Transdisciplinary Engineering for Complex Socio-technical Systems – Real-life Applications
J. Pokojski et al. (Eds.)

doi:10.3233/ATDE200116

Use of Textual Elements to Improve Reliability Prediction for Aircraft Component Behavior

Wim J.C. VERHAGEN[a,1] and Thijs OUDKERK[b]
[a] *RMIT University, Aerospace Engineering & Aviation*
[b] *TU Delft, Air Transport & Operations*

Abstract. Unplanned maintenance is a costly factor in aircraft operations. Predictive maintenance models aim to provide greater insight into future component and system behaviour. In the state of the art, a variety of statistical models and machine learning techniques, amongst others, are used to estimate component remaining useful life. These approaches commonly leverage technical information, such as sensor data. However, the use of data and techniques from other domains is not prevalent. One such example is the application of natural language processing to incorporate textual information, e.g. derived from pilot complaint data. In other words, does the presence and specific content of pilot complaints have potential to improve the predictability of component removals? In this research, data integration and processing from multiple disciplines are combined to address this question. Relevant words from pilot complaints are identified using a term frequency–inverse document frequency (TF-IDF) numerical analysis, after which the most relevant words are used as covariates in a proportional hazards model. Left truncation and right censoring is applied to limit the time-invariant nature of these covariates. The results in the form of hazard ratios indicate a hazard increase of several orders of magnitude with respect to baseline hazard, pointing towards potential value of including these words as predictive parameters.

Keywords. Predictive Maintenance, Natural language processing, Proportional Hazard Models

Introduction

In aviation, maintenance plays a crucial role in ensuring continued aircraft airworthiness, allowing for safe operations of worldwide aircraft fleets. Beyond safety, maintenance also is crucial in determining the economic feasibility of aircraft operations; the right level of maintenance will prevent unscheduled and costly interventions, while allowing for smoothly aligned scheduled interventions. To enable this, insight into when an aircraft – or more typically, one of its systems or components – will fail is key. The field that studies the remaining lifetime of an object is known as survival analysis in general, and usually referred to as reliability analysis in engineering applications. Scientists and practitioners in this field have come up with a variety of statistical models to give insight into remaining life-time. A substantial part of these models are parametric, such as the Weibull distribution [1]. Where more flexibility is required, non-parametric models such

[1] Corresponding Author, Mail: wim.verhagen@rmit.edu.au.

as the Kaplan-Meier Estimator [2] can be employed. Though many models are univariate, a subset of models allow for a multivariate approach to survival analysis, like the semi-parametric Proportional Hazards Model, or Cox model after its inventor Sir David Cox [3], which has been applied predominantly in healthcare, with some examples in the engineering domain [4]. In such applications, operational parameters or physical parameters such as engine oil condition [5] are typically used. In more recent research, machine learning techniques are employed to leverage the sharply increasing availability of sensor data to more fully understand component deterioration and failure, as part of the fields of predictive maintenance and prognostics.

Despite these advances, one could argue that the most comprehensive sensor of them all, the pilot, has been overlooked as a source of data for research purposes. The pilot produces information in the form of natural language which is captured in pilot reports and pilot complaints. This information has significant potential not only for use in airline operations and maintenance (where this potential is largely realised through current-day regulations, procedures and processes), but also for predictive purposes. Through the application of natural language processing to incorporate textual information derived from pilot complaint data, it may be possible to provide improved predictability regarding upcoming component failures. In other words, does the presence and specific content of pilot complaints have the potential to improve in-service performance? In this research, data integration and processing from multiple disciplines are combined to address this question in the form of a proof-of-concept approach.

This is an example of where it is necessary and valuable to consider the inclusion of methodologies and stakeholders from multiple disciplines. As such, the problem at hand provides an example that falls under the banner of transdisciplinary research. In particular, several essential characteristics of transdisciplinarity are met by the problem at hand, namely 1) a process that starts from a real-world problem; 2) collaboration between and contribution of knowledge from different disciplines; 3) a shared overarchting goal from research and practice [6-8].

In summary, the purpose of this research is gauge the usability of the pilot complaints as an external source of data, and thereby test the applicability and effect of using textual information to improve reliability estimation. The structure of the remainder of this paper reflects this focus. First, the approach employed to tackle this problem is discussed in more detail in Section 1. Subsequently, this approach is applied towards a dataset comprising component removals and associated pilot complaint data. Section 2 discusses the characteristics of the dataset and the subsequent implementation of the approach, as well as the results. Finally, conclusions and recommendations for future research are given.

1. Methodology

To tackle the inclusion of pilot complaint information within component failure prediction, the methodology as set out in Figure 1 is proposed. It comprises a number of distinct steps, which are discussed below, with attendant theoretical concepts being further explained where deemed necessary.

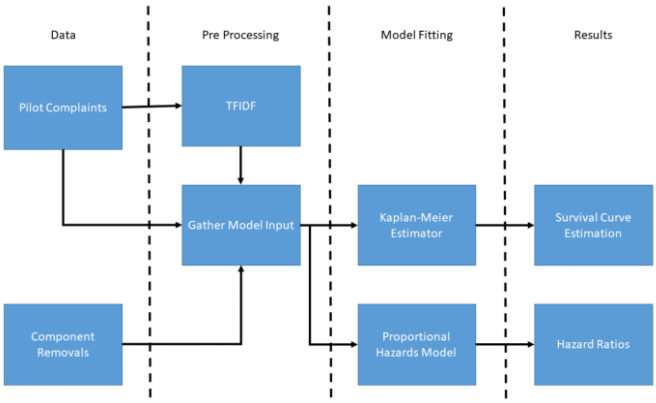

Figure 1. Methodological approach.

- **Data gathering:** the necessary data for estimating component reliability must first be gathered. This comprises pilot complaints and component removal data. The prior contain the textual data which must be processed to gather relevant modelling covariates (see below). The latter is primarily quantitative data which can be used to calculate baseline reliability characteristics. Section 2 describes the specific datasets used in this research in more detail.
- **Pre-processing:** before application in modelling and estimation, the data must first be pre-processed. For the component removal data, pre-processing involves cleaning, matching and selection of appropriate entries. More detail is again given in Section 2.
  To be able to use pilot complaint text as explanatory covariates within reliability models (see below), this text must first be converted to a numerical value. This pre-processing falls under the banner of natural language processing (NLP) [9]. In this research, pre-processing is performed using a basic natural language processing (NLP) technique, namely frequency-inverse document frequency (TF-IDF) as describes by Sparck Jones [10]. TF-IDF yields a measure of relevance for the processed text. In terms of the subject at hand, TF-IDF is used to score words based on how frequently they occur in the pilot complaints leading up to a removal, while correcting for its frequency in the entire corpus of pilot complaints. The scoring equation associated with the application of TF-IDF in this study is expressed in Equation 1 below.

$$a_{ij} = \log\left(tf_{ij} + 1\right) * \log\left(\frac{N+1}{n_j}\right) \quad (1)$$

With a_{ij} being the score of term j in document i, tf_{ij} being the term frequency of term j in document i, N being the total number of documents in the corpus, and n_j being the number of documents that term j appears in.

Model fitting: model fitting comprises the use of the pre-processed data to ensure two things: 1) modelling and estimating baseline reliability characteristics for comparative evaluation; 2) modelling and estimation reliability characteristics including the TF-IDF findings. With respect to 1), baseline reliability estimates are obtained by employing Kaplan-Meier

Estimators. The Kaplan-Meier Estimator, given below in Equation 2, is a method to estimate the survival function bases on (censored) lifetime data. As opposed to many statistical lifetime distributions, the Kaplan-Meier Estimator is non-parametric, meaning it does not adhere to a specific shape of distribution and therefore enjoys more flexibility.

$$\hat{S}(t) = \prod_{i:\, t_i \leq t}(1 - \frac{d_i}{n_i}) \tag{2}$$

With $\hat{S}(t)$ being the estimation of the survival curve, t_i being the time at which at least one removal occurs, d_i being the number of component removals, and n_i being the number of components that have not yet been removed.
With respect to 2), to allow for the inclusion of explanatory covariates, the Proportional Hazards Model is employed [3]. The PHM model models survival time while taking into account the effect of one or more explanatory variables, or covariates. The PHM model assumes time-independent covariates but time-dependent extensions are available in literature. In mathematical form, it can be represented as given in Equation 3. In this research, the main terms identified in the TF-IDF analysis are used as covariates within a PHM model (see Section 2) while assuming time-independence.

$$h(t|x) = b_0(t)\exp\left(\sum_{i=1}^{n} b_i(x_i - \bar{x_i})\right) \tag{3}$$

with $h(t|x)$ being the hazard function, $b_0(t)$ representing the baseline hazard, b_i representing the regression coefficients, x_i giving the covariate values, and $\bar{x_i}$ representing the covariates' average values.

- **Results:** application of Kaplan-Meier Estimators gives rise to estimated survival curves, representing reliability behaviour without incorporating the effect of pilot complaint-derived explanatory variables. In constrast, the PHM model output yields hazard ratios, which quantify the positive or negative influence of explanatory variables on the component hazard function (i.e., the instantaneous probability of failure).

The methodological approach is implemented and applied in a case study as described in Section 2.

2. Case study

To investigate whether the inclusion of textual information (in the 'raw' form of pilot complaints) could be used to improve reliability estimation and prediction, a case study has been carried out on the basis of a dataset provided by an independent Maintenance, Repair and Overhaul (MRO) organisation. For confidentiality reasons, the company is not identified. The dataset in question is described in more detail below. Modelling assumptions and application of the modelling approach is briefly discussed before moving towards the results for a specific component within the broader dataset, which serves as an example of the opportunities and pitfalls of applying the approach as proposed in Section 1.

2.1. Dataset characteristics

The involved maintenance service provider has performed maintenance activities across multiple airlines, aircraft manufacturers and aircraft types. Records are kept of all parameters relevant to the technical state of the aircraft and span a period of time from 1987 – 2016.

The relevant subsets within the overall dataset concern component removals and pilot complaints respectively, as they form the main source of information for further analysis. The relations between these two tables are depicted in Figure 2. Each data entry has a unique identifier (primary key), being the ”CompId” for a component removal and a ”PilotId” for pilot complaints. Both tables have the ”AircraftSerialNumber” as foreign key, being the entry used to link the data entry to a data entry in a foreign table. The most important information the data is the ”Date”, as insight into the date of a component removal is crucial towards reliability modelling.

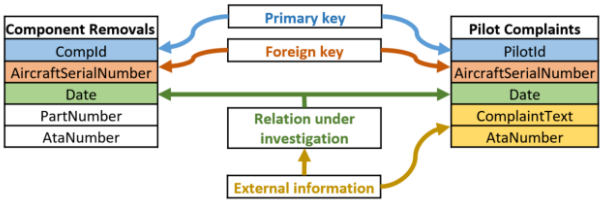

Figure 2. Primary datasets and attributes.

The component removals dataset spans a total of 476262 unique entries, whereas the pilot complaints comprises 428737 entries. However, not all of the data is suitable for further analysis as several issues contribute to significant dataset cleaning. The major three issues are 1) the appearance of non-English entries, where it has been decided to keep other languages than English out of the analysis to not complicate the NLP efforts; 2) absent aircraft serial numbers, where data entry has not been complete and therefore precludes linking specific component removals to specific pilot complaints; 3) quality of pilot complaint data, where the dataset has been constrained to post-2010 entries given that prior pilot complaint entries were sparsely and inconsistently captured.

In terms of NLP application, the pilot complaints that have been evaluated are typically comprised of sparse, keyword-like entries, sometimes including one or two brief sentences covering operational observations. To parse this information, entries have been made case-insensitive, with punctuation removed from textual entries. Furthermore, several synonyms (e.g. a/c, aircraft, airplane, etc.) have been merged into the analysis.

After cleaning, the datasets have been further reduced to enable a well-scoped, consistent data representation for use in analysis. The main steps here have been to constrain the datasets to occurrences from a single airline, within a single aircraft type, and selecting a top-five of components in terms of frequency of removals to arrive at a feasible scope of analysis. The results of the cleaning and reduction processes are given in Tables 1-2, with Table 3 providing an overview of the selected components and their removal numbers.

Table 1. Size of component removal data after various filtering and sampling steps.

	Component removals	Relative	Absolute
Raw	476.262	100%	100%
Filter dates	132.351	28%	28%
Missing registrations	102.451	77%	22%

Sample airline	21.761	21%	5%
Sample type	20.222	93%	4%
Sample components	3.101	15%	0.65%

Table 2. Size of pilot complaint data after various filtering and sampling steps.

	Component removals	Relative	Absolute
Raw	428.737	100%	100%
Filter dates	299.212	70%	70%
Missing registrations	295.746	99%	69%
Sample airline	96.951	33%	23%
Sample type	89.986	93%	21%

Table 3. Selected components (by frequency of removals).

Name	Description	Removals
Component 1	Oxygen bottle	2516
Component 2	Flow control valve	207
Component 3	Display unit	196
Component 4	Pressure Regulating Shut-off Valve	194
Component 5	Landing Light	176

2.2. Model assumptions and application

The main assumption to consider in the Proportional Hazards Model is explicitly part of its name. The hazard is assumed to be proportional to the baseline hazard. Equation 3 shows that the partial hazard merely scales the baseline hazard. Another assumption that follows from the model definition and the proportionality assumption is the fact that the effect a covariate has on the baseline hazard is constant in time. This last assumption is challenging regarding the nature of this research, since information from pilot complaints is very time-variant. Information is presented at some moment in time while being unknown before, and this information might become less relevant in time. This phenomenom is further illustrated in Figure 3, which depicts the situation where the birth is defined as the moment of installation of a component. The information in the pilot complaints is added somewhere between birth and death, death being the moment of component removal. It is evident that this information was not yet known before the onset of the pilot complaint. The covariate representing the pilot complaint or its content is therefore time-variant. Figure 4 shows the situation where the birth moment coincides with the onset of the pilot complaint. The information presented in the pilot complaint is known during the entire timeline and conceptually is not in violation of the proportionality assumption. The birth is therefore defined as the onset of each pilot complaint.

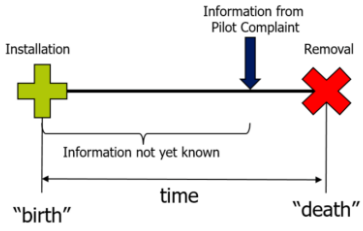

Figure 3. Moment of installation as birth.

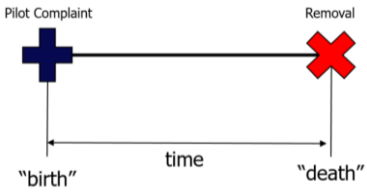

Figure 4. Moment of pilot complaint as birth

For each pilot complaint, the values of the following covariates are determined. The covariates consists of endogenous covariates and exogenous covariates. The former originate from the component removal data itself (and comprise variables "Year", "Fresh" and "Summer"), while the latter originate externally from the pilot complaint data. The following list give an overview of the covariates used in this research:

- **Year:** The year of complaint can gauge the effect time has on the hazard ratio.
- **Fresh:** This covariate has a value of 1 when the previous component removal was within two months of the pilot complaint under consideration. This is used to judge whether a recent installation has an effect on the hazard.
- **Summer:** This covariate has a value of 1 when the pilot complaint falls within the airline summer schedule and is used to discern any seasonality effects.
- **PN:** This covariate has a value of 1 if the part number is mentioned in the complaint text. This is used to analyze the lifetime patterns when it is known in advance that a removal will occur due to the pilot complaint in question, in effect serving as a validation set.
- **ATA:** This covariate is used to determine the effect that mentioning the specific subsystem has on the hazard.
- **word:***: This covariate has a value of 1 if the word represented by the asterisk is mentioned in the pilot complaint. This is used to measure the effect of certain words on the hazard.

2.3. Results

The TF-IDF complaint processing, Proportional Hazard Model and Kaplan-Meier Estimators have been applied to the selected 5 components (see Section 2.1). Here, some in-depth results are presented for component 5: landing light, as a representative case. Findings for the other components are briefly summarized at the end of this section.

The TF-IDF analysis for component 5 is represented in Table 4. It is clear that words that are functionally related to aircraft landings feature highly in the output. In this example, and in general as well, term relevancy quickly tapers off, indicating that specific words are relatively dominant in specific complaints.

Table 4. Overview of TF-IDF scores for Component 5: Landing Light.

Word	TF	DF	Score
Landing	30	3147	27.78
Extended	7	161	27.35
Retract	16	548	21.03
Lh	34	8061	20.28
light	5	20993	20.03

Results for the endogenous covariates are shown in Figure 5 and Table 5. It is clear that having information on the ATA chapter improves reliability estimation through the adjusted hazard rate - ATA scales the baseline hazard by 60% while being statistically significant. The same is true for the variable "year", though the effect is reversed and less pronounced in size.

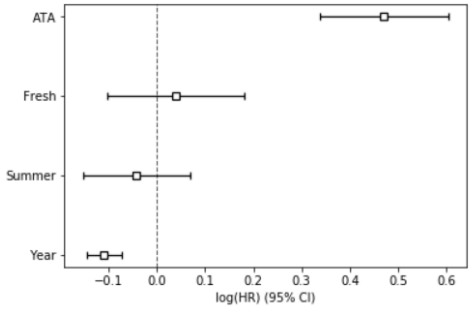

Figure 5. Forest plot of hazard ratios of Proportional Hazards Model fit of Component 5: Landing Light.

Table 5. Summary of Proportional Hazards Model characteristics for Component 5: Landing Light.

Endogenous variable	Coefficient	Exp(coefficient)	p-value	Proportionality assumption
ATA	0.47	1.60	<0.005	Met
Fresh	0.04	1.04	0.59	Met
Summer	-0.04	0.96	0.45	Met
Year	-0.11	0.90	<0.005	Met

Results for the exogenous covariates, i.e., the application of the Proportional Hazards Model for the four best scoring words in the TF-IDF analysis for this part, are shown in Figure 6. Note that the analysis for the word ”extended” is missing, due to excessive multicollinearity. Of the words, ”lh” misses statistical significance. The word ”retract”, although statistically significant, has a large standard deviation, as shown by the wide whiskers. The word ”landing” is the best performing word in this analysis, showing a hazard ratio of almost three, while being statistically significant and respecting the proportionality assumption.

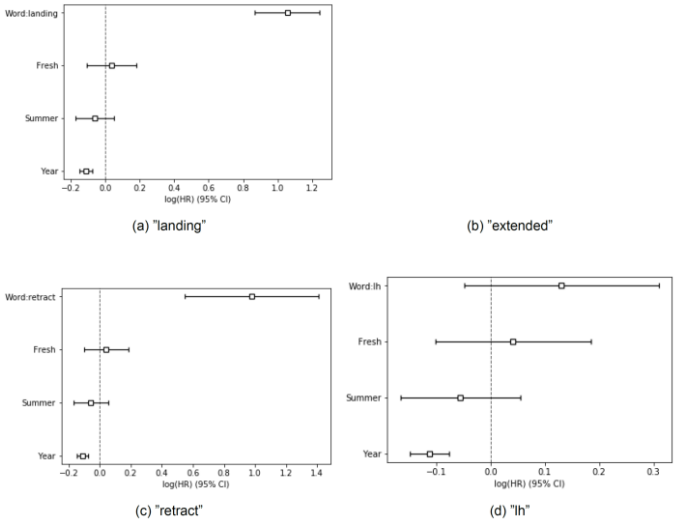

Figure 6. Forest plots of hazard ratios of Proportional Hazards Model fit of Component 5: Landing Light for different words.

Finally, Figure 7 shows a comparison between the Kaplan-Meier Estimator (KME), the adjusted hazard rate using the best-scoring word ("landing") and having the part number mentioned in the retrospectively added action in the pilot complaint. The latter is for validation purposes and shows the maximum predictive signal that one could obtain in a perfect world from textual entries. It is noticeable that including the word landing gives a slightly improved prediction of future failure behaviour, but does not come close towards full failure predictability as implied by having all information on complaint initiation, troubleshooting and resolution.

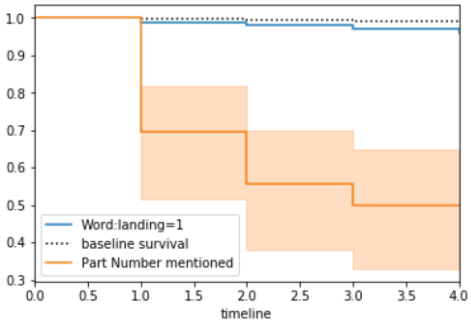

Figure 7. Comparison between Kaplan-Meier Estimator (baseline survival curve), KME with Part Number indication, and Proportional Hazards Model for Component 4: Landing Light for Word:landing.

Extending this analysis to other components shows similar patterns. For several components, the presence of multicollinearity precludes obtaining reliable results for specific words, but in general, the analysis of specific words in pilot complaints adds some increased predictability for component removals. However, the full potential of natural language processing of pilot complaints for forward-looking purposes is limited, unless the complaints are more detailed or incorporate maintenance troubleshooting information as well.

2.4. Discussion

Other situational factors may be at play which can further enrich the analysis provided here. For instance, the type, severity and frequency of complaints (especially in closely-spaced sequences) may help to distinguish slow- and fast-moving deterioration of components. Additional sources of textual information (such as maintenance inspection and shop findings) may further enhance reliability estimation by correlating pilot complaints with detailed characteristics (such as observed failure modes) associated with component removals, though sample sizes may be too small to find statistically meaningful results.

In terms of in-service implications, one critical aspect may be to consider the benefit of having an 'early-warning' function through real-time analysis of incoming pilot complaints. From this perspective, a knowledge-based diagnostic capability may be constructed by matching pilot complaints (or other sources of textual information) with prior cases and associated maintenance tasks.

3. Conclusions and recommendations

This research has presented a successful proof of concept, highlighting the potential use of textual information to enrich and improve reliability estimation. The hazard ratios resulting from the Proportional Hazards Model provide strong evidence for a statistically significant effect the information from the pilot complaint has on the hazard of a component removal. This effect however, is measured with respect to the baseline hazard. While the hazard in some cases increases more than sevenfold, one must also consider the absolute effect this has on the expected ”mortality”, which is severely limited by the short period under observation (in order to preserve time-independence). Furthermore, the limited descriptive content in the pilot complaints and the very simple NLP approach tested here do not provide deep insight into predicting towards future removal events.

These limitations can be addressed to some extent by considering more advanced NLP techniques in analyzing textual information that may be relevant towards component removals, especially techniques that (automatically) group synonyms or syntactically similar words. Furthermore, the major assumptions that had to be made with respect to time-invariant behaviour could be resolved by considering time-variant proportional hazard models. However, this would come at the cost of computational performance.

References

[1] W. Weibull, A statistical distribution function of wide applicability, *Journal of applied mechanics*, Vol. 18, No. 3, 1951, pp. 293–297.

[2] E.L. Kaplan and P. Meier, Nonparametric estimation from incomplete observations, *Journal of the American statistical association*, Vol. 53, No. 282, 1958, pp. 457–481.

[3] D.R. Cox, Regression models and life-tables, *Journal of the Royal Statistical Society: Series B (Methodological)*, Vol. 34, No. 2, 1972, pp.187–202.

[4] W.J.C. Verhagen and L.W.M. De Boer, Predictive maintenance for aircraft components using proportional hazard models, *Journal of Industrial Information Integration*, Vol. 12, 2018, pp. 23–30.

[5] A.K.S. Jardine, P.M. Anderson and D.S. Mann, Application of the weibull proportional hazards model to aircraft and marine engine failure data, *Quality and reliability engineering international*, Vol. 3, No. 2, 1987, pp. 77–82.

[6] R.W. Scholz and G. Steiner, The real type and ideal type of transdisciplinary processes: part I – theoretical foundations, *Sustainability Science*, Vol. 10, No. 4, 2015, pp. 527-544.

[7] A. Ertas, Understanding transdiscipline and transdisciplinary process, *Transdisciplinary Journal of Engineering Science,* Vol. 1, No. 1, 2010, pp. 55-73.

[8] N. Wognum, C. Bil, F. Elgh, M. Peruzzini, J. Stjepandić, and W.J.C. Verhagen, Transdisciplinary systems engineering: implications, challenges and research agenda, *International Journal of Agile Systems and Management*, Vol. 12, No. 1, 2019, pp. 58-89

[9] C.D. Manning and H. Schütze. *Foundations of statistical natural language processing*. MIT press, Boston, 1999.

[10] K. Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, Vol. 28, No. 1, 1972, pp.11–21.