

Despite their higher accuracy the majority of the data-driven models exhibited a computational performance inferior to the Weibull model. As shown in Tabel 3, only the linear model (LM) had similar performance to the Weibull model. This can be explained by the fact that the Weibull and remaining models were based on two different data sets $-$ while the data-driven models were based on fault messages, the Weibull model was based on removal times. This means a data-driven model had to process 2000 more data than the Weibull model given the ratio of 2000 messages to 1 removal in the data set.

5. DISCUSSION

This study aimed to show that data-driven models based on aircraft-derived fault messages, could provide significant benefits over the traditional Weibull approach. Our initial results appear quite promising.

The use of a set of data-driven techniques, such as linear regression, support vector machines and boosting trees, combined with the fault message data, yielded a performance improvement of up to 41% and 48% in regards to Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). Despite this performance improvement, an analysis of the residual plots in Figure 9 shows that both approaches do not exhibit an optimal predictive behavior.

In an optimal prediction model residuals are expected to be randomly distributed around zero. In contrast, both the Weibull and data-driven approach exhibited a clear trend of increasing absolute errors with time to removal (Figure 9). This trend was however, not as dramatic for data-driven models as for Weibull models. Our interpretation of this finding is that despite their superiority, our proposed data-driven models can be further improved, most probably with additional features and/or model sophistication.

An analysis of the time series of removals and messages suggests the best features to further improve relate to fault messages and not necessarily to removals. In fact, the application of Ljung-Box tests on the sequence of removals of each aircraft did not provide enough evidence to reject the null hypothesis of randomness in all except two samples. Conversely, Ljung-Box tests attested that the sequence of messages exhibited a non-random pattern ($p < 0.01$) for all samples.

Regarding the selection of a data-driven technique there are no clear cut answers $-$ the tested data-driven methods exhibited similar performance (see Figure 7 and 8) with the exception of the Neural Networks (NN) and Nearest Neighbor (KNN) that had slightly worse results in MAE and RMSE.

It is also important to note that the computational efficiency of SVM deviated considerably from the other data-driven approaches. Despite the SVM technique having had acceptable MAE results and a comparatively good residual distribution (Figure 9h), its computational efficiency was poor, taking around 1.3 days to complete an experiment with our data set of 260 removals (500 thousand messages). This finding could lead to the detriment of this technique over alternatives such as boosting trees (see Table 3), which exhibited similar performance and superior computational efficiency.

6. CONCLUSION

Accurate predictions of equipment failure times are central to maintenance and inventory decisions in the aeronautics field. Most of the existing decision models focus on equipment life data from a single population, such as failure time distributions, to establish inspection, maintenance and repair policies. Since these distributions are unaffected by the under-

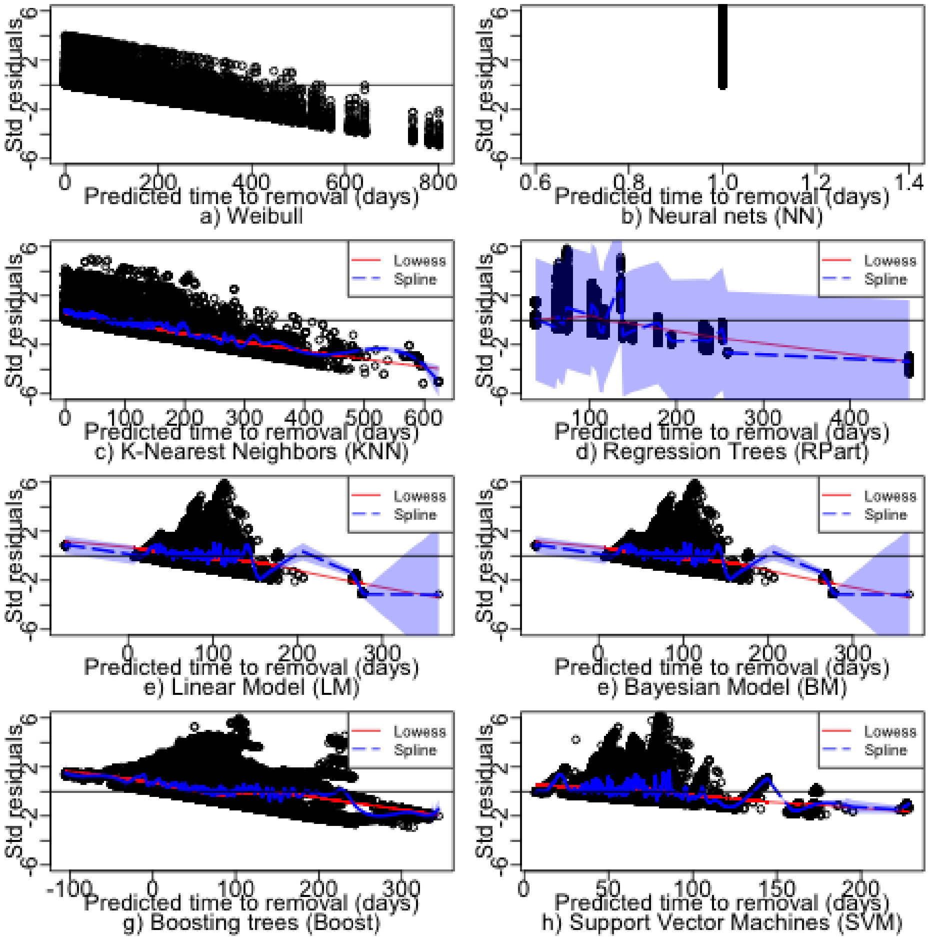


Figure 9. Standardized residual plots (residual = observed - simulated).

lying physical degradation processes, these are often unable to capture the particular degradation behavior of individual components of the population. This results in less accurate predictive power and hence less accurate replacement decisions.

Several works have attempted to explore sensor data streams (sensory signals) to enable the dynamic update of replacement decisions based on the physical condition of the equipment. This type of data-driven model shows promising results but also suffers from diverse limitations such as high volume data processing and low data quality.

In this paper, we propose an alternative data-driven model based on data derived from the aircraft processing computers. In particular, we study a kind of data which is commonly used by airline and aircraft manufacturers as the basis for their maintenance and repair decisions – aircraft-derived fault messages exchanged between the aircraft central computer and ground forces to report significant or abnormal conditions such as a temperature rise or excessive vibration.

The results of our study are encouraging reporting a performance improvement of up to 41% and 48% in Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) over the traditional life usage prognostics. Importantly, our results result from a scenario of a two component system. Accurate predictions for this kind of scenario are more difficult to obtain in comparison to the traditional single component system due to the complex interactions which exist between the recorded removal times. In view of our results, we hypothesize that our findings can be generalized, most probably with more success, to a scenario of a single unit system. Also, we assume that similar results can be obtained for field failure data or other type of maintenance data which not removals such as engine overhauls.

Future research should investigate which properties of the data-driven models based on fault messages most influence their predictive power. Most importantly, it is important to perform the following head-on-head comparisons: (1) life usage model based on time to removal versus machine learning models based on time to removal (2) life usage model based on fault data versus machine learning models based on fault data. Currently, strictly speaking, the improved performance cannot be uniquely attributed to the data-driven algorithms or fault data. A question rises: are the newly proposed models better (i) because the data-driven techniques are better than Weibull analysis, or (ii) because we included the fault data. Further research on this topic is needed.

It is also important to further understand the relative importance of predictors in our proposed data-driven models. Analyzing the influence of these predictors may provide useful insight into how to improve these models and why they perform better than the traditional Weibull model.

In this study, fault messages were shown to be a promising base for data-driven models. It is not clear however, the advantages and disadvantages it will bring over data-driven models based on other types of data such as data retrieved from aircraft sensors. For instance, fault-based models may lose accuracy to sensory based models but may compensate this in computational efficiency.

NOMENCLATURE

Nomenclature used in paper follows.

T	Remaining time to removal
MRO	Maintenance and Repair Operations
HM	Health Monitoring
PHM	Prognostics and Health Monitoring
JSF	Joint Strike Fighter
LRU	Line Replaceable Unit
GPM	General Path Model
CMC	Central Maintenance Computer
PDF	Probability Density Function
CDF	Cumulative Distribution Function
RUL	Remaining useful life
SVM	Support Vector Machines
NN	Neural Networks
MAE	Mean Absolute Error
$RMSE$	Root Mean Squared Error
ME	Mean Error (Standard Error)
$RPart$	Regression Trees - RPART R Package
$Boost$	Gradient Boosting with Regression Trees
$Bayes$	Linear Bayesian Model
$Linear$	Linear Regression
KNN	K-Nearest Neighbors

REFERENCES

Abernethy, R. B. (1996). *The New Weibull Handbook*. RB Abernethy.

Ahmad, R., & Kamaruddin, S. (2012). An Overview of Time-based and Condition-based Maintenance in Industrial Application. *Computers & Industrial Engineering*, *63*(1), 135–149.

Arlot, S., Celisse, A., et al. (2010). A Survey of Cross-Validation Procedures for Model Selection. *Statistics surveys*, *4*, 40–79.

Brotherton, T., Jahns, G., Jacobs, J., & Wroblewski, D. (2000). Prognosis of Faults in Gas Turbine Engines. In *Aerospace conference* (Vol. 6, pp. 163–171).

Brotherton, T., & Johnson, T. (2001). Anomaly Detection for Advanced Military Aircraft using Neural Networks. In *Aerospace conference* (Vol. 6, pp. 3113–3123).

Byington, C. S., & Roemer, M. J. (2002). Prognostic Enhancements to Diagnostic Systems for Improved Condition-based Maintenance (Military Aircraft). In *Aerospace conference* (Vol. 6, pp. 6–2815).

Camci, F. (2005). Process Monitoring, Diagnostics and Prognostics using Support Vector Machines and Hidden Markov Models. *Graduate School of Wanye State University, Detroit*.

Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2012). Nbclust Package: Finding the Relevant Number of Clusters in a Dataset. In *UseR! 2012*.

Coble, J., & Hines, J. W. (2011). Applying the General Path Model to Estimation of Remaining Useful Life. *International Journal of Prognostics and Health Management*, *2*, 71.

Frangopol, D. M., Kallen, M.-J., & Van Noortwijk, J. M. (2004). Probabilistic Models for Life-Cycle Performance of Deteriorating Structures: Review and Future Directions. *Progress in Structural Engineering and Materials*, *6*(4), 197–212.

Galar, D., Palo, M., Van Horenbeek, A., & Pintelon, L. (2012). Integration of Disparate Data Sources to Perform Maintenance Prognosis and Optimal Decision Making. *Insight-non-destructive Testing and Condition Monitoring*, *54*(8), 440–445.

Garvey, D. R., & Hines, J. W. (2007). Dynamic Prognoser Architecture via the Path Classification and Estimation (PACE) Model. In *Artificial Intelligence for Prognostics. in: AAAI fall symposium* (pp. 44–49).

GensymWebsite. (2007). Retrieved from `http://www.gensym.com`

Goebel, K., Saha, B., & Saxena, A. (2008). A Comparison of Three Data-driven Techniques for Prognostics. In *62nd Meeting of the Society for Machinery Failure Prevention Technology (MFPT)* (pp. 119–131).

He, F., & Shi, W. (2002). WPT-SVMs based Approach for Fault Detection of Valves in Reciprocating Pumps. In *American Control Conference* (Vol. 6, pp. 4566–4570).

Heng, A., Zhang, S., Tan, A. C., & Mathew, J. (2009). Rotating Machinery Prognostics: State of the Art, Challenges and Opportunities. *Mechanical Systems and Signal Processing*, *23*(3), 724–739.

Hubert, M., & Vandervieren, E. (2008). An Adjusted Boxplot for Skewed Distributions. *Computational Statistics &*

Data Analysis, *52*(12), 5186–5201.

Isermann, R. (2006). *Fault-diagnosis Systems: An Introduction from Fault Detection to Fault Tolerance*. Springer Science & Business Media.

Isermann, R., & Balle, P. (1997). Trends in the Application of Model-based Fault Detection and Diagnosis of Technical Processes. *Control Engineering Practice*, *5*(5), 709–719.

James, M. L., & Atkinson, D. J. (1990). Software for Development of Expert Systems.

Knotts, R. M. (1999). Civil Aircraft Maintenance and Support Fault Diagnosis from a Business Perspective. *Journal of Quality in Maintenance Engineering*, *5*(4), 335–348.

Kurien, J., & Nayak, P. P. (2000). Back to the Future for Consistency-based Trajectory Tracking. In *AAAI/IAAI* (pp. 370–377).

Loyer, J.-L., Henriques, E., & Wiseall, S. (2014). Comparison of Binary Classifiers for Data-driven Prognosis of Jet Engines Health. In *European Conference of the Prognostics and Health Management Society* (p. 1-12).

Lu, C. J., & Meeker, W. O. (1993). Using Degradation Measures to Estimate a Time-to-failure Distribution. *Technometrics*, *35*(2), 161–174.

Mathur, A., Cavanaugh, K. F., Pattipati, K. R., Willett, P. K., & Galie, T. R. (2001). Reasoning and Modeling Systems in Diagnosis and Prognosis. In *Aerospace/Defense Sensing, Simulation, and Controls* (pp. 194–203).

Moreira, R. d. P., & Nascimento Jr., C. L. (2012). Prognostics of Aircraft Bleed Valves using a SVM Classification Algorithm. In *Aerospace Conference* (pp. 1–8).

Osborne, J., & Waters, E. (2002). Four Assumptions of Multiple Regression that Researchers should always Test. *Practical Assessment, Research & Evaluation*, *8*(2), 1–9.

Pecht, M. (2008). *Prognostics and Health Management of Electronics*. Wiley Online Library.

Rausand, M., & Høyland, A. (2004). *System Reliability Theory: Models, Statistical Methods, and Applications* (Vol. 396). John Wiley & Sons.

Sankavaram, C., Pattipati, B., Kodali, A., Pattipati, K., Azam, M., Kumar, S., & Pecht, M. (2009). Model-based and Data-driven Prognosis of Automotive and Electronic Systems. In *Automation Science and Engineering* (pp. 96–101).

Schwabacher, M., & Goebel, K. (2007). A Survey of Artificial Intelligence for Prognostics. In *AAAI Fall Symposium* (pp. 107–114).

Si, X.-S., Wang, W., Hu, C.-H., & Zhou, D.-H. (2011). Remaining Useful Life Estimation – A Review on the Statistical Data Driven Approaches. *European Journal of Operational Research*, *213*(1), 1–14.

Steidle, C. E. (1997). The Joint Strike Fighter Program. *Johns Hopkins APL Technical Digest*, *18*(1), 7.

Strong, E. A. (2014). Development of a Method for Incorporating Fault Codes in Prognostic Analysis.

Tukey, J. W. (1977). Exploratory Data Analysis.

Vachtsevanos, G., Lewis, F., Roemer, M., Hess, A., & Wu, B. (2006). Intelligent Fault Diagnosis and Prognosis for Engineering Systems. *Usa 454p Isbn*, *13*, 978–0.

Wang, P., & Vachtsevanos, G. (2001). Fault Prognostics using Dynamic Wavelet Neural Networks. *AI EDAM*, *15*(04), 349–365.

Weld, D. S., & Kleer, J. d. (1989). *Readings in Qualitative Reasoning about Physical Systems*. Morgan Kaufmann Publishers Inc.

Williams, B. C., & Nayak, P. P. (1996). A Model-based Approach to Reactive Self-configuring Systems. In *Proceedings of the National Conference on Artificial Intelligence* (pp. 971–978).

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., ... others (2008). Top 10 Algorithms in Data Mining. *Knowledge and Information Systems*, *14*(1), 1–37.

BIOGRAPHIES

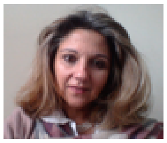
Marcia Lourenco Baptista (BS and MSc. in Informatics and Computer Engineering. Instituto Superior Tecnico, Lisbon, Portugal, September 2008) is a PhD candidate student at the Engineering Design and Advanced Manufacturing (EDAM) program under the umbrella of the MIT Portugal Program. Her research focuses on the development of prognostics techniques for aeronautics equipment in collaboration with Embraer Brazil. Her research interests include data-driven modeling, prognostics fusion, and uncertainty management.

Ivo Paixao de Medeiros Computer Engineer (2008, Federal University of Para), MSc. in Eletronics and Computer Engineering (2011, Aeronautics Institute of Technology) and Applied Computing DSc. candidate (National Institute for Space Research). His research focuses on Integrated Vehicle Health Management and Prognostics and Health Monitoring. He has been working for Embraer since 2010, first at Flight Safety Dept. and since 2012, at Technology Development Dept.

Joao Pedro Pinheiro Malere holds a bachelor degree in Control Engineering from Universidade Estadual de Campinas (Unicamp, 2004), Brazil, and a Master Degree in Aeronautical Engineering from Instituto Tecnologico de Aeronautica (ITA, 2007), Sao Jose dos Campos, Sao Paulo, Brazil. He is with EMBRAER S.A., So Jose dos Campos, So Paulo, Brazil, since 2006. He works as a Development Engineer in an R&T group at EMBRAER performing research on IVHM technology for application to aeronautical systems. His current research focus is on integrated health management systems.

Cairo Lucio Nascimento Junior received the B.S. degree in electrical engineering from the Federal University of Uberlandia (UFU, Uberlandia, Brazil) in 1984, the M.S. degree in electronics engineering from the Instituto Tecnologico de Aeronautica (ITA, So Jose dos Campos, Brazil) in 1988 and the Ph.D. degree in electrical engineering from the UMIST, Control Systems Centre (Manchester, UK) in 1994. Since 1986 he has been a lecturer with the Division of Electronic Engineering, ITA, and has supervised (or co-supervised) 9 Ph.D. and 27 M.Sc./Prof. M. students. He is the co-author of a book on intelligent systems and was the chairman for the 1999 Fourth Brazilian Conference on Neural Networks. His current research interests include autonomous systems, artificial intelligence, mobile robotics, control engineering, the development of internet-based remote-access laboratories for engineering education and data mining for knowledge extraction and application on fraud detection and PHM (Prognostics and Health Management).

Helmut Prendinger received his Master and Doctoral degrees in Logic and Artificial Intelligence from the University of Salzburg in 1994 and 1998, respectively. Since 2012, he is a full professor at the National Institute of Informatics (NII), Tokyo, after joining NII in 2004 as Associate Professor. Previously, he held positions as research associate (2000 2004) and JSPS postdoctoral fellow (1998 2000) at the University of Tokyo, Dept. of Information and Communication Engineering, Faculty of Engineering. In 1996-1997, he was a junior specialist at the University of California, Irvine. His research interests include artificial intelligence including machine learning, intelligent user interface, cyber-physical systems, and the melding of real and virtual worlds, in which areas he has published more than 220 peer-reviewed journal and conference papers. His vision is to apply his research to establishing the IT infrastructure for Unmanned Aerial Vehicles, or “drone”. He is a member of IEEE and ACM.



Elsa Maria Pires Henriques has a doctorate degree in Mechanical Engineering and is associated professor at Instituto Superior Tecnico in the University of Lisbon. She is responsible for the “Engineering Design and Advance Manufacturing (LTI/EDAM)” post-graduation. During the last fifteen years she has participated and/or coordinated several national and European R&D projects in collaboration with different industrial sectors, from tooling to automotive and aeronautics, mainly related to manufacturing, life cycle based decisions and management of complex design processes. She has a large number of scientific and technical publications in national and international conferences and journals. She was a national delegate in the 7th Framework Programme of the EU.