

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/325943740>

Proceedings of the 10th IMA International Conference on Modelling in Industrial Maintenance and Reliability

Book · June 2018

CITATIONS

0

READS

2,313

3 authors:



[Phil Scarf](#)

Cardiff University

149 PUBLICATIONS 4,524 CITATIONS

[SEE PROFILE](#)



[Shaomin Wu](#)

University of Kent

184 PUBLICATIONS 4,213 CITATIONS

[SEE PROFILE](#)



[Phuc Do](#)

University of Lorraine

109 PUBLICATIONS 2,390 CITATIONS

[SEE PROFILE](#)



Proceedings of the 10th IMA International Conference on Modelling
in Industrial Maintenance and Reliability
Manchester, UK
12-15 June, 2018

Edited by:
Philip Scarf, Shaomin Wu, Phuc Do

ISBN: 978-0-905091-32-7

Table of contents

- 1 Imprecise statistical inference for accelerated life testing data: imprecision related to the likelihood ratio test
A.A.H. Ahmadini and F.P.A. Coolen
- 7 New method for availability computing of complex systems using imprecise Markov models
J. Akrouche, M. Sallak, E. Chatelet, F. Abdallah and H. Hajj Chehade
- 13 Modelling the quality of maintenance of a protection system: a case study on isolation valves used in water distribution systems
A.R. Alberti, C.A.V. Cavalcante, P.A. Scarf and A.L.O. Silva
- 19 Mitigating warranty risk for automotive industry
A. Aljazea, M. Luo and S. Wu
- 25 Integration of reliability, availability and maintainability analysis and operational research modelling - a case study
J. Bishop, M. Niblett and M. McCarthy
- 31 Estimation and goodness-of-fit tests for imperfect maintenance models: single system vs multi-system
L. Breniere, L. Doyen and O. Gaudoin
- 36 Applications of the gamma process in storage models: energy harvesting systems
I.T. Castro, L. Landesa and A. Serna
- 42 Multicriteria model to support a definition of a maintenance - spare parts policy
P.P.L. Dantas, T.F.O. Rodriguez and C.A.V. Cavalcante
- 48 Remaining useful life estimation methods for predictive maintenance models: defining intervals and strategies for incomplete data
A. Delmas, M. Sallak, W. Schon and L. Zhao
- 54 A framework for maintenance 4.0
E. Jantunen, G. Di Orio, F. Larrinaga, M. Becker, M. Albano, and P. Malo
- 60 Modeling spare part demand data using transmuted or exponentiated Poisson distribution
R. Jiang and Z. Chen
- 66 Optimal sequential resource allocation under error-prone success assessment
K. Kalyanam, M. Pachter and D. Casbeer
- 73 Practical uses of the Weibull distribution for decision making
G. Knight
- 79 A pitch control system deterioration model considering wind speed turbulence intensity
J. Ma, M. Fouladirad and A. Grall
- 85 Optimising remedial outcomes for gas turbines through large scale data analysis
J. McGinty, S. Kollias and G. Leontidis
- 91 Vulnerability assessment modelling for railway networks
R. Meesit and J. Andrews
- 97 Cost effective component swapping to increase system reliability
A. Najem and F.P.A. Coolen
- 103 Reliability modelling of dispensing processes in community pharmacy
M. Naybour, R. Remenyte-Prescott and M.J. Boyd
- 109 Joint maintenance scheduling and routing optimization for geographically dispersed production systems
H.S.H. Nguyen, P. Do, B. Lung and H.C. Vu
- 116 Maintenance engineering management of a LNG plant critical system: a case study of gas turbine dry gas seal
E. Pereira, B.M. Alkali and O. Niculita
- 125 A selective maintenance model based on FITradeoff method
L. Ribeiro, P. Dantas and C.A.V. Cavalcante
- 131 Selective maintenance model for mission-oriented systems considering heterogeneous missions and budget constraints
L. Ribeiro and C.A.V. Cavalcante
- 135 Reliability demonstration tests considering performance degradation with measurement error
C. Ruiz, H. Liao and E.A. Pohl

- 141 Insight into cost of defective state using delay-time concept
A.C.J. Santos and C.A.V. Cavalcante
- 147 A two-dimensional availability-based warranty policy incorporating preventive maintenance
C. Su and L. Cheng
- 153 Computational fluid dynamics or Gaussian - is there a right way to model gas dispersion ?
A. Tripathi, C. Souprayan, A. Stanley and N. Warrilow
- 160 Research on voice interaction for augmented reality assisted maintenance
R. Wang, C. Lyu, S. Mei, C. Guo, J. Geng and Y. Xiao
- 166 An unpunctual preventive maintenance policy for repairable items sold with a two-dimensional warranty
X. Wang, L. Li and M. Xie
- 172 Decentralisation of reliability data through blockchain
G. Willoughby, T. Ingram, T. Byrne, D. Smith and N. Rahman
- 178 Assurance for sample size determination in binomial reliability demonstration testing
K.J. Wilson and M. Farrow
- 184 Maintenance centered matrix-based universal quality characteristics analysis
Y. Xiao, C. Lv, J. Geng, C. Guo and R. Wang
- 190 Dynamic imperfect condition-based maintenance for systems subject to nonlinear degradation paths
X. Zhao, Z. Liang, A. Parlikad and M. Xie
- 195 A visualization automated analysis and evaluation method based on virtual reality
Q. Zhou, D. Zhou, Y. Xie and W. Zhou
- 203 **Index of Authors**

Imprecise statistical inference for accelerated life testing data: imprecision related to the likelihood ratio test

Abdullah A.H. Ahmadini^{1,2} and Frank P.A. Coolen¹

^{1,2}*Department of Mathematics Science, Durham University, Durham, United Kingdom*

²*Department of Mathematics, Faculty of Science, Jazan University, Jazan, Saudi Arabia*

Abstract In this paper, we present a new imprecise statistical inference method for accelerated life testing data, where nonparametric predictive inferences at normal stress levels are integrated with a parametric Arrhenius-Weibull model. The method includes imprecision which provides robustness with regard to the model assumptions. The imprecision leads to observations at increased stress levels being transformed into interval-valued observations at the normal stress level, where the width of an interval is larger for observations from higher stress levels. Simulation studies are presented to investigate the performance of the proposed method.

1. Introduction

Accelerated life testing (ALT) has frequently been used to obtain information on the lifespan of devices. Testing items under normal conditions can require a great deal of time and expense. To determine the reliability of devices in a shorter period of time, and with lower costs, ALT is frequently used. In ALT, a unit is tested under levels of physical stress (e.g. temperature, voltage, or pressure) greater than normal operating conditions. Using this method, devices will fail more quickly, enabling estimation of the lifetime under normal conditions via extrapolations using an ALT model. There is a wide range of designs for ALT, including constant-, step- and progressive-stress testing [Nelson 1990]. Two components usually used to extrapolate a device's lifespan under normal use conditions from ALT results are the life time distribution (e.g. Weibull or log-normal) and the relationship between failure time and stress level (e.g. Arrhenius law or power law).

In this paper, we consider the Arrhenius model and a Weibull lifetime distribution for a constant-stress ALT. The Arrhenius model is based on a physical/chemical theory, and thus is often an appropriate model to use when the failure mechanism is driven by temperature [Nelson 1990]. The Weibull distribution is usually used for examining component, system or product life. The Arrhenius-Weibull model is adapted to the current research on imprecise statistical approaches to establish the use of imprecision in modelling the nature of the relationship between stress levels and unit failure rates. The main issue here is how to extrapolate the failure data from units tested at higher-than-normal stress levels to units operating at normal level.

To explain the Arrhenius-Weibull model assumptions, we must understand that it is based on differential failure times at varying stress levels. For example, K_0 represents the actual stress level, and there are $m \geq 1$ increased stress levels, with stress K_i at level $i \in \{1, \dots, m\}$. We parameterize the Weibull distribution such that its survival function is

$$P(T > t) = \exp \left[- \left(\frac{t}{\alpha} \right)^\beta \right],$$

with α the scale parameter and β the shape parameter. The Weibull distributions for different stress levels are assumed to have different scale parameters, α_i for level i , but the same shape parameter β . Therefore, the link function of scale parameter α_i should be used for establishing a connection between different stress levels i , and is assumed to satisfy the function

$$\alpha_i = \alpha_0 \exp(\gamma/K_i - \gamma/K_0). \quad (1)$$

Where using this model with temperature as stress levels, K_0 is the normal temperature (Kelvin) at stress level 0, K_i is the higher temperature (Kelvin) at stress levels i , and γ is the parameter of the Arrhenius link function model. So there are in total three parameters that need to be estimated, α_0 , β , and γ .

In this work, we will apply the pairwise likelihood ratio test to investigate equality of different scale parameters, with shape parameter being equal for all stress levels, as explained in Section 2 [Nelson 1990, Alam and Sudhir 2015]. In this paper, we assume the same β for all levels, so difference between stress levels is only modelled through the α_i , and hence in the model through the γ parameter of the link function between different

stress levels. We develop the use of likelihood ratio test to obtain the interval of values of the γ parameter of the link function for which the null hypothesis that two groups of failure data come from the same underlying distribution, is not rejected. In this paper, we assume that there are no right-censored observations. Note further that at the normal stress level 0 we assume to have failure observations. We briefly comment on this in the last section.

In this paper, we develop a new likelihood ratio test based method for analysing ALT data based on imprecise probabilities, where nonparametric predictive inferences (NPI) at a normal stress level are integrated with a parametric Arrhenius-Weibull model. This new method consist of two steps. We assume Arrhenius-Weibull as a complete model for all levels in the first step. Then we derive an interval for γ values of the Arrhenius link function model by pairwise likelihood ratio tests, which allows observations of increased stress levels to be transformed to interval-valued observations at the normal stress level in the second step. Then we use NPI at the normal stress level for predictive inference on the failure time of a future unit operating under normal stress.

Nonparametric Predictive Inference (NPI) is a statistical method, which provides lower and upper survival functions for a future observation based on past data using imprecise probability [Augustin, et al 2014, Coolen 2011]. [Hill 1968] proposed an assumption which gives direct conditional probabilities for a future random quantity which depend on the values of related random qualities [Augustin and Coolen 2004, Coolen 2006]. Given ordered observations $x_1 < x_2 < \dots < x_n$, and defining $x_0 = 0$ and $x_{n+1} = \infty$, the NPI lower and upper survival functions for a future observation X_{n+1} are

$$\underline{S}_{X_{n+1}}(t) = \frac{n-j}{n+1}, \text{ for } x \in (x_j, x_{j+1}), j = 0, \dots, n. \quad (2)$$

$$\bar{S}_{X_{n+1}}(t) = \frac{n+1-j}{n+1}, \text{ for } x \in (x_j, x_{j+1}), j = 0, \dots, n. \quad (3)$$

The imprecision, the difference between the upper and lower survival functions, reflects the amount of information in the data.

This paper is organized as follows: in Section 2, the novel method of imprecise predictive inference is introduced based on ALT failure data and pairwise likelihood ratio tests. In Section 3 we illustrate our proposed method in an example. Section 4 presents results of initial simulation studies that investigate the performance of the proposed method, where the data are simulated using the assumed Arrhenius-Weibull model. Section 5 presents concluding remarks.

2. Predictive inference based on ALT data and pairwise likelihood ratio test

The aim of this section is to combine NPI with the fully parametric model used in our new statistical method analysing data from ALT. The use of NPI here provides lower and upper survival functions for a future observation at the normal stress level, based on all failure data. This new statistical method for data in ALT divides neatly into two steps. First, the basic Arrhenius-Weibull model is adopted [Nelson 1990], and the pairwise likelihood ratio test is used between the stress levels K_i and stress level K_0 , to obtain the intervals $[\underline{\gamma}_i, \bar{\gamma}_i]$ for the parameter γ for which we do not reject the null hypothesis that the data transformed from stress level i to normal stress level 0, and the original data obtained at the normal stress level 0, come from the same Weibull distribution, where $i = 1, \dots, m$. With these m pairs $[\underline{\gamma}_i, \bar{\gamma}_i]$, we define $\underline{\gamma} = \min \underline{\gamma}_i$ and $\bar{\gamma} = \max \bar{\gamma}_i$.

The Arrhenius link function for scale parameters α_i should be identified to establish a connection between the different stress levels i . Regarding to this link function model, an observation t^i at the stress level i , subject to stress K_i , can be transformed to stress level 0, with the transformed observation from level i to level zero represented by the equation

$$t^{i \rightarrow 0} = t^i \exp \left(\frac{\gamma}{K_0} - \frac{\gamma}{K_i} \right). \quad (4)$$

In the second step, we apply the data transformation using $\underline{\gamma}$ for all $i = 1, \dots, m$ to get transformed data at level 0 which leads to NPI lower survival function $\underline{S}$. Similarly, we apply the data transformation using $\bar{\gamma}$ for all $i = 1, \dots, m$ to get transformed data at level 0 which leads to NPI upper survival function $\bar{S}$. Note that each

Table 1: Failure times at three temperature levels

Case	$K_0 = 283$	$K_1 = 313$	$K_2 = 353$	$K_1 = 313 (*1.4)$	$K_2 = 353 (*0.8)$
1	2692.596	241.853	74.557	338.595	59.645
2	3208.336	759.562	94.983	1063.387	75.987
3	3324.788	769.321	138.003	1077.050	110.402
4	5218.419	832.807	180.090	1165.930	144.072
5	5417.057	867.770	180.670	1214.878	144.560
6	5759.910	1066.956	187.721	1493.739	150.176
7	6973.130	1185.382	200.828	1659.535	160.662
8	7690.554	1189.763	211.913	1665.668	169.531
9	8189.063	1401.084	233.529	1961.517	186.823
10	9847.477	1445.231	298.036	2023.323	238.429

Table 2: Accepted γ for likelihood ratio

Significance level	0.99		0.95		0.90	
Stress level	$\underline{\gamma}_i$	$\bar{\gamma}_i$	$\underline{\gamma}_i$	$\bar{\gamma}_i$	$\underline{\gamma}_i$	$\bar{\gamma}_i$
Case 1: K_1 K_0	4060.018	6605.752	4424.881	6261.168	4593.700	6100.653
	K_2 K_0	4377.043	5602.321	4550.205	5434.908	4630.511
Case 2: $K_1 * (1.4)$, K_0	3066.539	5612.273	3431.402	5267.689	3600.221	5107.174
	$K_2 * (0.8)$, K_0	4695.498	5920.775	4868.660	5753.363	4948.965

observation at an increased stress level transforms into an interval-valued observation at the normal stress level 0, where the width of an interval is larger for an observation from a higher stress level.

Note that, if the model fits really well, we expect most $\underline{\gamma}_i$ values to be quite similar, as well as most $\bar{\gamma}_i$ values. If the model fits poorly, $\underline{\gamma}_i$ are most probably very different, or $\bar{\gamma}_i$ are very different, or both. The proposed method in Section 2 is illustrated by an example using simulated data in Section 3.

3. Example

In this section we present two cases. In case 1 we simulated data at all levels using the parametric model for the link function we assume for the analysis. In case 2 we change these data such that the assumed link function will not provide a good fit anymore and we show the change to the resulting interval $[\underline{\gamma}, \bar{\gamma}]$. We assumed the temperature stress levels to be $K_0 = 283$, $K_1 = 313$ and $K_2 = 353$ Kelvin. We generated ten observations from a Weibull distribution at each stress level integrated with the Arrhenius link function. The Weibull distribution at K_0 had $\beta = 3$ and $\alpha_0 = 7000$, and the Arrhenius parameter γ was set at 5200. This model keeps the same shape parameter at each temperature, and the scale parameters are linked by the Arrhenius relation, which led to $\alpha_1 = 1202.942$ at K_1 and $\alpha_2 = 183.0914$ at K_2 . Ten units were tested at each temperature, so a total of 30 units is used in the study. The failure times are given in Table 1.

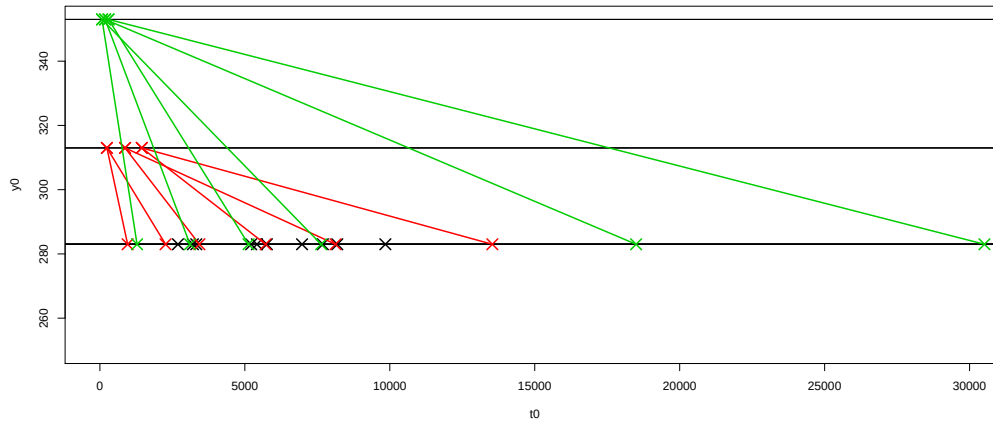


Figure1. Transformed data using [4060.018, 6605.752]

To illustrate the pairwise likelihood ratio tests method using these data, we assume the Weibul distribution at each stress level and the Arrhenius link function for the data. To obtain the accepted intervals $[\underline{\gamma}_i, \bar{\gamma}_i]$ of the values γ_i for which we do not reject the null hypothesis with regard to the well-mixed data transformation, we used the pairwise likelihood ratio test between K_1 and K_0 and between K_2 and K_0 . The results of the accepted intervals $[\underline{\gamma}_i, \bar{\gamma}_i]$ for three significance levels are giving in Table 2.

Then, we transformed the data using the $[\underline{\gamma}, \bar{\gamma}]$ values. All observations at the increased stress levels were transformed to the normal stress level. Therefore, the observations at the increased stress levels K_1 and K_2 are transformed to interval-valued observations at the normal stress level K_0 . We briefly illustrate this using only a small part of the data in Figure 1. We transformed the data from the higher stress levels K_1 and K_2 using [4060.018, 6605.752] with 0.99 level of significance, mixed with the original samples at the normal stress level K_0 . Note that, in Figure 1 the two largest transformed are actually the $\underline{\gamma}$ and $\bar{\gamma}$ transformations of the largest observation from level K_2 . So, this illustrates the key property of our method, that data transformed from higher levels tend to be wider intervals at the normal level.

The NPI lower survival function can be obtained by taking from the pairwise stress level K_1 to K_0 or K_2 to K_0 always the min of the $\underline{\gamma}_i$ with significance levels 0.99, 0.95 and 0.90 values, giving in Table 2. Similarly, the NPI upper survival function can be obtained by taking from the pairwise stress level K_1 to K_0 or K_2 to K_0 always the max of the $\bar{\gamma}_i$ with levels of significance 0.99, 0.95 and 0.90 values, given in Table 2. In case 1, we take $\underline{\gamma} = 4060.018$ and the $\bar{\gamma} = 6605.752$, $\underline{\gamma} = 4424.881$ and $\bar{\gamma} = 6261.168$, and $\underline{\gamma} = 4593.700$ and $\bar{\gamma} = 6100.653$ of the pairwise K_1, K_0 with 0.99, 0.95 and 0.90 significance levels, respectively. We used all the above $\underline{\gamma}$ and $\bar{\gamma}$ values to transform the data to the normal stress level 0, see Figure 2(a).

In case 2, we multiply the data at level K_1 by 1.4 and in addition we multiply the data at level K_2 by 0.8. The simulated data values are given in the last two columns in Table 1. We take $\underline{\gamma} = 3066.539$ and $\bar{\gamma} = 5920.775$, $\underline{\gamma} = 3431.402$ and $\bar{\gamma} = 5753.363$, and $\underline{\gamma} = 3600.221$ and $\bar{\gamma} = 5675.492$ for 0.99, 0.95 and 0.90 significance levels, respectively. We used all the above $\underline{\gamma}$ and $\bar{\gamma}$ values to transform the data to the normal stress level 0, see Figure 2(b). In these figures, the lower survival function $\underline{S}$ is labeled as $\underline{S}(\underline{\gamma}_i)$ and the upper survival function $\bar{S}$ is labeled as $\bar{S}(\bar{\gamma}_i)$. These figures show that higher significance level leads to more imprecision for the NPI lower and upper survival functions. Note that in Case 2, the observations at level K_1 have increased, leading to smaller $\underline{\gamma}_1$ and $\bar{\gamma}_1$ values, and this leads to the lower and upper survival functions decreasing in comparison to Case 1. Also, the observations at K_2 stress level in case 2, have increased, resulting in larger values for $\underline{\gamma}_2$ and $\bar{\gamma}_2$, and this leads to the lower and upper survival functions increasing in comparison to Case 1. The combined effect, based on $\underline{\gamma}$ and $\bar{\gamma}$, is therefore less obvious.

4. Simulation studies

The predictive statistical method we propose based on imprecise probabilities for ALT using the pairwise likelihood ratio tests with the parametric Arrhenius-Weibull model, is intuitively attractive because it shows

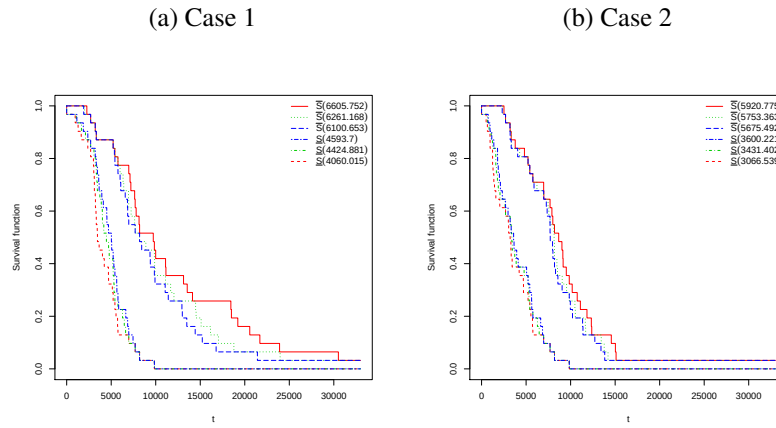


Figure2. The NPI lower and upper survival functions: Case 1 and Case 2.

some robustness, and demonstrates the effect of imprecision when we lack perfect information about the link between different stress levels. In this section, we apply simulation studies to investigate the performance of our proposed method when considering the assumed Arrhenius-Weibull model as the true underlying model.

To investigate the predictive performance inference of our new method for ALT data, we conduct a simulation study. We assumed the temperature stress levels to be $K_0 = 283$, $K_1 = 313$, and $K_2 = 353$. In this analysis, we run the simulation 10,000 times with the data simulated from the Weibull distribution and using the parameters shape parameter = 3, scale parameter = 7000, and the link function parameter $\gamma = 2000$, using the assumed Arrhenius link function model between different stress levels, where we have used $n = 10, 50, 100$ assigned to each stress level. We applied the method described in Section 2, with level of significance 0.95. For good performance of the method, we check the results by taking into account a future observation with actual data transformed to the normal stress level K_0 and how well the future observation mixes among them. We examined the performance in the quartiles of NPI lower and upper survival functions for $q = 0.25, 0.50, 0.75$, and whether or not the future observation at the normal stress level has exceeded these quartiles in the right proportion. One can use different quartiles but these quartiles provided a good indicator of the performance of our method. We conducted simulation with assumed Arrhenius-Weibull model, hence with data generated from the same model.

For good performance of the method, we require that the future observation for each run at the normal stress level has exceeded the first, second, and third quartiles of the NPI lower survival functions just over proportions 0.75, 0.50 and 0.25 of all runs, respectively, and for the NPI upper survival functions just under proportions 0.75, 0.50 and 0.25. Figure 3 presents the results of the performance of our method when the Arrhenius model is the assumed model for the simulated data and for the analysis. All of these figures show an overall good performance of the proposed method, with level of significance 0.95. Note that the first, second, and third quartiles in these figures are denoted $qL0.25$ and $qU0.25$, $qL0.50$ and $qU0.50$, and $qL0.75$ and $qU0.75$ corresponding to the NPI lower and upper survival functions, respectively. We note that for corresponding proportion with larger values of n , the differences between lower and upper survival functions tend to decrease. That means that when based on more data, the NPI lower and upper survival functions allow more precise inference. From these simulations, we conclude that the proposed approach using the Arrhenius-Weibull model provides suitable predictive inference if the model assumptions are fully correct. We will investigate robustness in case of model misspecification in simulation in the future.

5. Concluding remarks

This paper has presented statistical methods for ALT using the Arrhenius-Weibull model under constant stress testing, supported by the theory of imprecise probability, where the imprecision results from the use of the likelihood ratio test. The proposed method applies the use of the likelihood ratio test to compare the survival distribution of pairwise stress levels, in combination with the Arrhenius model finding the accepted interval of γ values according to the null hypothesis. The observations at the increased stress levels were transformed to interval-valued observations at the normal stress level by developing imprecision in the link function of

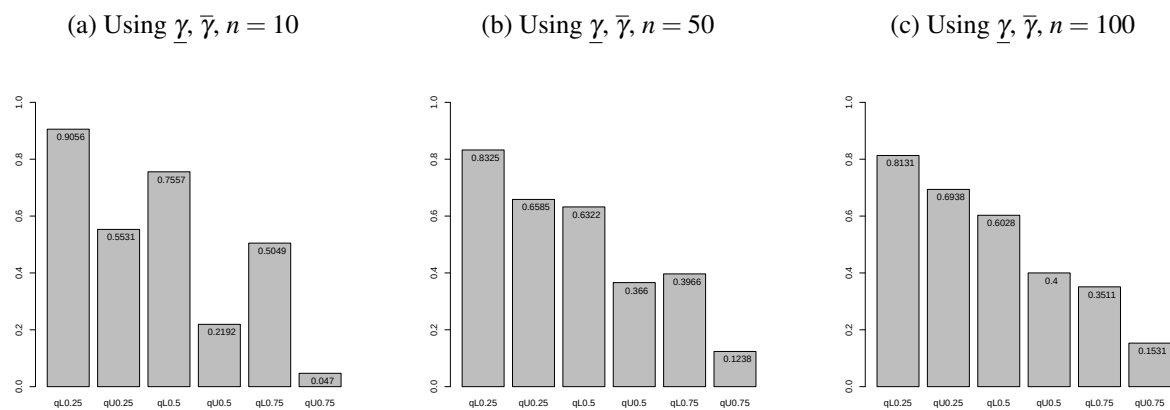


Figure3. Proportion of runs with future observation greater than the quartiles.

the Arrhenius model via a classical test between the pairwise stress levels. We found that using an interval of values for the parameter in the link function between different stress levels enabled us to achieve a greater level of robustness than if we were to use a single point for the parameter. Using the Arrhenius model, we linked the data at different stress levels to the normal stress level, after which NPI can be used at normal stress level.

The use of pairwise tests instead of one test on all stress levels simultaneously, is discussed in detail by [Ahmadini and Coolen 2018], who do not assume a Weibull distribution at each stress level and use the non-parametric log-rank test instead of the likelihood ratio test. The argument for use of the pairwise test is the same: if the model fits poorly, a single test on all stress levels would result in less imprecision while our proposed method, combining pairwise tests tends to, result in more imprecision. The application of our novel method in this paper assumed that we have failure times observed at the normal stress level K_0 . The assumption of having failure data at the normal stress level K_0 may not be realistic in real world applications. In this case, we can apply our method with basic function to a higher stress level that is above the normal stress level K_0 . The combined data at that level are then transformed all together to the normal stress level K_0 . Investigating this is a topic for future research. Also the presented method can be applied similarly if the data contain some right-censored observations, this will also be considered in detail in future research.

Acknowledgements

Abdullah A.H. Ahmadini gratefully acknowledges the financial support from Jazan University in Saudi Arabia and the Saudi Arabian Cultural Bureau (SACB) in London for pursuing his PhD at Durham University.

References

- Ahmadini, A., Coolen, F. (2018). Imprecise statistical method for accelerated life testing data using the Arrhenius model: imprecision related to the log-rank test. In submission.
- Alam, K., Sudhir, P. (2015). Testing equality of scale parameters of two Weibull distributions in the presence of unequal shape parameters. *Acta et Commentationes Universitatis Tartuensis de Mathematica*, **19**, 11–26.
- Augustin, T., Coolen, F. (2004). Nonparametric predictive inference and interval probability. *Journal of Statistical Planning and Inference*, **124**, 251–272.
- Augustin, T., Coolen, F., de Cooman, G., Troffaes, M. (2014). *Introduction to Imprecise Probabilities*. Wiley.
- Coolen, F. (2006). On nonparametric predictive inference and objective Bayesianism. *Journal of Logic, Language and Information*, **15**, 21–47.
- Coolen, F. (2011). Nonparametric predictive inference. *International Encyclopedia of Statistical Science*, pp. 968–970. Berlin: Springer.
- Hill, B. (1968). Posterior distribution of percentiles: Bayes' theorem for sampling from a population. *Journal of the American Statistical Association*, **63**, 677–691.
- Nelson, W. (1990). *Accelerated Testing: Statistical Models, Test Plans, and Data Analysis*. Wiley.

New method for availability computing of complex systems using Imprecise Markov models

J. Akrouche*, M. Sallak*, E. Châtelet**, F. Abdallah*** and H. Hajj Chehade***

*Sorbonne universités, Université de technologie de Compiègne, CNRS, Heudiasyc, CS 60 319, 60 203 Compiègne cedex, France. (joanna.akrouche, mohamed.sallak)@hds.utc.fr

** Université de Champagne, Université de technologie de Troyes, CNRS, Institut Charles Delaunay, 12 rue Marie Curie, CS 42060, 10004 Troyes cedex, France. eric.chatelet@utt.fr

*** Lebanese University, Beirut-Hadath, Lebanon. (fahed.abdallah76, hiba.hajchhade)@gmail.com

Abstract. The system availability is defined as the probability that the system is operational at a given time. To compute the availability of a multi-states system (the system and its components could have multi-states), we have to sum over all the probabilities of the components working states, therefore these probability precise values are required. In some cases (rare event failures, new components, ...), it isn't possible to obtain the working probabilities precisely because of the lack of data. In this work, we propose to apply imprecise continuous Markov chain where the failure and repair rates are imprecise. Only few works were developed using this concept. The precise initial distributions are replaced by intervals, which represents the unknown initial probabilities and the unknown transition matrix. The interval constraint propagation method is exploited for the first time, in availability modeling, to compute the imprecise multi-states system availability. The probability interval bounds associated to real variables are contracted, without removing any value that may be consistent with the set of constraints. The proposed methodology is guaranteed, and different examples of complex systems with some properties (convergence, ergodicity, ...) are studied. All the numerical examples and results will be discussed in the paper.

1. Introduction

When we study the dependability of a system, we are confronted in many cases with the fact that the system and its components may have different states or modes of functioning with different performance levels. Such systems are called multi-states systems (MMS). Applied to MSSs, availability analysis calculates the ability of the system to provide a required level of performance based on its level of degradation at a time t . Several methods have been proposed to calculate the availability of an MSS: Universal Generator Function (UGF) [1], inclusion-exclusion method [2], Monte Carlo simulations, and Markovian models[3]. The UGF is applicable only in the case of systems with simple structures: series, parallel, series-parallel, ... The inclusion-exclusion method is adopted only for systems where the components have binary states. Monte Carlo simulations are expensive in terms of time. In this paper, we have chosen to use Markovian model since it is a commonly used method in dependability domain for reliability and availability calculation of an MSS with repair rates [4].

On the other hand, in dependability, the knowledge that we have about the components data is generally uncertain. The validity of the results of the studies depends on the total or partial taking into account of the imperfection of the used knowledge. This requires methods that allow modeling and manipulation of these uncertainties such as probability theory, fuzzy set theory, belief functions theory, ... Among the works that have done to calculate the availability of an MSS, those based on the imprecise UGF method [5] using the arithmetic interval probability approach applied on the UGF, and the work proposed by Destercke et al. [6] proposing the use of the belief functions applied on the UGF method. In this work, we propose the use of imprecise Markov models for the calculation of the availability of an MMS. In a previous paper [7], we have some preliminary work about this subject.

2. Problem statement

The main issue in our work is to calculate the availability of complex MSS in presence of different types of uncertainties. Indeed, we are interested in uncertainties about reliability data (failure rate λ and repair rate μ of each component), due to the difficulty to estimate these data (new components, rare components failures, expensive components, ...) and the transitions rates could be variable over time and affected by several factors. The proposed way to cope with these uncertainties is to use the theory of imprecise probabilities, and particularly interval probabilities. We consider our transition rates as not being precise, but instead being bounded by intervals. That's why we propose to treat the data that we have in the terms of intervals and to apply interval analysis on imprecise Markov approaches, so we could calculate the imprecise availability of a system.

3. Contribution

In our study, we consider a complex MSS where the components and the system are multi-states and the number of components is important. The repair rates and the failure rates are represented in terms of intervals. To calculate the imprecise availability, we chose to use Markovian approaches.

3.1 Imprecise Markovian approaches

Markovian models represent a class of stochastic processes where system is "memoryless" [8], i.e., the future state of the system is independent of the previous states. The time in our case is continuous. The transition rates q_{ij} are in terms of failure rates and repair rates. Q the transition matrix is the matrix containing these elements. Note that the sum of the elements of each line of Q is equal to 0. Since we focus on the imprecision of failure and repair rates, we will take each element q_{ij} as an interval $q_{ij} = [q_{ij} ; \overline{q_{ij}}]$, where q_{ij} and $\overline{q_{ij}}$ represent respectively the lower and upper bounds of the interval the interval. Thus, the transition matrix Q will be in the form of an interval $Q = [\underline{Q} ; \overline{Q}]$.

3.2 Evaluation of the Availability

Based on the Chapman-Kolmogorov equations, we get the following formula

$$\dot{P}(t) = P(t) \cdot Q \quad (1)$$

Where $P(t)$ is the probability states vector of the system at a time t . We are interested in the stationary states probabilities that's why we need to compute the asymptotic availability. Using eq.(1), we obtain the following asymptotic solution

$$\Pi \cdot Q = 0 \quad (2)$$

where $\Pi = [\pi_i]$ is the steady probability vector of the system to be in a state i ; $\lim_{t \rightarrow \infty} P_i(t) = \pi_i$ and we always have the property $\sum_{i=1}^n \pi_i = 1$. Eq. (2) is a system of equations that we need to solve in order to find the intervals $\pi_i = [\underline{\pi_i} ; \overline{\pi_i}]$. Then, the asymptotic availability will be computed as the sum of the probabilities π_i corresponding to the working states. Solving the system of these equations isn't simple, therefore we introduce a method inspired from interval analysis : the technique of contraction [9].

3.2 The technique of contraction

The contraction technique [9], was introduced for the first time in 1970 in the artificial intelligence field. It is a method of constraint satisfaction problem and was later developed in interval analysis methods. This technique makes it possible to contract an interval $[x]$ by operators named contractors in order to obtain an interval $[x']$, such that $S \subset [x'] \subset [x]$ where S is the exact interval (note that this technique is more detailed in the appendix). We will use the contractor Forward-Backward propagation [9] for several reasons. It is applicable for all kinds of equations, besides calculations are much simpler than

with other contractors, the obtained intervals after the contraction are guaranteed that they belong to the initials intervals. The Forward-Backward propagation technique consists of two steps, the first is the "Forward" : it proposes to make an arithmetic calculation of the intervals. The second step is the "Backward" : it proposes to perform the same arithmetic calculation already made but in the opposite order, which means from the last operation made to the first one, taking into account in each computation step to take the intersections of $[x']$ and $[x]$. In our case the intervals $[x]$ are the $[\pi_i]$ that we need to contract them to obtain the smallest intervals, so we could calculate the availability of the system.

4. Methodology of the proposed method

To calculate the availability of a complex MSS with multi-state components, we first need to illustrate its Markov graph. So, we could present all the transitions from one state to another, from the total working state to the total failure state of the system. We define the interval transition matrix in terms of failure rates and repair rates of dimension $(n \times n)$.

To determine the availability of the system, first we take the initial vector Π as $\Pi = [[0,1][0,1] \dots [0,1]]$, and we apply eq. (2) where we obtain a system of n equations with the last equation representing the fact that $\sum_{i=1}^n \pi_i = 1$. We apply the Forward-Backward on the intervals $\pi_i = [\pi_i; \bar{\pi}_i]$, to reduce Π as much as it is possible. We keep repeating the contraction on all the equations and several times, until the vector Π converges. When we obtain Π , we calculate the imprecise availability of the system A by computing the sum of π_i over all the working states.

To verify if the results obtained by the technique of contraction is accurate, we will compare it to the availability obtained by taking all the possible transition matrix by doing a combination between the lower bounds and upper bounds of the failure rates interval and the repair rates interval, so for each combination we obtain a transition matrix, where we solve the system from eq.(2) and find the corresponding Π and the availability, after that we compare all the obtained availabilities from all combinations, so we can choose the corresponding availability interval where its bounds are the lowest availability (lower bound) and the highest availability (upper bound). We will call this method the "Exact method". In addition, we can verify the results, by taking the center of each interval of our data, so we will get one transition matrix formed from all the midpoint values of the data, we solve eq(2) to get the precise vector Π , where we could finally calculate the precise availability, which it has to belong to th interval obtained by the contraction technique and to the interval obtained by the exact method. We will call this method the "Precise method". Figure1 illustrate the steps of the technique of contraction.

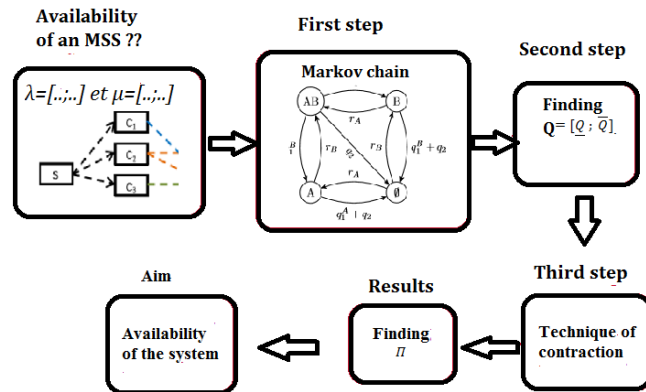


Figure 1. Methodology of the technique of contraction

5. Numerical example

To illustrate the technique of contraction with the Forward-Backward propagation, we will apply it on the following example. We aim to compute the availability of a system presented in Figure 2, and composed of eight components (A, B, C, D, E, F, G and H).

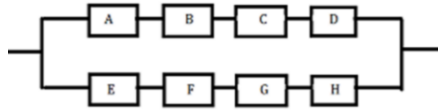


Figure 2. MSS composed of eight components.

The transmission of the system is from left to right. We assume that components B, C, E and H are binaries, i.e., they have only two possible states: a state of total failure, and a state of full functioning. Components A, F and G have three possible states: a state of total failure, a state of full functioning, and a state of partial failure. Component D has four possible degraded states: a state of total failure, a state of full functioning, a state of degraded functioning of type 1, and a state of degraded functioning of type 2. Thus, the total number of states for the system is 1728. All the information about repair rates and failure rates of each components are presented in Table 1.

Component A	$\lambda_{1,1} = [3 \times 10^{-5}; 4.5 \times 10^{-5}]$ $\lambda_{1,2} = [2.6 \times 10^{-5}; 7 \times 10^{-5}]$	$\mu_{1,1} = [4.1 \times 10^{-1}; 7 \times 10^{-1}]$ $\mu_{1,2} = [3 \times 10^{-1}; 6.2 \times 10^{-1}]$
Component B	$\lambda_2 = [2.6 \times 10^{-5}; 3.1 \times 10^{-5}]$	$\mu_2 = [1.6 \times 10^{-1}; 4 \times 10^{-1}]$
Component C	$\lambda_3 = [3.5 \times 10^{-5}; 4.6 \times 10^{-5}]$	$\mu_3 = [2.1 \times 10^{-1}; 3 \times 10^{-1}]$
Component D	$\lambda_{4,1} = [2.6 \times 10^{-5}; 3.5 \times 10^{-5}]$ $\lambda_{4,2} = [2.9 \times 10^{-5}; 3.7 \times 10^{-5}]$ $\lambda_{4,3} = [2.7 \times 10^{-5}; 3.4 \times 10^{-5}]$	$\mu_{4,1} = [3.1 \times 10^{-1}; 3.5 \times 10^{-1}]$ $\mu_{4,2} = [2.9 \times 10^{-1}; 3.3 \times 10^{-1}]$ $\mu_{4,3} = [3 \times 10^{-1}; 3.5 \times 10^{-1}]$
Component E	$\lambda_5 = [3 \times 10^{-5}; 4.1 \times 10^{-5}]$	$\mu_5 = [2.5 \times 10^{-1}; 3 \times 10^{-1}]$
Component F	$\lambda_{6,1} = [3.1 \times 10^{-5}; 5.6 \times 10^{-5}]$ $\lambda_{6,2} = [2.9 \times 10^{-5}; 4.6 \times 10^{-5}]$	$\mu_{6,1} = [2 \times 10^{-1}; 2.4 \times 10^{-1}]$ $\mu_{6,2} = [2.6 \times 10^{-1}; 3 \times 10^{-1}]$
Component G	$\lambda_{7,1} = [4.1 \times 10^{-5}; 4.7 \times 10^{-5}]$ $\lambda_{7,2} = [2.9 \times 10^{-5}; 3.2 \times 10^{-5}]$	$\mu_{7,1} = [2.3 \times 10^{-1}; 2.9 \times 10^{-1}]$ $\mu_{7,2} = [2.7 \times 10^{-1}; 3.1 \times 10^{-1}]$
Component H	$\lambda_8 = [4.1 \times 10^{-5}; 5 \times 10^{-5}]$	$\mu_8 = [2.7 \times 10^{-1}; 3 \times 10^{-1}]$

Table 1. Failure rates and repair rates form a state to another for each components

Where $\lambda_{k,i}$ and $\mu_{k,i}$ represent the failure rate i and repair rate i of component k .

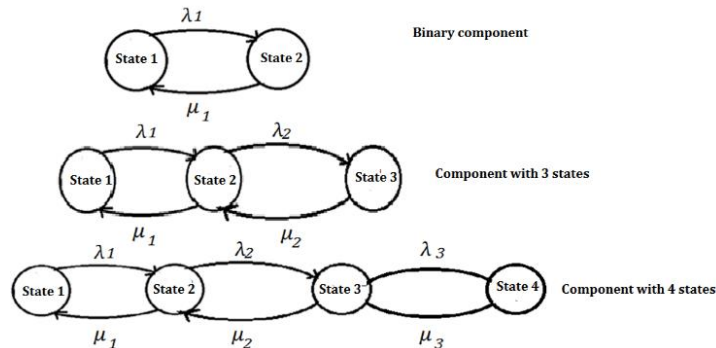


Figure 3. The corresponding Markov graph for each component

Since in this example we consider that the components are independents, so for each component, we will construct its Markov graph (cf. Figure 3), to determine its interval transition matrix. Then, we will

apply eq.(2) to find Π_k for each component, with $k = \{A, \dots, H\}$. Finally, we compute the availability of each component by computing the sum of $\Pi_{k,i}$ over all the working states. In this example, the availability of the entire system A_{system} is presented in the following equation:

$$A_{system} = 1 - (1 - A_A A_B A_C A_D) (1 - A_E A_F A_G A_H) \quad (3)$$

To find Π_k for each component we will apply the technique of contraction. We propose to compare the obtained availability with the availability of the system obtained by the exact method and by the precise one. All the results are presented in Table 2.

Contraction technique	Exact method	Precise method
[0.9999991; 0.99999999]	[0.9999996; 0.99999995]	0.99999992

Table 2. The availability of the system by applying different methods.

6. Conclusion

The main contribution in this paper is to compute the availability of a complex MSS. We have proposed the use of imprecise Markovian approaches in presence of imprecise failure and repair rates. Using the Contraction methods, and particularly the forward-backward contraction method on the equations obtained by the transition matrix of a Markov chain representing the MSSs. The main goal of the Contraction method is to reduce the initial interval to its most possible minimum size. After testing the method on several examples, one of them the one presented in this paper, and comparing the results of the obtained availability by the exact method and the precise one, we found that the interval obtained by the technique of contraction is more conservative than the interval obtained by the exact method. We have also find that precise availability is always belong to the intervals. Based on this method, we can study the availability of larger systems, where the number of components and the number of possible states for each element is higher, and then we can proceed to our future objective: the optimization of the imprecise availability of systems when considering imprecise failure and repair rates.

Appendix

Consider n_x variables $x_i \in \mathbb{R}$, $i = 1, \dots, n_x$, linked by n_f relations (or constraints) [9] of the form: $f_j(x_1, \dots, x_{n_x}) = 0$, $j = 1, \dots, n_f$, it can be written in vector notation as: $f(x) = 0$. Each variable x_i is known to belong to a domain X_i . For simplicity, these domains will be intervals, denoted by $[x_i]$. Define the vector $\mathbf{x} = (x_1, \dots, x_{n_x})^T$ and the prior domain for $\mathbf{x}$ is a box as: $[\mathbf{x}] = [x_1] \times \dots \times [x_{n_x}]$. Let $\mathbf{f}$ be the function whose coordinate functions are the f_j 's. Equation $f(x) = 0$ corresponds to a constraint satisfaction problem (CSP) H , which can be formulated as: $(H: f(x) = 0, \mathbf{x} \in [\mathbf{x}])$. The solution set of H is defined as: $S = \{\mathbf{x} \in [\mathbf{x}] | f(\mathbf{x}) = 0\}$. Contracting H means replacing $[\mathbf{x}]$ by a smaller domain $[\mathbf{x}']$ such that the solution set remains unchanged, i.e. $S \subset [\mathbf{x}'] \subset [\mathbf{x}]$. There exists an optimal contraction of H , which corresponds to replacing $[\mathbf{x}]$ by the smallest box that contains S . A contractor for H is any operator that can be used to contract it. A contractor C is defined as an operator used to contract the initial domain of the CSP, and thus to provide a new box. Several contractors exist, each works in a different manner and is efficient only for specific CSPs and for certain cases [9]. The most popular contractor, which will be used in our approach, is the "Forward-backward propagation (FBP) contractor". This technique is known for its simplicity and easiness, it is also more general than the others since it works on all type of systems [9]. It also gives guaranteed results which means that during the contraction we always get an interval belonging to the initial interval. The Forward-backward (FBP) contractor $C \downarrow \uparrow$, is a classical algorithm in constraint programming for contracting. This contractor makes it possible to contract the domains of the CSP H by taking into account each one of the n_f constraints apart. The algorithm works in two steps [9]. The forward step applies interval arithmetic to each operator of the function $y=f(x)$, from the variable's domain $([x])$ up to the function's domain $([y])$, this step considers the direct forms of the

equations. The backward step sets the interval associated to the new function's domain $[y]$ to $[0, 0]$ (imposes constraint satisfaction, since we are solving $f(x)=0$) and, then, applies backward arithmetic from the function's domain to the variable's domain, which means using the inverse of the functions that appear in the equations $f(x)$. The following example explains the procedure of the FBP technique. *Example:* Consider the constraint $y = -5x_1 + 2x_2 = 0$ and the initial box-domain $[x] = [1, 4] \times [-3, 7]$. This constraint can be decomposed as shown in eq.(4) into three primitive constraints (i.e. constraints associated with a unique elementary function: multiplication or addition) by introducing two intermediate variables a_1 and a_2 defined as: $a_1 = -5x_1$ and $a_2 = 2x_2$. Initial domains for these variables are determined as follows:

$$\begin{aligned} a_1 &= -5x_1 = -5 \times [1, 4] = [-20, -5] \\ a_2 &= 2x_2 = 2 \times [-3, 7] = [-6, 14] \\ y &= a_1 + a_2 = [-20, 5] + [-6, 14] = [-26, 9] \end{aligned} \quad (4)$$

and this step is called the "forward propagation". A method for contracting H with respect to the constraint $f(x) = -5x_1 + 2x_2 = 0$ is to contract each of the primitive constraints in eq.(4) until the contractors become inefficient. For this example: Since $f(x) = 0$, the domain for y should be taken equal to $\{0\}$, so we can add the step: $[y] := [y] \cap \{0\}$. If $[y]$ as computed in eq.(4) turns out to be empty, then the CSP has no solution. Else, $[y]$ is replaced by 0 , which is the case in this example. After, a backward propagation is performed, updating the domains associated with all the variables to get:

$$\begin{aligned} [a_1] &:= ([y] - [a_2]) \cap [a_1] = [-14, -5] \\ [a_2] &:= ([y] - [a_1]) \cap [a_2] = [5, 14] \\ [x_1] &:= ([a_1]) / -5 \cap [x_1] = [1, 14/5] \\ [x_2] &:= ([a_2]) / 2 \cap [x_2] = [5/2, 7] \end{aligned} \quad (5)$$

Thus, we obtain the new box: $[x](1) = [1, 14/5] \times [5/2, 7]$, which is the result of the first FBP contraction. Iterating this procedure, the resulting sequence of boxes $[x](k)$ converges towards the smallest possible domain, after which the domains no longer change following another iteration of FBP.

References

- [1] Gregory Levitin (2005). *The universal generating function in reliability analysis and optimization*. Springer.
- [2] Marija Mihova and Natasha Maksimova (2011). Estimation of minimal path vectors of multi state two terminal networks with cycles control. *Mathematica Balkanica*, 25(4): 437–447
- [3] Matthias C.M. Troffaes, Jacob Gledhill, Damjan Skulj, Simon Blake (2015). Using Imprecise Continuous Time Markov Chains for Assessing the Reliability of Power Networks with Common Cause Failure and Non-Immediate Repair. 9th International Symposium on Imprecise Probability: Theories and Application, Pescara, Italy.
- [4] A. Lisnianski and G. Levitin (2003), *Multi-State System Reliability: Assessment, Optimization and Applications*. World Scientific Publishing Co Pte Ltd.
- [5] C.-Y. Li, X. Chen, X.-S. Yi, and J.-Y. Tao (2011), "Interval-valued reliability analysis of multi-state systems," *IEEE Transactions on Reliability*, 60, 323–330.
- [6] S. Destercke and M. Sallak (2013), An extension of Universal Generating Function in Multi-State Systems Considering Epistemic Uncertainties, *IEEE Transactions on Reliability*, 62(2), 504–514.
- [7] J. M. Akrouche, M. Sallak, E. Châtelet, F. A. Abdallah and H. Z. Haj Chhadé. A contribution to the evaluation of imprecise availability of complex systems using Markov models. UNCECOMP 2017, Greece 1517 June 2017.
- [8] Mazen El Falou and Eric Châtelet (2011), Analytical Uncertainty Propagation for Availability Assessment of Stochastic Multi-state Systems, *International Journal of Performability Engineering*, 7(3): 251–261.
- [9] Luc Jaulin, Michel Kieffer, Olivier Didrit, Eric Walter (2001). *Applied Interval Analysis*. Springer.

Modelling the quality of maintenance of a protection system: a case study on isolation valves used in water distribution systems.

A. R. Alberti*, C. A. V. Cavalcante**, P.A. Scarf ***, A. L. O. Silva****

* Departamento de Eng. de Produção, Universidade Federal de Pernambuco, Brazil: alexandre.emc091@gmail.com

** Departamento de Eng. de Produção, Universidade Federal de Pernambuco, Brazil: cristiano@ufpe.br

*** Salford Business School, University of Salford, Manchester M5 4WT, UK: p.a.scarf@salford.ac.uk

**** COMPESA, Companhia Pernambucana de Saneamento, Brazil

Abstract. A hybrid inspection and replacement policy for a protection system is described by a mathematical model that incorporates a set of maintenance quality aspects. A three-state component failure model is assumed, with a defective state preceding failure. The quality of maintenance intervention is modelled by supposing that inspection may induce a defect and is subject to errors in the defective state judgement (false positives and false negatives). The quality of replacement is modelled by supposing that a component arises from a heterogeneous population, composed of weak and strong items. This paper presents a case study on isolation valves used in water distribution systems, that motivated the development of the model. We evaluate the impact of these quality aspects upon cost and production losses. A detailing of the context is presented, highlighting important aspects of the analysis, which should be observed when other studies are carried out in different contexts. We present the results in order to evaluate the impact of the quality aspects regarded, verify conditions that justify investment in higher quality maintenance, and obtain guidance on which aspects should be prioritized in the investment. Defect induction is found to be a key determinant of the cost-optimal policy.

1. Introduction

This paper presents a discussion about a case study on isolation valves carried out at a Brazilian sanitation company. This case study motivated the development of a new model to evaluate the impact of different aspects of maintenance quality on the performance of protection systems, which was recently presented in Alberti et al. (2018). Here, our objective is present a discussion about the case study accomplishment, highlighting important aspects of the analysis, which should be observed when other studies are carried out in different contexts. Some results of an application are present in order to show how the model can be used to obtain some guidelines for decision-making in maintenance.

2. The case study

Isolation valves play an important role in a water distribution system by enabling the isolation of certain segments in the system, which is necessary in some cases such as repair of broken pipes or confinement of contaminants within a controlled area (Jun et al., 2007). Walski (1993) defined “segment” as the smallest area in a water distribution system for which supply can be interrupted with the closing of the valves without affecting the rest of the system.

Due to its operation characteristics, isolation valves can be classified as a protection system, that is, a system for which operation is only required in the case of a specific demand. In that case, demands can be understood as any event that requires the isolation of a segment in the distribution network. Two types of demands are specifically highlighted in the literature: leaks or related preventive measures, and water contamination (Jun et al., 2007; Giustolisi and Savic, 2010). Goulter and Coals (1986) suggested a Poisson distribution with parameter λ (failures/km/year) as a probability distribution that better represents the occurrences of pipeline breakdowns. In regards to water contamination events, a Poisson distribution is assumed to also represent them well since those are occasioned by external and potentially random factors.

When a demand event occurs and the valves required are fully functional, the isolation of the segment involved is done without delay, leaving only a limited number of clients without service and avoiding greater damage to the region. On the other hand, when a demand occurs and any of the isolation valves required fail,

a new containment plan is needed and requires the isolation of adjacent segments, leaving a greater number of clients without service for a longer period.

The damage caused by pipeline breakdowns can be significant depending on the operational pressure of the segment and the failure of an isolation valve can make matters worse due to the greater time needed to carry out containment plans, thus contributing to greater waste of water, more damage to surrounding structures potentially leading to erosion. It is even more critical to control contamination events, which have a greater impact on cost and service disruption. Failure of the isolation system can lead to contaminants being spread to more extensive areas, contributing to costlier economic damages, as in the case of substances that can damage the network and clients' infrastructure, or even magnify the effects of sabotage actions.

We can see that the impact resulting from an unmet demand can be considerably negative, which justifies the concerns with isolation valve maintenance planning. Next, some information obtained from an investigation about failure aspects and maintenance practices is presented.

2.1 Failure aspects

An isolation valve is considered to be failed when it is not possible to close it completely to stop water flow through it. As it is common with mechanical equipment, the valve's physical condition or "state" declines over time due to a number of factors such as fatigue, corrosion, wearing of material, or accumulation of debris. The deterioration rate depends on the valve and its specific operational context (Marlow, Beale and Mashfort, 2012). In our case study, we discovered that the most failure prone component is the reduction mechanism of the valve's activation system, its criticality is due to the fact that it wears off faster and is more sensitive to factors related to maintenance quality. Such components are used in valves of great nominal diameter and high operation pressure.

In practice, it is noted that before failure, the equipment presents a malfunctioning state, that is, even though it can fulfill its role, it shows signs that should serve as a warning of imminent failure such as, resistance when turning the valve's lever and/or small passing of water. Such valve functioning information can be gathered through inspections thus, the delay time modelling becomes very useful for describing failures in this context (Marlow, Beale and Mashfort, 2012; Christer, 1999).

2.2 Maintenance

Since the reduction mechanism that activates isolation valves is critical, the main focus of our analysis is the replacement of this component in special. Because it is a component with a high deterioration rate compared to others, and is greatly affected by maintenance quality aspects (especially when it comes to defect induction), it can be presumed that a maintenance policy defined based in its characteristics is capable of serving the system well as a whole. Thus, preventive and corrective maintenance has to entail the substitution of the valve's reduction mechanism, and other activities that are done with the purpose of improving the state of the equipment as a whole and restoring it to a like-new state.

Due to safety concerns, it is not possible to perform maintenance on the isolation valve while it is under normal operation conditions. Therefore, it is necessary to interrupt water flow by isolating the surrounding region. In that case, maintenance of the isolation valve system causes the main system to be partially unavailable, and it is plausible to assume that the system's activation will not be required during that time. That stands in contrast to the example presented by Berrade, Cavalcante and Scarf (2012), where protection system maintenance does not lead to the unavailability of the main system; however, there is the risk of a demand not being met by the protection system while in maintenance.

The quality of the maintenance service may influence the valve's performance: if maintenance is performed incorrectly, or with poor quality materials, the valve can become more vulnerable to premature failure with a short life span (this aspect is classified as bad installation).

2.3 Inspection

The only purpose of inspections is to identify the equipment's working status. The inspection procedure is done in the following manner at the company in which the case study was conducted: an inspector tries to close an isolation valve and verifies, through pressure gauges downstream, if the water flow was interrupted.

In this case, the inspection does not cause the protection system to be unavailable and the impact on the operation of the main system is negligible.

Some important observations were made:

- (1) The pressure gauges show pressure differences when water flow is high enough to fill the duct's transverse section, which happens when the valve is failed. Small quantities of passing water are not detected, making it difficult to identify a defective state.
- (2) The perception of the resistance when closing a valve (a characteristic of defective valves) varies among inspectors, depending on personal physical strength. Nonetheless, it is possible to tell for sure when a valve cannot be fully closed (failed valve).
- (3) The use of tools such as levers to close the valves more easily is common practice. That practice increases the torque applied, but the inspector may end up using more force than necessary and damage the components of the activation mechanism of the valve, accelerating wear and tear of the component.

By making these observations, we concluded that the quality of the inspection influences the accuracy of the judgement of the valve's functioning status, besides potentially contributing to valve malfunction.

As seen in items (1) and (2), it can be concluded that there is no error in judgement when it comes to identifying signs of failed state; however, the inspection is subject to error in judgement when identifying the defective state. Besides that, as seen in item (3), a poor quality inspection can lead to defects.

2.4 The model

Based on these observations, we developed a delay time model for a single component protection system, which regards the maintenance quality aspects aforementioned:

- In order to model the replacement quality, we assumed that the components come from a heterogeneous population made out of strong and weak items, with a mix parameter p indicating the percentage of weak items (Scarf et al., 2009).
- To model inspection quality, we considered the probabilities w of a false positive (false alarm when the valve is fully functional) and q of a false negative (as approached by Berrade, Cavalcante and Scarf, 2012). Therefore, we assume that there is a probability r of inducing defect in a good working valve (as approached by Scarf and Cavalcante, 2012).

We considered a hybrid inspection and replacement policy similar to the one suggested by Vaurio (1999), which advises performing M inspections with time interval T between them such that, when an inspection finds a component in defective or failed state, its replacement for a new unit is arranged and in the occasion of the M -th inspection, replacement is indicated regardless of its state.

The model was presented in Alberti et al. (2018).

3. Application

For the application, we consider as decision criterion the cost-rate (C_∞) obtained as function of the maintenance policy, additionally we present as an auxiliary criterion the loss-rate (Sh_∞ - long-run production loss per unit time), which here represents the quantify of water that is not delivered.

3.1 Specification of the parameters

To apply the model, we studied an isolation valve in a water distribution network that is installed in a high-pressure pipeline, located in an easily accessible site. Also, the segments adjacent to that valve cover a region that is at high risk for contamination. The parameters were estimated based on expert knowledge obtained from the engineers of the network operator and maintainer.

There are premature failures in approximately 10% of the valves installed due to poor installation, or items that return to the company's inventory after a poor maintenance. Therefore, we established $p=0.1$. Weibull distributions were used to distinguish the X (time to defect arrival) and H (delay time) distributions (characteristics of gear reducers), with the parameters displayed in Table 1.

In the initial scenario observed in the case study, we found a low probability of a false positive occurring followed by a high probability of a false negative occurring, while the probability of a defect induction was

low, albeit not significant. In that way, we estimated the following values for the variables that characterize the quality of the maintenance: $w = 0.05$, $q = 0.3$ and $r = 0.05$.

Table 1 – Probability distribution parameters of X and H.

Variable	Scale parameter	Shape parameter
X – weak components	$\eta 1 = 1$ year	$\beta 1 = 2$
X – strong components	$\eta 2 = 4$ years	$\beta 2 = 3$
H	$\eta 3 = 0.5$ year	$\beta 3 = 1$

An isolation valve is subject to demands coming from the different segments adjacent to it: when there is a need to isolate a segment, the adjacent segment can also be affected depending on the state of the valve. Each segment has their specific rates of demand occurrences and the consequences associated with it are also different in each case. For a direct application of the model presented, we simplified the scenario by regarding the valve as a protection system subject to two kinds of demands and by assuming that the consequences of each demand, depending of its functional state, are constant and known. The value used to characterize the consequences of the demands, met or not, may be an average of what is actually seen.

Therefore, concerning operations, we consider that demands due to leaks occur on average once every 120 days ($\mu_1 = 3.042$ demands/year), while demands resulting from contaminations occur on average once every two years ($\mu_2 = 0.5$ demands/years). If a demand due to leakage occurs and the valve is functional, a region with an average daily consumption of 25.000 m³ of water (D_1) is without supply for 1 day ($T1_m$), but if the valve fails, a greater region with daily consumption of 50.000 m³ (D_2) is without supply for 2 days ($T1_u$). On the other hand, if a demand due to contamination occurs and the valve is fully functional the time without supply is 2 days ($T2_m$), and if the valve fails, 4 days ($T2_u$). To maintain the valve, it is necessary to interrupt the supply to a region with a daily consumption of 50.000 m³ (D_3) for 1 day ($S_p = S_d = S_f$). Besides that, as observed in subsection 3.2: $S_i = 0$.

The costs required for the model were established based on the cost of preventive replacement, so that $c_p = 1$ un. (quantity not specified in monetary units). The remaining costs were then specified in the following way: $c_i = 0.04$ un., inspection cost is significantly lower than preventive replacement, but it is not insignificant due to the cost of transportation to the valve's location. With regards to replacement costs: $c_d = 1,4$ un. and $c_f = 2$ un., it is plausible to consider these as being different because we are considering the cost of refurbishing the component to return it to the company's inventory, which varies according to its state ($c_f > c_d > c_p$). The extra costs resulting from a unmet leakage demand is $CUD_1 = 20$ un., and the extra costs resulting from a unmet contamination demand is $CUD_2 = 50$ un., these values take into consideration the direct costs of the containment and the impact in terms of clients left without water supply.

3.2 Results and discussion

The model is an approximation to reality, and to obtain the most suitable values for the maintenance quality parameters can be a difficult activity. Thus, a complete analysis of results obtained from the model has to be carried out, in order to obtain some guidelines for decision-making in maintenance. In this section, we present a possible way to realize this analysis.

First, we aimed to analyze the effect of the variation of the maintenance quality parameters on the definition and performance of the optimal maintenance policy. Beside this, we aimed to evaluate if worth to adopt a policy that advises the replacement in case of identification of a defective state. Some results are presented in Table 2.

From the results, we can conclude that, given the level of inspection quality, this practice is not advantageous, so it is more advisable to adopt a non-inspection preventive replacement policy, even if quite conservative, with little use of the useful component's life. The results were more sensitive to variation of r , what evidence that, if there is some uncertainty about the value of this parameter, it can be better to consider the scenario more conservative. Another trend observed is that the practice of inspections becomes more relevant the greater the percentage of weak items in the population of components (p).

In the lines 13 and 14 we consider a policy where the replacement is advised only when the failed state is identified (for this, it is sufficient to set $w = 0$ and $q = 1$, that is, not “perceiving” a defective state). From the results, we can see that it is better to consider the policy as presented previously. Furthermore, from the lines 15 and 16, it gets evident that mistakes can be made if the maintenance quality aspects considered are ignored.

Table 2 - Comparison of optimal policies for various values of the quality parameters. T in years, C_∞ in un., Sh_∞ in $10^5 m^3$ (adapted from Alberti et al., 2018).

Case	Quality parameters				Optimal policy				
	p	w	q	r	M	T	MT	C_∞^*	Sh_∞
1 Base case	0.10	0.05	0.30	0.05	1	0.785	0.785	1.856	1.659
2 Worse p	0.20	0.05	0.30	0.05	24	0.151	3.624	2.134	1.425
3 Best p	0	0.05	0.30	0.05	1	1.239	1.239	1.072	1.419
4 Worse r	0.10	0.05	0.30	0.10	1	0.785	0.785	1.856	1.659
5 Better r	0.10	0.05	0.30	0.03	20	0.162	3.240	1.644	1.353
6 Better r	0.10	0.05	0.30	0.01	15	0.184	2.760	1.253	1.314
7 Best r	0.10	0.05	0.30	0	17	0.178	3.026	1.051	1.288
8 Worse w	0.10	0.10	0.30	0.05	1	0.785	0.785	1.856	1.659
9 Best w	0.10	0	0.30	0.05	1	0.785	0.785	1.856	1.659
10 Worse q	0.10	0.05	0.50	0.05	1	0.785	0.785	1.856	1.659
11 Best q	0.10	0.05	0	0.05	20	0.178	3.560	1.805	1.371
12 Best w and q	0.10	0	0	0.05	25	0.146	3.650	1.656	1.295
13 Ignoring defects, better r	0.10	0	1	0.01	14	0.133	1.862	1.687	1.321
14 Ignoring defects, best r	0.10	0	1	0	20	0.109	2.180	1.309	1.264
15 Perfect inspection	0.10	0	0	0	20	0.164	3.280	0.818	1.197
16 Perfect maintenance	0	0	0	0	15	0.192	2.880	0.749	1.201

The model also can be used to demonstrate the benefit of investment in higher quality maintenance, as in the analysis presented in Table 3.

Table 3 - An analysis of the benefit of investment in higher quality maintenance. T in years, C_∞ in un., Sh_∞ in $10^5 m^3$ (adapted from Alberti et al., 2018).

Case	Quality parameters				Costs				Optimum policy				
	p	w	q	r	c_i	c_p	c_d	c_f	M	T	MT	C_∞	Sh_∞
1 Base case	0.10	0.05	0.30	0.05	0.04	1	1.40	2	1	0.785	0.785	1.856	1.659
2 Improve classification	0.10	0.01	0.10	0.05	0.05	1	1.40	2	24	0.151	3.624	1.816	1.314
3 Improve classification	0.10	0.01	0.10	0.05	0.06	1	1.40	2	21	0.161	3.381	1.879	1.312
4 Reduce defect induction	0.10	0.05	0.30	0.02	0.08	1	1.40	2	13	0.207	2.691	1.662	1.324
5 Reduce defect induction	0.10	0.05	0.30	0.03	0.08	1	1.40	2	13	0.203	2.639	1.864	1.344
6 Improve inspection	0.10	0.01	0.10	0.01	0.12	1	1.40	2	11	0.244	2.684	1.448	1.251
7 Improve inspection	0.10	0.01	0.10	0.02	0.12	1	1.40	2	12	0.229	2.748	1.645	1.263
8 Improve inspection	0.10	0.01	0.10	0.03	0.12	1	1.40	2	12	0.224	2.688	1.837	1.281
9 Reduce weak items	0.05	0.05	0.30	0.05	0.04	1.10	1.54	2.20	1	0.954	0.954	1.670	1.546
10 Reduce weak items	0.05	0.05	0.30	0.05	0.04	1.50	2.10	3	1	1.042	1.042	2.082	1.506
11 Joint improvement	0.05	0.01	0.10	0.01	0.12	1.50	2.10	3	11	0.255	2.805	1.648	1.237

The analysis shows that the investment in higher quality maintenance can present a good cost benefit in the long term. The best results are obtained with investment in inspection quality, especially when focused on reducing the level of defect induction.

4. Conclusion

In this paper we discuss an interesting, real problem that motivated the development of a new model to evaluate the impact of maintenance quality aspects on the performance of a protection system. We bring some details of the case study in order to present the elements considered to build the model.

Through the application, we show how the model can be used to obtain important guidelines for decision-making in maintenance. From the results, we are able to answer some questions: what maintenance quality aspects have the greatest impact on the system's performance? Is it worth to invest in higher quality

maintenance? Which aspects must be prioritized in that investment? The level of defect induction during inspections has the most significant impact on the performance of the maintenance policy, and its impact becomes even more prominent when combined with judgement errors made during inspections. By investing in higher quality maintenance, it is possible to arrive at a more economical scenario in the long term. That investment should be directed mainly towards improving inspection quality, especially when it comes to reducing defect induction. Furthermore, if efforts to reduce judgement errors are included, even greater results can be achieved.

In our case study, defect induction was a frequently mentioned topic by those interviewed; however, a clearly identified perception, and that can be translated into reliable estimates, cannot always be verified. Hence, the application of techniques to investigate trends that affect the system performance can be useful.

Acknowledgements

The work of Cristiano Cavalcante has been supported by CNPq (Brazilian Research Council). The work of Alexandre Alberti has been supported by FACEPE (Foundation for Science and Technology of the State of Pernambuco). The work of Phil Scarf has been supported by the College of Science, King Saud University, Saudi Arabia.

References

- Alberti AR, Cavalcante CAV, Scarf PA, Oliveira ALO (2018) Modelling inspection and replacement quality for a protection system. *Reliability Engineering & System Safety*, In Press, DOI: <https://doi.org/10.1016/j.ress.2018.04.002>.
- Berrade MD, Cavalcante CAV, Scarf PA (2012) Maintenance scheduling of a protection system subject to imperfect inspection and replacement. *European Journal of Operational Research* **218**, 716-725.
- Giustolisi O, Savic D (2010) Identification of segments and optimal isolation valve system design in water distribution networks. *Urban Water Journal* **7(1)**, 1-15.
- Goulter C, Coals AV (1986) Quantitative approaches to reliability assessment in pipe networks. *Journal of Transportation Engineering* **122(3)**, 287-301.
- Jun H, Loganathan GV, Deb AK, Grayman W, Snyder J (2007) Valve Distribution and Impact Analysis in Water Distribution Systems. *Journal of Environmental Engineering* **133**, 790-799.
- Marlow DR, Beale DJ, Mashfort JS (2012) Risk-based prioritization and its application to inspection of valves in the water sector. *Reliability Engineering & System Safety* **100**, 67-74.
- Scarf PA, Cavalcante CAV (2012) Modelling quality in replacement and inspection maintenance. *International Journal of Production Economics* **135**, 372-381.
- Scarf PA, Cavalcante CAV, Dwight RA, Gordon, P (2009) An Age-Based Inspection and Replacement Policy. *IEEE Transactions on Reliability* **58(4)**, 641-648.
- Vaurio JK (1999) Availability and cost functions for periodically inspected preventively maintained units. *Reliability Engineering & System Safety* **63**, 133-140,
- Walski TM (1993) Water distribution valve topology for reliability analysis. *Reliability Engineering & System Safety* **42**, 21-27.

Mitigating Warranty Risk for Automotive Industry

Ahmed Aljazea*, Ming Luo** and Shaomin Wu***

* Kent Business School, Canterbury, Kent CT2 7FS, UK. Email: ama65@kent.ac.uk

**Kent Business School, Canterbury, Kent CT2 7FS, UK. Email: ml472@kent.ac.uk

***Kent Business School, Canterbury, Kent CT2 7FS, UK. Email: s.m.wu@kent.ac.uk

Abstract Warranty is offered by manufacturers as protection and promotional tools. It also gives customers a certain degree of insurance against product failures for a certain period. Although bringing those benefits, it involves various risks originating from various perspectives of the product lifetime cycle. To prepare risk mitigation plans is therefore needed, which is challenging due to the increasing complexity of the product designs and the long warranty period. Accordingly, the decision made to select the mitigation plan involves a high level of uncertainty. This paper develops a plan to mitigate warranty risk based on cumulative prospect theory, which helps warranty decision makers in selecting the optimal mitigation plan.

1. Introduction

Although the literature of warranty management is vast, it focuses on making financial plans such as estimating the number of the future warranty claims or costs, optimising maintenance policies, among others. In contrast, warranty risk management (WaRM) has received little attention and only been mentioned as an off topic by Díaz & Márquez (2011), González-Prida & Márquez (2012) and Costantino et al. (2012).

Mitigating warranty risk as a decision problem is another challenge that has not been investigated in the literature. The risk mitigation process requires decision makers to opt the optimal mitigation plan fulfilled the optimal cost and result based on predefined criteria. In the warranty context, the decision made to select such a plan may involve a high level of uncertainty and risk due to the increasing complexity of the new invitations and long warranty period.

Different multi-criteria decision-making tools (MCDM) such as multi-attribute utility theory (MAUT) and analytic hierarchy process (AHP) can be used to select the optimal warranty risk mitigation plan (Velasquez & Hester, 2013). However, the existing methods are unable to capture some behavioural aspects of the decision makers towards making the decision under uncertainty and risk, which can influence the final decision.

There are some behavioural decision-making methods such as prospect theory (Kahneman, 1979), regret theory (Bell, 1982), disappointment theory (Bell, 1985), third generation prospect theory (Schmidt et al., 2008) and cumulative prospect theory (Tversky & Kahneman, 1992) were therefore developed. The cumulative prospect theory (CPT), however, shows its superiority through well describing the decision makers' attitude in four elements: (1) reference dependence; (2) loss aversion; (3) diminishing sensitivity, and (4) probability weighting, and providing the formulas required to calculate values and weights in a clear logic and simple computation process. For this reason, CPT is adapted in this paper to address the decision problem in mitigating warranty risk. Interested readers can refer to (Tversky & Kahneman, 1992) for more details.

In warranty management practice, some automotive manufacturers may invalidate warranty if the customer fails to comply with the recommended service schedule or if the vehicle continues to be driven when the fault is apparent¹. It implies that the periodic check-ups are necessary to detect failures at the early stage; after a certain period or after a certain amount of failures reported, WDMs may decide to implement a mitigation plan selected from a pre-specified protocol in response to the occurred warranty claims. The chosen plan at time t , however, can affect the future decisions concerning the selection of

¹ For example, BMW (2018): <https://www.bmw.co.uk/bmw-ownership/servicing-and-repairs/bmw-warranties/new-car-warranty-guidelines>

the optimal mitigation plan at the future check-ups. Accordingly, the reference point of the mitigation plans and the WDMs' behaviour may be changed. For example, the loss aversion can be greater at t_{i+1} than t_{i-1} due to some losses experienced in the recent past. This paper models the decision process of risk mitigation with the CPT considering time-varying variables such as the reference point of the costs of warranty mitigation plans and the parameters that capture WDMs' behaviour).

2. Introduction to CPT, Assumptions and notations

Cumulative prospect theory (CPT) is a decision making model for descriptive decisions under risk and is introduced by Tversky & Kahneman (1992).

We make the following assumptions in this paper.

- (1) The warranty decision makers (WDM) need to select different mitigation plans to treat warranty related hazards at different time points based on two criteria (a) warranty cost, and (b) manufacturer reputation.
- (2) The time to an implement mitigation plan is pre-scheduled.
- (3) Three levels of mitigation plans are considered: (a) product recall, (b) selective recall, and (c) corrective rectification.
- (4) A mitigation plan can lead to more than one outcome with different probabilities.

Notations:

$O = \{o_i\}, i = 1, 2, \dots, m$	vector of mitigation plan costs at different levels.
$P = \{p_j\}, j = 1, 2, \dots, n$	vector of state probability, where p_j denotes the probability of different states of outcomes and $\sum_{j=1}^n p_j = 1$.
$C = \{C_k\}, k = 1, 2, \dots, q$	criteria that should be considered during the warranty risk analysis, where C_k is the k th criterion. The set of C represents the main potential elements that will be influenced once warranty risks have occurred. Two types of criteria are often considered in the decision analysis, costs and benefits (Peng et al., 2011). Thus, we denote C^c and C^b as the criterion sets of cost and benefit type, respectively, where $C^c \cup C^b = C$ and $C^c \cap C^b = \emptyset$. Accordingly, we denote I^c and I^b as the subscript sets of C^c and C^b respectively, where $I^c \cup I^b = \{1, 2, \dots, q\}$ and $I^c \cap I^b = \emptyset$.
$R = \{R_{ijk}\}, i = 1, 2, \dots, m; j = 1, 2, \dots, n; k = 1, 2, \dots, q$	a set of the results of applying a mitigation plan to the impact of a state of the emerged warranty event on C_k . For example, for warranty cost, a positive value means the expected aggregate warranty cost in the next period is increased.
o^r	the reference point of the cost of the mitigation plans. Here, the average warranty cost in the industry can be used as a reference point or the average warranty cost of the product based on the historical record.
$R^r = \{R_k^r\}, k = 1, 2, \dots, q$	vector of the criterion C_k reference point, where the average of C_k values is deemed as the reference point R_k^r . Two types of criteria are often considered in the decision analysis, costs and benefits (Peng et al., 2011). Thus, we denote C^c and C^b as the criterion sets of cost and benefit type, respectively, where $C^c \cup C^b = C$ and $C^c \cap C^b = \emptyset$. Accordingly, we denote I^c and I^b as the subscript sets of C^c and C^b respectively, where $I^c \cup I^b = \{1, 2, \dots, q\}$ and $I^c \cap I^b = \emptyset$.
$\Phi = (\mu^c, \mu^r)$	weight vector of mitigation plan cost and result, where the μ^c denotes the degree of importance of the spending of different mitigation plans, and the μ^r denotes the importance degree of the mitigation plan result; $\mu^c + \mu^r = 1, 0 \leq \mu^c, \mu^r \leq 1$. The value of μ^c, μ^r can be assigned by decision makers.
t_i	Time of making i th decision.
S_i	Expected aggregate warranty cost during $[t_{i-1}, t_i)$.

3. Model Development

The warranty risk mitigation plans will be ranked from the greatest to the smallest, considering $P, O, R, o^r(t), R^r$ and Φ . The solution procedure of the warranty decision problem under uncertainty is briefly discussed as the follow:

- (1) As the impact on manufacturer reputation criterion is qualitatively determined (ordinal data), we need to transform it into interval data in order to use it in the later steps. Likewise, the impact on warranty cost criterion, it will be transformed into interval data (Guilford, 1975).
- (2) The criterion values and mitigation plan costs and their reference points will be normalised through utilising the normalised approach (Kou et al., 2012).
- (3) The value function of CPT will be utilised to calculate the value of R_{ijk} .
- (4) As there are potential interdependences between criteria, i.e. the outcomes of one decision under different criteria are non-additive. Then, the Choquet integral is used to aggregate the outcomes under different criteria of each decision to compare the decisions.
- (5) The weighting function of CPT will be used to determine the weight of the objective probability of each potential result obtained from (4).
- (6) Based on (4) and (5), the prospect value of each mitigation plan treating a certain state can be calculated.
- (7) Based on the prospect values derived from (6), the overall prospect value of each mitigation plan is calculated.
- (8) The ranking of warranty risk mitigation plans can be achieved based on the overall prospect value derived from (7).

In the warranty practice, WDMs may schedule several time points to review the warranty claims related issues and respond to the emerged hazards by selecting the optimal mitigation plan, as discussed above. The chosen plan at this time, however, can affect the future decision concerning the determination of the optimal mitigation plan at the future review points. Accordingly, the reference point of the mitigation plans o^r and the parameters values g_{t_i}, l_{t_i} and λ_{t_i} may be changed. For example, the loss aversion can be greater at t_{i+1} due to some losses experienced in the recent past.

The criterion values and reference point will be normalised into numbers between 0 and 1 in order to unify them. Let $\bar{o}_i$ denote the normalised cost of a warranty mitigation plan. It is then presented as:

$$\bar{o}_i = \frac{o^{\max} - o_i}{o^{\max} - o^{\min}}, \quad i = 1, 2, \dots, m \quad (1)$$

where

$$o^{\max} = \max\{o^r(t), o_1, o_2, \dots, o_m\}, \quad (2)$$

$$o^{\min} = \min\{o^r(t), o_1, o_2, \dots, o_m\}, \quad (3)$$

It can be noticed that $o^r(t)$ is embedded in the Eqs. (2) and (3) in order to maintain the comparability between the reference point of the mitigation plans costs and the o_i values. In Eqs. (1)–(3), the costs of mitigation plans are unified into benefit type (Peng et al., 2011; Kou et al., 2012). In other words, the greater $\bar{o}_i$ is, the better mitigation plan will be.

Let $\bar{R}_{ijk}$ denote the normalisation of the results R with respect to C_k . It is calculated as: $\bar{R}_{ijk} = \frac{R_{ijk} - R_k^{\min}}{R_k^{\max} - R_k^{\min}}$, for $i = 1, 2, \dots, m; j = 1, 2, \dots, n; k \in I^B$, and $\bar{R}_{ijk} = \frac{R_k^{\max} - R_{ijk}}{R_k^{\max} - R_k^{\min}}$, for $i = 1, 2, \dots, m; j = 1, 2, \dots, n; k \in I^C$, where $R_k^{\max} = \max\{R_k^r, R_{11k}, \dots, R_{1nk}, \dots, R_{m1k}, \dots, R_{mnk}\}$, $R_k^{\min} = \min\{R_k^r, R_{11k}, \dots, R_{1nk}, \dots, R_{m1k}, \dots, R_{mnk}\}$, and the reference point R_k^r is embedded to maintain the comparison between the R values and their reference points. Similarly, the criterion values are unified into benefit type (Kou et al., 2012). In other words, the greater $\bar{R}_{ijk}$ is the better result R will be. For example, if R_{ijk} denotes the change of warranty cost, a positive value means the cost increased, and a negative value means the cost decreased; then R_{ijk} should be normalized by the cost type, $\bar{R}_{ijk} = \frac{R_k^{\max} - R_{ijk}}{R_k^{\max} - R_k^{\min}}$, as higher $\bar{R}_{ijk}$ can indicate lower cost. If R_{ijk} denotes the change of reputation; then it

should be normalized by the benefit type, $\bar{R}_{ijk} = \frac{R_{ijk} - R_k^{\min}}{R_k^{\max} - R_k^{\min}}$, as higher $\bar{R}_{ijk}$ can indicate higher reputation. The purpose of normalization is to measure the criteria in the same scale that higher score indicates a better position the decision located in.

In order to determine the new reference point o_{t_i} , WDMs may assess the expected value S_i of the aggregated warranty cost for the interval $[t_{i-1}, t_i]$. This can be computed as (Luo & Wu, 2018): $E[S_i(t_i)] = N(t_i)Z_i$ where N is the number of warranty claims which assumed to follow the non-homogenous Poisson process and Z is the warranty cost.

The reference point of the mitigation plan at the time t_i is determined based on the expected aggregate warranty cost S_i .

$$o^r(t) = f(S_i)$$

where $f()$ can represent the decision maker's attitude on the incurred warranty cost. The function can be monotone increasing, decreasing or other non-monotone forms. Let $\bar{o}_i^r$ denote the normalised reference point pertaining the mitigation plan cost, it is computed as

$$\bar{o}^r = \frac{o^{\max} - o^r(t)}{o^{\max} - o^{\min}} \quad (7)$$

Let $\bar{R}_k^r$ denote the normalised reference point of the C_k criterion, it is calculated as: $\bar{R}_k^r = \frac{R_k^r - R_k^{\min}}{R_k^{\max} - R_k^{\min}}$, for $k \in I^B$, and $\bar{R}_k^r = \frac{R_k^{\max} - R_k^r}{R_k^{\max} - R_k^{\min}}$, for $k \in I^C$.

Thus, based on the values of $\bar{R}_{ijk}$ and $\bar{R}_k^r$, the value of R concerning the criterion C_k can be calculated. If $\bar{R}_{ijk} \geq \bar{R}_k^r$ then R is deemed as a *gain* in respect of C_k criterion. This *gain* can be calculated as $\bar{R}_{ijk} - \bar{R}_k^r$. In contrast, if $\bar{R}_{ijk} < \bar{R}_k^r$ then R is deemed as *loss* in respect of C_k criterion. This *loss* can be calculated as $\bar{R}_k^r - \bar{R}_{ijk}$. In this research, the average of R_k is used as a reference point for each criterion. Terzi et al. (2016) found that the population average payoff level is the modal reference point for individuals when they attempt to make a decision under risk.

Let v_{ijk} denote the value of R with respect to C_k criterion. Thus, based on the value function of CPT the calculation of v_{ijk} at time t_i is computed as:

$$v_{ijk}(t) = \begin{cases} (\bar{R}_{ijk} - \bar{R}_k^r)^{g_k(t)}, & \bar{R}_{ijk} \geq \bar{R}_k^r \\ -\lambda_k(t)(\bar{R}_k^r - \bar{R}_{ijk})^{l_k(t)}, & \bar{R}_{ijk} < \bar{R}_k^r \end{cases}, \quad i = 1, 2, \dots, m; j = 1, 2, \dots, n; k = 1, 2, \dots, q \quad (9)$$

where $g_k(t)$ and $l_k(t)$ are time-varying parameters which represent the time-varying levels of risk-aversion of WDMs over the *gain* and risk-seeking over the *loss* with respect to C_k criterion, respectively. The $\lambda_k(t)$ is a time-varying parameter representing the loss-aversion attitude of WDMs with respect to C_k criterion, where $0 < g_k(t), l_k(t) \leq 1, \lambda_k(t) > 1, k = 1, 2, \dots, q$. Such parameters can be determined through experiments.

The R can be then assessed based on the values resulting from Eq. (9). Let v_{ij} denote the value of the result R_{ij} , then it can be assessed based on the aggregation of $v_{ij1}, v_{ij2}, \dots, v_{ijq}$. However, we assume there is an interdependence between criteria, e.g., between warranty cost criterion and manufacturer reputation criterion. Therefore, the Choquet integration model is used to aggregate $v_{ij1}, v_{ij2}, \dots, v_{ijq}$ into the value of R_{ij} . Accordingly, these values are ranked in a decreasing order through a comparison between them. As a result, the ranking order is presented as $v_{ij(1)} \leq v_{ij(2)} \leq \dots \leq v_{ij(q)}$, where $v_{ij(k)}$ denote the k th smallest result among $v_{ij1}, v_{ij2}, \dots, v_{ijq}, k = 1, 2, \dots, n$. Let $C_{ij(k)}$ denote the criterion corresponding to $v_{ij(k)}$, where $C_{ij(k)} \in C$. Let $u(E_{ij(k)})$ denote the importance degree of criterion subset $E_{ij(k)} = \{C_{ij(k)}, C_{ij(k+1)}, \dots, C_{ij(q)}\}$. The importance can be assigned directly by decision makers or indirectly by using some existing methods such as analytic hierarchy process method (AHP) (Saaty, 2008). It is clear that $E_{ij(1)} = \{C_1, C_2, \dots, C_q\} = \Omega$ and $E_{ij(q+1)} = \emptyset$. Hence, the value of $u(E_{ij(1)}) = 1$ and $u(E_{ij(q+1)}) = 0$. Based on that, the potential value of $v_{ij}(t)$ at time t_i can be calculated as:

$$v_{ij}(t) = \sum_{k=1}^q v_{ij,k} [u(E_{ij,k}) - u(E_{ij,k+1})], \quad i = 1, 2, \dots, m; j = 1, 2, \dots, n \quad (10)$$

The prospect value of each mitigation plan can be calculated based on the values of $v_{i1}, v_{i2}, \dots, v_{in}$ obtained from Eq. (10). To do so, these values are ranked in an increasing order of v_{ij} which is determined through comparing $v_{i1}, v_{i2}, \dots, v_{in}$. Hence, the ranking result should be presented as $v_{i(1)} \geq v_{i(2)} \geq \dots \geq v_{i(n)}$ or $(v_{i(n)} \geq v_{i(n-1)} \geq \dots \geq 0 \geq \dots \geq v_{i(1-m)}, v_{i(-m)})$ where $v_{i(k)}$ is the k th greatest among $v_{i(1)}, v_{i(2)}, \dots, v_{i(n)}$, $k = 1, 2, \dots, n$. Correspondingly, the probabilities of the states are re-indexed as $p_{i(1)}, p_{i(2)}, \dots, p_{i(n)}$, $p_{i(k)} \in P$, $k = 1, 2, \dots, n$. The prospect value V_i of each mitigation plan at time t_i is then computed as:

$$V_i(t) = \sum_{k=0}^n v_{i,k} \pi_{i,k}^+ + \sum_{k=-m}^0 v_{i,k} \pi_{i,k}^-, \quad i = 1, 2, \dots, m, \quad (11)$$

Denote weighting function as $(x, \mu) = x^\mu / (x^\mu + (1-x)^\mu)^{1/\mu}$, and hence the decision weights $\pi_{i,k}^+$ and $\pi_{i,k}^-$ for gains and losses, respectively, can be computed as the following (Tversky & Kahneman, 1992):

$$\pi_{i,k}^+ = w\left(\sum_{z=k}^n p_{i,z}, \mu\right) - w\left(\sum_{z=k+1}^n p_{i,z}, \mu\right), \quad 0 \leq k \leq n-1 \quad (12)$$

$$\pi_{i,k}^- = w\left(\sum_{z=-m}^k p_{i,z}, \mu\right) - w\left(\sum_{z=-m}^{k-1} p_{i,z}, \mu\right), \quad 1-m \leq k < 0 \quad (13)$$

where x is the objective probabilities and $0.27 < \mu < 1$, and $w(\cdot)$ is monotonic depicting the inverse S-shapes (Tversky & Kahneman, 1992). These parameter μ reflects the behaviour of decision makers such as overweighting the outcomes with low probabilities and underweighting the outcomes with moderate and high probabilities, and it is determined through experiments (Tversky & Kahneman, 1992).

Now, we need to calculate the cost of each mitigation plan to assess the overall prospect value OV_i . To do so, we need $\bar{o}_i$ and $\bar{o}^r$. It is important to emphasise that these costs are transformed into benefit type based on the Eqs. (1) and (7). If the $\bar{o}_i \geq \bar{o}^r$, then there is $\bar{o}_i \leq \bar{o}^r$ and hence the mitigation plan can be considered as a *gain* computing as $\bar{o}_i - \bar{o}^r$, while if $\bar{o}_i < \bar{o}^r$, then there is $\bar{o}_i > \bar{o}^r$ which can be deemed as a *loss* calculating as $\bar{o}^r - \bar{o}_i$. Accordingly, let V_i^c denote the value cost of a mitigation plan, according to Eq.(2) the V_i^c can be computed as:

$$V_i^c = \begin{cases} (\bar{o}_i - \bar{o}^r)^{g^c}, & \bar{o}_i \geq \bar{o}^r \\ -\lambda^c (\bar{o}^r - \bar{o}_i)^{l^c}, & \bar{o}_i < \bar{o}^r \end{cases} \quad i = 1, 2, \dots, m, \quad (14)$$

As the cost of each mitigation plan is certain, then the value of cost can be treated as the prospect value of cost. Therefore, the overall prospect value OV_i of each mitigation plan at time t_i can be computed by:

$$OV_i(t) = \mu^r V_i + \mu^c V_i^c, \quad i = 1, 2, \dots, m \quad (15)$$

Based on the result of OV_i , the greater overall prospect value is the better mitigation plan will be. Accordingly, the mitigation plans will be ranked according to their overall prospect value and selected the greatest one to mitigate the warranty risks.

Then, the integrated mitigation policy including the mitigation plans selected at each time can be determined. The overall value of an integrated mitigation policy is $OV = \sum_i^n OV_i$. the optimal integrated mitigation policy can be determined in further research. The time-varying property of this model is represented by the change in the reference point of the cost of the mitigation plans $\bar{o}_i^r$ and the time-varying parameters related the decision maker's behaviour under risk and uncertainty. In further research, for the optimization problem, the objective function is $OV = \sum_i^n OV_i$, which should be maximized with the constraint is $\sum_{i=1}^k o_i \leq \gamma$, where γ is a prespecified total budget.

4. Conclusion

Although warranty plays a significant role nowadays, it may also cause substantial financial and reputational losses for manufacturers. As such, a better planning for warranty risks is extremely important. WDMs therefore should consider proactive plans for mitigating warranty risks when they have occurred.

As such, this paper focuses mainly on developing a tool aiding WDMs to select the suitable mitigation plan based on different criteria. The following points, therefore, have been achieved:

- 1) The average warranty cost in the industry is used to be the reference point for the mitigation plan costs.
- 2) Since the reference point of mitigation plans costs and the estimated parameters capturing the WDMs behaviour can change over the warranty period, the time dependence is integrated to the CPT-WaRM to respond to such changes.

This model may aid WDMs in evaluating and ranking different mitigation plans when considering those psychological aspects that may affect the final decision such as reference dependence, loss aversion, diminishing sensitivity, and probability weighting. In addition, the optimal mitigation plan can be determined at different revision points of warranty claims based on the first selected plan.

References

- Bell, D. E. 1982. Regret in decision making under uncertainty. *Operations Research*, 30, 961-981.
- Bell, D. E. 1985. Disappointment in decision making under uncertainty. *Operations Research*, 33, 1-27.
- Costantino, F., De Minicis, M., González-Prida, V. & Crespo, A. 2012. On the use of Quality Function Deployment (QFD) for the identification of risks associated to warranty programs. *ESREL Conference*.
- Díaz, V. G. & Márquez, A. C. 2011. On the risks and costs methodologies applied for the improvement of the warranty management. *Journal of Service Science and Management*, 4, 191.
- González-Prida, V. & Márquez, A. C. 2012. A framework for warranty management in industrial assets. *Computers in Industry*, 63, 960-971.
- Guilford, J. P. 1975. *Psychometric methods*, Bombay, Tata McGraw-Hill.
- Kahneman, D. 1979. Prospect theory: An analysis of decisions under risk. *Econometrica*, 47, 278.
- Kou, G., Lu, Y., Peng, Y. & Shi, Y. 2012. Evaluation of classification algorithms using MCDM and rank correlation. *International Journal of Information Technology & Decision Making*, 11, 197-225.
- Luo, M. & Wu, S. 2018. A mean-variance optimisation approach to collectively pricing warranty policies. *International Journal of Production Economics*, 196, 101-112.
- Peng, Y., Kou, G., Wang, G. & Shi, Y. 2011. FAMCDM: A fusion approach of MCDM methods to rank multiclass classification algorithms. *Omega*, 39, 677-689.
- Saaty, T. L. 2008. Decision making with the analytic hierarchy process. *International journal of Services Sciences*, 1, 83-98.
- Schmidt, U., Starmer, C. & Sugden, R. 2008. Third-generation prospect theory. *Journal of Risk and Uncertainty*, 36, 203.
- Terzi, A., Koedijk, K., Noussair, C. N. & Pownall, R. 2016. Reference point heterogeneity. *Frontiers in Psychology*, 7, 1347.
- Tversky, A. & Kahneman, D. 1992. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5, 297-323.
- Velasquez, M. & Hester, P. T. 2013. An analysis of multi-criteria decision making methods. *International Journal of Operations Research*, 10, 56-66.

Integration of Reliability, Availability and Maintainability Analysis and Operational Research Modelling – A Case Study

J. Bishop*, M. Niblett** and M. McCarthy***

*Sellafield Ltd., jacqueline.l.bishop@sellafieldsites.com

**ZUP Badura Ltd., External Consultants to Sellafield Ltd., mike.niblett@zupbadura.com

***Risk Based Decisions Ltd., External Consultants to Sellafield Ltd

Abstract. A facility on the Sellafield site is being designed which will receive waste from several of the site's high hazard facilities during their decommissioning.

Sellafield Ltd. uses a process called T-RAM (Throughput – Reliability, Availability and Maintainability) to develop a reasoned auditable argument that a new plant, and its systems and subsystems, will meet its performance targets. T-RAM activity is focused on areas of the plant that represent the most T-RAM risk as determined through a top-down proactive approach to risk identification. Identified T-RAM risk is mitigated through improved design supported by the development of effective operations, maintenance and asset care strategies using techniques including Failure Mode Effects and Criticality Analysis (FMECA); Reliability Centered Maintenance (RCM) and Reliability Block Diagrams (RBD).

In order to underpin the “Throughput” capability of the facility, the Operational Research (OR) Group has developed a simulation model of the facility which has been used to feedback the expected throughput based on current best knowledge of plant configuration, equipment cycle times, reliability and maintainability data. This is an iterative process, with high risk bottleneck areas identified by the simulation model becoming the focus for T-RAM progressive risk reduction activities. The project is an excellent example of the OR and T-RAM teams working together to influence facility design.

1. Background

Sellafield Ltd is responsible for the safe and secure operation and clean-up of the Sellafield nuclear site. Focus is on the decommissioning of the two highest hazard facilities; the Magnox Swarf Storage Silo (MSSS) and First Generation Magnox Storage Pond (FGMSP).

MSSS is a 50 year old radioactive waste store on the Sellafield site, operational from 1964 until the final routine waste “tip” in 1992, Sellafield Ltd. (2015). The 22 compartment building now contains radioactive waste from the UK's first generation Magnox power stations. FGMSP was operational for 26 years, processing around 27,000 tonnes of nuclear fuel, Sellafield Ltd. (2016). It now contains a mix of radioactive sludge waste and solid nuclear fuel.

A facility is being designed and built to receive and process waste from these, and other, high hazard legacy facilities to enable safe interim storage pending a final disposal route.

2. The T-RAM Process

T-RAM is a progressive and proactive process used within Sellafield Ltd. to identify and mitigate risks that would otherwise threaten the required performance of physical assets throughout their intended life. The focus of the analysis is to assure that asset(s) have sufficient capability (functional performance) and availability to contribute to the successful achievement of throughput requirements.

The asset being analysed could be something as small as a component or alternatively it could be a sub-system, a whole Plant Based System, a Facility or even a whole Value Stream – the question of whether assets are meeting their T-RAM requirements can effectively be asked at any level.

For example T-RAM can be applied to a major project to design and build a brand new processing facility to remediate waste from legacy ponds and silos, right through to a much more modest plant modification to replace the type of pump used on an existing plant. In the former case the T-RAM process will seek to progressively substantiate that the high level T-RAM requirements detailed in a Project Functional Specification can be achieved, whilst in the latter the T-RAM process may simply seek to verify that changing the pump will mean that local sub-system availability can be improved to meet performance targets. Of course, even in the case of the pump replacement the local piece of T-RAM analysis is effectively

carried out in the context of the bigger picture – i.e. how improving the T-RAM of the specific pump will contribute to improving the T-RAM of the facility in which the pump resides.

Regardless of the scope, the application of T-RAM can be considered through answering the following three basic questions:

1. What do the assets need to deliver?
2. What is the risk of them not delivering?
3. How can the risk of non-delivery be managed/mitigated to acceptable levels?

In the case of building a processing facility, question 1 starts with ensuring that all the high level requirements detailed within the Project Functional Specification (implicitly including T-RAM requirements) are suitably resolved and progressively disseminated down to areas/functions, systems and ultimately to sub-systems/components, where appropriate, as part of the project requirements management process.

The disseminated T-RAM requirements ensure that individual design teams are targeted on developing a design that will contribute to delivering overall facility throughput requirements (as well as meeting appropriate statutory or regulatory requirements). Verification that T-RAM requirements are being achieved is sought as early as possible within the project lifecycle i.e. commencing with design supported by further and progressive verification managed appropriately through purchasing (e.g. vendor requirements), manufacture (e.g. quality assurance), works testing/factory acceptance testing (e.g. functional performance, condition assessment/baselining, maintainability testing), construction and commissioning and into operations.

A range of methods and tools are deployed in a graded way to proactively identify T-RAM risks as early as possible and to provide insight into how they may be mitigated or eliminated in order to reduce performance risk to acceptable levels and progressively verify that project T-RAM requirements should be met. A fundamental and overarching tool that is used within Sellafield major projects to help achieve this is throughput modelling which may take the form of relatively simple spreadsheet based models early on in the project life cycle followed by the development of more complex discrete event simulation models as the design progresses.

In order to support timely completion of the processing facility, the various systems of design and associated procurement packages vary in terms of level of progression through the project lifecycle stages. For example, some packages of equipment have been procured and are undergoing construction whilst some systems are still in the latter stages of detailed design. Appropriate T-RAM analysis through the use of tools including FMECA, RBD, Support Options Analysis (SOA) and RCM are being employed in a timely manner on a system by system basis using a risk based graded approach taking account of Operational Experience from the use of similar assets both inside and outside of Sellafield. Where relevant the T-RAM analysis is now being augmented with Original Equipment Manufacturer (OEM) and vendor information through appropriate detailed design and procurement phases. As the maturity of equipment performance (e.g. cycle time), reliability and maintainability progresses the information is periodically updated within the project OR (Throughput) model to simulate and analyse the behavior of the plant both with equipment breakdowns disabled (to predict the demonstrable throughput) and with equipment breakdowns enabled (to predict the effective throughput). This includes the appropriate use of more specific bottleneck, sensitivity and scenario analysis to assist with the development of the remaining engineering design and associated support strategies - in support of the project T-RAM case.

3. The Processing Facility Operational Research Model

A discrete event simulation model of the processing facility has been developed by the Sellafield Operational Research (OR) Group. Its main function is to help prove or otherwise the ability of the facility to meet its design functional requirements and where gaps are identified to help target areas for improvement. The analysis includes an assessment of import, processing and export rates between the processing plant, its associated product store and its donor facilities. A schematic of the facility is shown in figure 1, and an illustration of the simulation in figure 2.

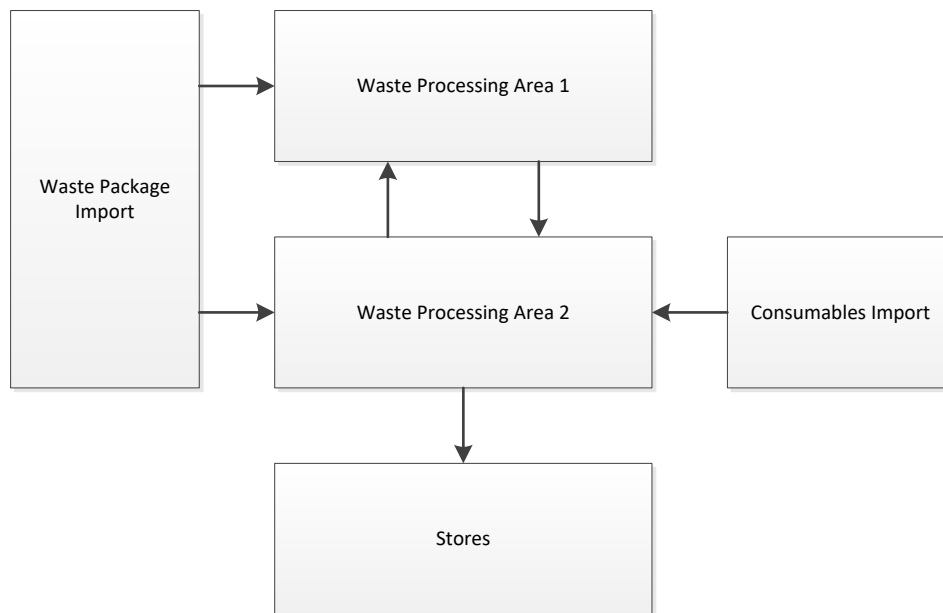


Figure 1. High Level Schematic of the Processing Facility

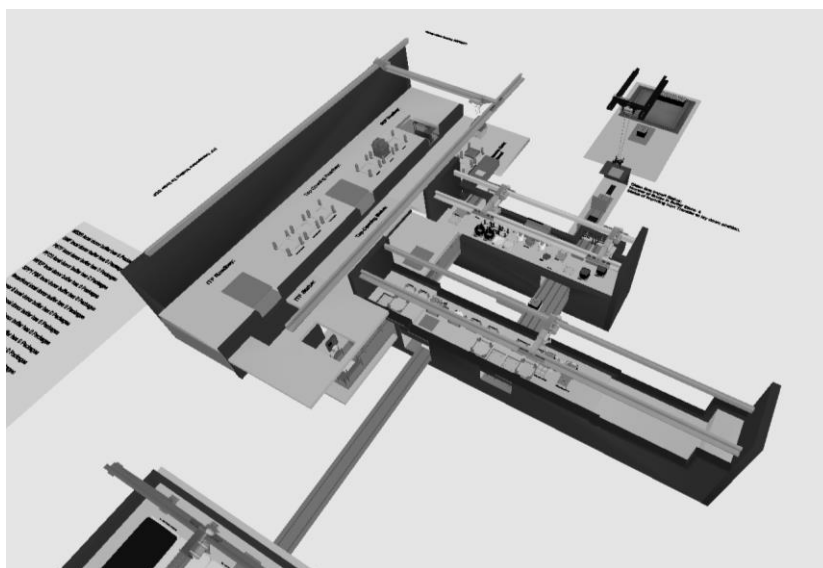


Figure 2. High Level View of the Processing Facility Simulation Model

The model inputs cover a wide range of data types, with over 500 parameters which can be adjusted to represent a vast range of scenarios. The data being used to drive the model is being progressively developed by the project T-RAM team using the methods described above. The model, together with simplified earlier model variants, has already been used to support project review milestones by providing progressive throughput analysis relating to the various waste streams that the facility is tasked with processing.

The following sections present specific (non-classified) examples of the type of analysis undertaken to help verify that the plant, as designed, should be capable of meeting its T-RAM requirements including specific scenario analysis to assess the impact of robot redundancy in one of the main processing areas and an improved means of recovery from a roadbay door failure.

3.1 FGMSP waste processing

The OR model has been used to predict whether the processing facility should be capable of supporting performance requirements for the import, processing and export to store of packages of waste received from the FGMSP waste stream. To achieve this, the model was run in a number of different modes and scenarios including:

- An unlimited number of waste packages made available to be imported to the model on demand
- A fixed and discrete daily schedule of waste packages presented to the model
- Scenarios both with equipment breakdowns enabled and breakdowns disabled
- Scenarios covering a credible range of process variations including effluent settling times and grout curing times (sensitivity analysis)

Key conclusions drawn from the analysis included:

- With baseline process cycle times the plant was not capable of consistently achieving the required daily throughput
- Despite this, a significant headroom in plant capacity was demonstrated against the foreseeable peak yearly processing requirements
- Sensitivity analysis demonstrated that a circa 30% reduction in grout curing times generated a 62-65% throughput performance increase
- By delivering waste packages to the model against a regular daily schedule and taking credit for reduced grout curing and effluent settling times allowed daily throughput requirements to be fully demonstrated

3.2 MSSS waste processing

A key requirement of the processing facility is to support the decommissioning of MSSS through timely receipt and processing of waste retrieved, packaged and dispatched from the MSSS facility. The inability of the facility to receive MSSS waste packages due to equipment unavailability, for example, can lead to the undesirable effect of constraining the MSSS retrievals process.

3.2.1 Robot Redundancy

The baseline design includes two robots to perform in-cell waste-processing duties, with a defined list of activities assigned to each robot. Due to the nature of the nuclear environment in which they reside, repair or replacement of either one of the robots in the event of failure requires significant recovery time (measured in days rather than hours) in order to remotely prepare, disconnect and remove the failed robot and then replace it with a spare one. Initial model analysis was performed assuming that upon failure of either robot all operations in the processing cell would stop until safe recovery and replacement could take place.

A change in MSSS waste processing requirements, to a simplified sequence of mainly automated processing steps, presented the opportunity for the robot sequences to be reconsidered in such a way that each robot, whilst configured slightly differently, could perform the required MSSS processing tasks single handedly. This presents the opportunity of leaving a failed robot in-situ whilst continuing MSSS processing with the other robot until a suitable planned or forced outage occurs to enable the failed robot to be replaced without directly impacting throughput. The potential benefits of adopting this pseudo duty/standby robot arrangement for processing the MSSS waste stream was configured and tested using the OR model by adjusting the baseline robot recovery and reinstatement time down to only 2 hours – to instead represent the anticipated time it would take to manipulate the failed robot into a safe (retracted) position and configure the second robot to take up the failed robot's duties.

3.2.2 Roadbay Door Failure Mitigation

Vehicular (transporter) access to the plant is gained via two roadbay doors in the Waste Package Import Area (Figure 1) with one door designated for entry and the other for exit. Similar to the robots the baseline T-RAM analysis indicated potentially significant repair times in the event of roadbay door failure and baseline model analysis shows the doors failing during operation and therefore blocking a waiting transporter from either entering or leaving the facility.

Subsequent T-RAM analysis of the door design indicated that, in the event of the majority of failures, the doors could be manually operated or, for more significant failures, that vehicle entry/exit might still be possible via controlled use of the remaining operational door. To test the potential impact of these proposed operational workarounds using the model each of the two roadbay door repair times was adjusted down to 2 hours to represent the time it would take to implement a change in operational practice to enable continued transporter entry and exit from the facility. The intent being that repair of the failed door could be managed in an opportune and timely manner with little or no direct impact on throughput..

3.2.3 Model results

Table 1 below summarises the model output covering the two scenarios described in 3.2.1 and 3.2.2 above in terms of percentage improvement against baseline. Package rejection rate (i.e. percentage of package arrivals from MSSS which could not enter the processing facility in a timely fashion) was used as a measure of facility unavailability and associated extension impact on the MSSS retrievals programme.

Table 1. OR Model Results

Scenario description	Percentage improvement against baseline package rejection rate
Baseline except with 2 hour repair times against each of the two robots (to represent changeover of duty rather than repair)	6.7%
Baseline except with 2 hour repair times on both robots (as above) plus 2 hour repair times on each of the two roadbay doors (to represent a change in operational practice rather than repair)	20%

The model results show that improvements to the operations and maintenance strategy of the robots and the roadbay doors combined could have a significant impact on plant availability and throughput (measured in terms of a 20% reduction in the percentage of packages rejected over baseline). The results also indicate that the roadbay door improvements contribute more than the robot changes. This is not unexpected as delays in transporter access to the plant because of erstwhile roadbay door unavailability would very quickly impact import rates. In contrast, the robots are located further along within the process enabling buffer positions and work in progress to act to reduce the impact of robot unavailability on facility import rates.

3.4 Distributed Simulation

The OR model was initially developed to provide an assessment of stand-alone facility performance across a number of waste feeds including but not limited to the MSSS waste stream. More recently the model has been further enhanced and dynamically linked to the MSSS facility discrete event simulation model using an innovative technique called Distributed Simulation. This allows the OR team to not only understand the

performance of individual facilities, but also how they interact with each other to provide the business with a Value Stream throughput assessment.

4. Conclusions and Next Steps

The OR model for the processing facility contributes an essential element of the developing T-RAM case in that it allows progressive assessment of throughput capability and scenario/sensitivity analysis as the project moves through its lifecycle stages from concept design, through preliminary and detailed design to procurement, installation and commissioning.

In addition to updating the model to more accurately reflect current design and agreed operational strategy the next phase of work using the OR model is to further improve understanding of the impact of key equipment repair times on package import rates (e.g. to identify the maximum repair times for key equipment which, if achieved, would have little or no impact on throughput). This will allow the T-RAM team to focus attention on key areas of the plant and optimise spares and maintenance strategy development activities to maximise package import availability in support of high hazard reduction within donor facilities such as MSSS an FGMSP.

References

- Sellafield Ltd (2015) Sellafield Magazine: Issue 1. *In Focus: Magnox Swarf Storage Silo*,
<https://www.gov.uk/government/publications/sellafield-magazine-issue-1>, pp.14-17.
- Sellafield Ltd (2016) Sellafield Magazine: Issue 4. *In Focus: First Generation Magnox Storage Pond*,
<https://www.gov.uk/government/publications/sellafield-magazine-issue-4>, pp.20-23.

Estimation and goodness-of-fit tests for imperfect maintenance models: single system vs multi-system

L. Brenière*, L. Doyen**, O. Gaudoin***

*Univ. Grenoble Alpes, CNRS, Grenoble INP, email address: lea.breniere@univ-grenoble-alpes.fr

**Univ. Grenoble Alpes, CNRS, Grenoble INP, email address: laurent.doyen@univ-grenoble-alpes.fr

***Univ. Grenoble Alpes, CNRS, Grenoble INP, email address: olivier.gaudoin@univ-grenoble-alpes.fr

Abstract Let us consider repairable systems subject to failures and maintenances. Observations can come from only one system or from several identical and independent systems. The maintenance is supposed to be imperfect. This is an intermediate situation between the case where a maintenance renews the system, called As Good As New, and the case when the sole effect of maintenance is to get the system working again, called As Bad As Old. Imperfect maintenance models have a parameter that can be interpreted as the maintenance efficiency. Other parameters lie in the baseline hazard rate of the first failure time. All these parameters can be estimated through maximisation of the log-likelihood, which takes into account the single system or multi-system nature of the data. The first aim of this paper is to compare the quality of parameter estimation in the single and multi-system cases. The second aim is to compare the quality of goodness-of-fit tests in this framework. Moreover, new goodness-of-fit tests are introduced in the multi-system case.

1. Introduction

Throughout its lifetime, a repairable industrial system undergoes maintenances and failures. After each unforeseeable failure, the system is repaired. This is usually known as corrective maintenance. The repairing times are neglected. In this article, we will not consider preventive maintenances. Two types of data sets will be considered and compared. The single system case corresponds to the observation of the n first failure times of a single system. The multi-system case corresponds to the observation of v identical and independent systems with n failures each. The failure times are denoted $(T_{j,i})_{1 \leq j \leq v, 1 \leq i \leq n}$. The single system is the particular case for which $v = 1$. Then, for the sake of simplicity, $T_{1,i}$ will be denoted T_i for single systems.

This paper focuses on imperfect maintenance models such as the virtual age models of Kijima (1989). These models depend on parameters which both characterize the intrinsic wear out and the maintenance efficiency. When a model is applied to a data set, these parameters have to be estimated. We can legitimately wonder if this estimation is of good quality. This has already been done in Doyen and Gaudoin (2004) for the single system case, but no study exists for the multi-system case. Even though we have as much data in a single system with 15 failures as in 3 systems with 5 failures each, the information given will not be the same in both cases. Therefore, our aim is to compare different situations with the same total number of failures $v \times n$. Simulation studies have been done with the R package VAM (<https://rpackages.u-ga.fr/VAM>) developed by LJK laboratory in Grenoble.

We can also ponder on the fact that the chosen model is adapted or not to the data. One way to address this question is to perform goodness-of-fit (GoF) tests, likewise in Chauvel et al. (2016a). In that article, only one system was considered at a time. Here we will develop GoF tests able to take into account the multi-system nature of our data and we will compare the GoF tests in the single and multi-system cases.

2. Imperfect maintenance models

Let us assume that we observe the n first failure times of a single repairable system $T_1, T_2, \dots, T_n$, with $T_0 = 0$. Let $N = (N_t)_{t \geq 0}$ denote the counting process of failures. A stochastic model for this process of recurrent events is completely characterized by the intensity $\lambda = (\lambda_t)_{t \geq 0}$ of the failure counting process defined as

$$\lambda_t = \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \mathbb{P}(N_{t+\Delta t} - N_t = 1 | \mathcal{H}_t^-), \quad t \geq 0, \quad (1)$$

where $\mathcal{H}_t^-$ is the past of the process just before t and N_t^- represents the left-hand limit of N_t .

An imperfect maintenance model is composed of two parts. The first part of the model is the initial intensity, which corresponds to the hazard rate of the first time to failure and expresses the intrinsic wear of the new

system before the first maintenance. Usually, the first failure time is assumed to follow a Weibull distribution, so that the initial intensity is the intensity of a Power Law Process (PLP), $h(t) = \frac{\beta}{\eta} \left(\frac{t}{\eta}\right)^{\beta-1}$, for $t \geq 0$ and $\eta, \beta > 0$. β is a shape parameter while η is a scale parameter. The system is wearing for $\beta > 1$ and improving for $\beta < 1$. The second part of the model characterizes the maintenance effect on the system. The repair can be minimal, in which case the system is As Bad As Old (ABAO) and the counting process is a Non Homogeneous Poisson Process of intensity $\lambda_t = h(t)$, $t \geq 0$. The repair can also leave the system As Good As New (AGAN). The corresponding counting process is a Renewal Process of intensity $\lambda_t = h(t - T_{N_t-})$, $t \geq 0$. Between these two extreme cases, there is a wide range of models corresponding to imperfect maintenances.

In the Brown-Proschan (BP) model proposed in Brown and Proschan (1983), each maintenance is AGAN with probability $p \in [0, 1]$ and ABAO with probability $1 - p$.

Another usual imperfect maintenance model is the Quasi-Renewal (QR) model, proposed in Wang and Pham (1996). The corresponding counting process is also known as the Geometric Process and its intensity is

$$\lambda_t = q^{-N_t-} h(q^{-N_t-} (t - T_{N_t-})), \quad t \geq 0, \quad (2)$$

where $q > 0$ is a parameter characterizing the effect of the repair. The case $q = 1$ corresponds to the AGAN model. If $q \in]0, 1[$, the system deteriorates due to stochastically decreasing inter-failure times and if $q > 1$ the system improves.

The last class of imperfect maintenance models presented in this paper is the class of virtual age models, exposed in Kijima (1989). The models of this class depend on a sequence of positive random variables $(A_i)_{i \in \mathbb{N}^*}$ called effective ages. After the i th failure, the system behaves like a new system which has not failed until A_i , for $i \in \mathbb{N}^*$. The failure intensity is

$$\lambda_t = h(A_{N_t-} + t - T_{N_t-}). \quad (3)$$

A virtual age model is defined by a particular expression of the effective ages. For instance, in the ARA_∞ model or Kijima type II model with deterministic constant effect exposed in Doyen and Gaudoin (2004), the repair is supposed to reduce the effective age by a factor $\rho \leq 1$. The corresponding intensity is

$$\lambda_t = h\left(t - \rho \sum_{k=0}^{N_t-1} (1 - \rho)^k T_{N_t-k}\right), \quad t \geq 0. \quad (4)$$

In the ARA_1 model or Kijima type I with deterministic constant effect exposed in Doyen and Gaudoin (2004), the supplement of age since the last failure is reduced by a factor $\rho \leq 1$ and

$$\lambda_t = h(t - \rho T_{N_t-}), \quad t \geq 0. \quad (5)$$

Between the ARA_1 and ARA_∞ model lie the ARA_m models, for $m \in \mathbb{N}^*$ (the memory of the model), with intensity

$$\lambda_t = h\left(t - \rho \sum_{k=0}^{\text{Min}(m-1, N_t-1)} (1 - \rho)^k T_{N_t-k}\right), \quad t \geq 0. \quad (6)$$

In all ARA models, the value of ρ represents the effect of repair. If $\rho \in]0, 1[$ the repair is efficient, if $\rho = 1$ the repair is optimal (AGAN), if $\rho = 0$ the repair is minimal (ABAO) and if $\rho < 0$ the repair is harmful. In this paper, we will not consider harmful repairs so that $\rho \in [0, 1]$.

3. Quality of estimation

As we have seen, imperfect maintenance models depend on some parameters that have to be estimated from data. This is actually done through numerical maximisation of the log-likelihood. The likelihood function associated to the observation of the n first failure times of v systems $(T_{j,i})_{1 \leq j \leq v, 1 \leq i \leq n}$ is

$$L_{v,n} = \left(\prod_{j=1}^v \prod_{i=1}^n \lambda_{T_{j,i}} \right) \exp \left(- \sum_{j=1}^v \int_0^{T_{j,n}} \lambda_t dt \right). \quad (7)$$

The maximisation of the log-likelihood is done under box constraints for β and ρ , so that we have $\beta > 1$ (in order to have a baseline hazard rate corresponding to a wearing system) and $0 < \rho < 1$ (corresponding to an efficient repair).

We have performed a simulation study in order to assess the quality of the parameter estimations both in the single and multi-system cases. The framework of the study is described hereafter.

- Simulated models: $ARA_1, ARA_2, ARA_3, ARA_\infty, QR$.
- Parameters values: $\eta = 1, \beta = 2.5$, and ρ (and q) depending on the simulated model, in order to obtain a comparable wear-out.
- Single system case ($v = 1$): $n \in \{9, 15, 21, 30, 60, 90\}$.
- Multi-system case: $v \in \{3, 5, 7, 10, 20, 30\}$ and $n = 3$.

We compare the two cases with an equal amount of data: for example, the estimation of parameters with a system of 21 failures is contrasted with the case of 7 systems facing 3 failures each. In each configuration, 5000 replications are simulated. The quality of the parameter estimations is assessed by computing the means, standard deviations, mean squared errors and boxplots of the 5000 estimates.

The parameter η is always very well estimated: bias and standard deviation are low and decrease with the number of data. This is a significant progress with respect to the estimation results obtained with the historical parameterization $h(t) = \alpha\beta t^{\beta-1}$ used in Doyen and Gaudoin (2004). The results for β are also satisfying but show an overestimation. The estimation of ρ is clearly worse with an overestimation and, in the single system case, a bias and standard deviation that seem to stabilize even when the number of data increases.

For both η and β , the estimations are of much better quality in the multi-system case than in the single system case. Indeed, the estimations are not only closer to the true value, but also clearly convergent with the increase of the number of systems. η and β are better estimated certainly because with multiple systems, there are multiple realizations of the first failure time, and therefore it is observed as many times as there are systems.

The conclusion for ρ is not so sharp. On one hand, the stabilization of the estimation quality observed in the single system case is no longer true in the multi-system case. On the other hand, the estimation quality is not necessarily better in the multi-system case. In fact, a rather large number of failures per system is needed in order to observe and then estimate correctly the efficiency of maintenance.

The fact that β and ρ are simultaneously overestimated can come from a compensation between the two parameters: a system with a very efficient repair and a strong wearing is similar to a system with a poor repair but with a low wearing.

Finally, the parameters of the QR model are better estimated than the ones of the ARA models. This is probably due to the fact that for this model, the inter-failure times are independent.

4. Goodness-of-fit tests: theory

We need to know whether a given imperfect maintenance model is adapted or not to a given data set. The authors in Chauvel et al. (2016a) developed Goodness-of-Fit (GoF) tests to address this issue. This work has been presented at MIMAR 2016 (see Chauvel et al. (2016b)). However, they have not considered multi-system data. Therefore, our goal is to find GoF tests for the multi-system case.

Given that a point process model is characterized by its intensity, an imperfect maintenance model can be denoted $\mathcal{C} = \{\lambda(\theta), \theta \in \Theta \subset \mathbb{R}^d\}$, where θ is the model parameter. We want to determine whether $\mathcal{C}$ is a relevant model for the observed data $(T_{j,i})_{1 \leq j \leq v, 1 \leq i \leq n}$. A GoF test is a statistical test of $H_0 : \lambda \in \mathcal{C}$ versus $H_1 : \lambda \notin \mathcal{C}$.

Our first idea was to extend one of the classes of tests exposed in Chauvel et al. (2016a): the statistics based on the Probability Integral Transform (PIT). Since the intensity is parametric, let

$$\Lambda_t(\theta) = \int_0^t \lambda_s(\theta) ds, \quad \forall t \geq 0, \quad (8)$$

be the cumulative intensity of the failure process for a given parameter θ . For a single system, the principle is that under H_0 , the random variables $\Lambda_{T_{i+1}}(\theta) - \Lambda_{T_i}(\theta)$ are i.i.d. with standard exponential distribution. Then,

the authors proposed to transform these variables into uniform ones by applying:

$$U_i := \exp(-\Lambda_{T_{i+1}}(\theta) + \Lambda_{T_i}(\theta)), \quad 0 \leq i \leq n-1. \quad (9)$$

Their idea was that if the model is not the proper one for the data, these U_i are no longer uniform variables. Therefore, the main point of the test procedure was to test their uniformity. The authors proposed three tests statistics: Kolmogorov-Smirnov, Cramér-von Mises and Anderson-Darling.

To extend this procedure to the multi-system case, we compute

$$U_{j,i} := \exp(-\Lambda_{T_{j,i+1}}(\theta) + \Lambda_{T_{j,i}}(\theta)), \quad 1 \leq j \leq v, 0 \leq i \leq n-1. \quad (10)$$

Under H_0 , these variables should still be i.i.d. with uniform distribution, since our systems are independent. Our idea was then to compute the three test statistics, and to perform the test in the same way as in Chauvel et al. (2016a). However, as explained later, this procedure performed very poorly compared to the initial one. Therefore, we looked for another way to use the $U_{j,i}$ with their matricial structure.

We chose to focus on the standard deviation of the $U_{j,i}$. On the $v \times n$ matrix $(U_{j,i})_{1 \leq j \leq v, 0 \leq i \leq n-1}$, we compute line by line the standard deviation for each system. Their mean is

$$M^S := \frac{1}{v} \sum_{j=1}^v \sqrt{\frac{1}{n-1} \sum_{i=0}^{n-1} (U_{j,i} - \overline{U_{j,\bullet}})^2}. \quad (11)$$

Then we shuffle all the $U_{j,i}$ and reconstruct a $v \times n$ matrix with these shuffled variables. We compute M_1^S in the same way we computed M^S but with the shuffled $U_{j,i}$. We repeat this procedure 1000 times so that we get $(M_k^S)_{1 \leq k \leq 1000}$. If the $U_{j,i}$ are independent, which should be the case under H_0 , then the shuffling operation will not affect the distribution of M_k^S . Therefore, the idea of the new GoF test is to compare the $(M_k^S)_{1 \leq k \leq 1000}$ to the value of M^S computed on the initial $U_{j,i}$. The test statistic is the mean of the absolute difference between M^S and the $(M_k^S)_{1 \leq k \leq 1000}$:

$$S_S = \frac{1}{1000} \sum_{k=1}^{1000} |M^S - M_k^S|. \quad (12)$$

The null hypothesis H_0 will be rejected if S_S is too large.

The S_F statistic is computed similarly to S_S but we transpose the matrix: we compute column by column the standard deviation for each failure index. We have

$$M^F := \frac{1}{n} \sum_{i=0}^{n-1} \sqrt{\frac{1}{v-1} \sum_{j=1}^v (U_{j,i} - \overline{U_{\bullet,i}})^2}, \quad (13)$$

$$S_F = \frac{1}{1000} \sum_{k=1}^{1000} |M^F - M_k^F|. \quad (14)$$

Because of the estimation of θ , the $U_{j,i}$ can not be computed. Then, the test statistics will be computed by replacing θ by $\hat{\theta}$. Their distribution under H_0 is unknown, so we cannot use classical tables to determine the critical region. Therefore, we use a parametric bootstrap approach to compute the quantiles of the test statistics distributions under the null hypothesis. This test procedure is described in detail in Chauvel et al. (2016a). The idea is to compute the test statistic from the initial dataset, to simulate a large number of datasets under the estimated model and to compute the test statistic for all these new datasets. We can compute a p-value as the proportion of newly computed statistics that are greater than the statistic computed from the initial dataset.

5. Goodness-of-fit tests: simulation

We have performed a simulation study in order to assess the quality of the GoF tests in the multi-system case. The framework of the study is described hereafter.

- Simulated models: ARA₁, ARA_∞, QR, BP.
- Tested models: ARA₁, ARA_∞, QR.
- Multi-system case: $v = 6$ and $n = 5$.

The procedure to compute the empirical powers of the GoF tests is the following: 1000 data sets of 6×5 failures are simulated under a given model. For each data set, the test statistics are computed and the p-value for the null hypothesis is calculated with a parametric bootstrap. The empirical power is then the proportion of rejection of the null hypothesis, that is the proportion of p-values under the test level (here 5 %) when data are simulated according to a model different from the one that is tested.

First of all, the GoF tests proposed in Chauvel et al. (2016a) applied to the multi-system case appear to be significantly less powerful than in the single system case. That is why we have proposed the new GoF tests based on the matricial structure of the observations. It appears that S_F and S_S have opposite behaviors: depending on the cases, one of them is more powerful than the original tests, and the other has low power. At the end it is unfortunately not possible to say that one test is globally better than the other. However, as in Chauvel et al. (2016a), the test powers remain globally quite low.

It seems difficult to differentiate ARA_1 and ARA_∞ models. The difficulty can come from the too few failures (only 5) that are in each system of our simulated data sets. Indeed, the first two failure times of these models have the same probability distribution. Therefore, there are only three failures left per system to differentiate the two models. Indeed, with 10 failures per system the tests powers increase significantly.

Globally, the tests involving the QR model show much better performances, as in the single system case.

6. Conclusion

We have compared the quality of estimation of the parameters of imperfect maintenance models in the single and multi-system cases. Since the initial intensity is linked to the first failure time, in order to obtain good estimations of η and β , it is better to observe several independent replications of this first failure time, as in the multi-system case. Conversely, since p (or q) corresponds to the maintenance efficiency, to have a good estimation quality, we must have enough successive failure times per system.

Since the goodness-of-fit tests presented here assess the goodness-of-fit of the maintenance part of the virtual age model, it is understandable that it is difficult to distinguish between models in the multi-system case, because the sequence of failures for each system is too short to see the differences between these very similar and flexible models. If we try to discriminate between different initial intensities, we may have better results than in the single system case, as for the quality of estimation.

Acknowledgements

This work has been partially supported by the LabEx PERSYVAL-Lab (ANR-11-LABX-0025-01) funded by the French program Investissement d'avenir.

References

- Brown, M. and Proschan, F. (1983). Imperfect repair. *Journal of Applied probability*, 20(4):851–859.
- Chauvel, C., Dauxois, J.-Y., Doyen, L., and Gaudoin, O. (2016a). Parametric bootstrap goodness-of-fit tests for imperfect maintenance models. *IEEE Transactions on Reliability*, 65(3):1343–1359.
- Chauvel, C., Dauxois, J.-Y., Doyen, L., and Gaudoin, O. (2016b). Parametric bootstrap goodness-of-fit tests for imperfect maintenance models. In *MIMAR 2016 - 9th IMA International Conference on Modelling in Industrial Maintenance and Reliability*, London, United Kingdom.
- Doyen, L. and Gaudoin, O. (2004). Classes of imperfect repair models based on reduction of failure intensity or virtual age. *Reliability Engineering & System Safety*, 84(1):45–56.
- Kijima, M. (1989). Some results for repairable systems with general repair. *Journal of Applied probability*, 26(1):89–102.
- Wang, H. and Pham, H. (1996). A quasi renewal process and its applications in imperfect maintenance. *International journal of systems science*, 27(10):1055–1062.

Applications of the gamma process in storage models: Energy Harvesting Systems

I.T. Castro*, L. Landesa**, A. Serna***

Escuela Politécnica, Universidad de Extremadura (Spain)

inmatorres@unex.es, llandesa@unex.es**, serna@unex.es****

Abstract The gamma process is a stochastic process extensively used to model the degradation of a system over the time. In this paper, gamma process is used in a different perspective: to model the harvested energy for an Energy Harvesting System (EHS). Energy harvesting refers to harnessing energy from the environment or other energy sources and converting it to electrical energy. Specifically, in this work, gamma process is used to model the harvested energy for a Radio Frequency Energy Harvesting System. A management policy of this EHS is described.

1. Introduction

In general terms, an energy harvesting system is a system that captures small amounts of readily available energy in the environment and converts it into usable electrical energy. Nowadays, EHS's are emerging as a promising technology for charging batteries used to supply ultra low power micro controllers and sensors (Pantelopoulous and Pantelopoulous, 2010). Applied to sensor nodes, energy from external sources can be harvested to power the nodes and in turn, increase their lifetime and capability (Sudevalayam and Kulkarni, 2011).

In an EHS, the incoming energy is managed using two points of view.

- Harvest-store-use (HSU) where harvested energy can be accumulated for future use. It requires an energy storage device or battery along with an appropriate management of the stored energy.
- Harvest-use (HU) where energy can not be stored and it is used immediately.

Focusing on the HSU architecture and depending on the capacity of the storage device, finite and infinite capacity of the battery can be considered.

- Infinite capacity implies that the incoming energy can always be saved in the battery. It can be, for example, devices with a very large energy storage capacity compared to the rate of the harvested energy flow (Ozel and Ulukus 2012).
- HSU with finite capacity of the storage battery implies that the harvested energy can be stored up to a limit. Beyond this limit, incoming energy can not be stored and would be lost (Tutuncuoglu and Aylin, 2012).

In this paper, an EHS system with a HSU architecture is analyzed. The EHS system contains

- (a) An antenna that gathers energy from the environment.
- (b) A battery with finite capacity to store the harvested energy.
- (c) A device powered by the battery. The device received energy from the battery at certain times called *feeding times*.

The harvested energy can be treated as a stochastic process due to the random nature of the energy source. Common stochastic processes in the literature to describe the harvested energy include Poisson processes (Yang *et. al* 2016) and compound Poisson processes (Khuzani and Mitran, 2014). However, in this paper, the process used to model the harvested energy by the EHS is the gamma process. Gamma process is a versatile stochastic process used in many engineering fields to model, for example, the system degradation over the time, the water storage by dams and the aggregate claims process in classical collective risk theory (van Noortwijk, 2009). From a theoretical point of view, a gamma process can be thought as a compound Poisson process with Poisson rate tends to infinity and increment sizes tend to zero in proportion. The definition of gamma process is first recalled.

DEFINITION 1. We say that $\{X(t), t \geq 0\}$ is a homogeneous gamma process with parameters α and β if it fulfills the following properties.

- $P(X(0) = 0) = 1$
- Increments $X(t+s) - X(t)$ are independent for all t and s .
- Increments $X(t+s) - X(s)$ follow a gamma distribution with parameters αt and β .

The technical justification for the use of a gamma process to model the harvested energy for a Radio Frequency Energy Harvesting System is described in detail in Castro *et. al* (2018).

2. System Description

We assume a harvesting system which contains an antenna and a battery to store the harvested energy. The device is powered by the battery at certain times called *feeding times*. Let $T_1, T_2, \dots$, the successive feeding times.

The system functioning is described as follows.

(a) Let $E_H(t)$ be the harvested energy by the system at time t . We assume that $\{E_H(t), t \geq 0\}$ follows an homogeneous gamma process with parameters α and β . For $s < t$, we get that the probability distribution function of the harvested energy in the interval $(s, s+t)$, $E_H(s, s+t)$, is given by

$$f_{E_H(s, s+t)}(x) = \frac{\beta^{\alpha t}}{\Gamma(\alpha t)} x^{\alpha t - 1} \exp(-\beta x), \quad x > 0, \quad (1)$$

where $\Gamma(\cdot)$ denotes the gamma function given by

$$\Gamma(\alpha t) = \int_0^\infty x^{\alpha t - 1} e^{-x} dx. \quad (2)$$

(b) Let $E_B(t)$ be the stored energy in the battery at time t with $E_B(0) = E_0$ where E_0 denotes the initial stored energy. There is a *safety threshold* in the battery that corresponds to the minimum energy reserve of the battery that can not be consumed. It means that the level of stored energy must always exceed a threshold level E_M for security reasons with $E_0 > E_M$.

(c) The capacity of the battery is finite and equals to E_K hence the stored energy in the battery at time t is bounded

$$E_M \leq E_B(t) \leq E_K, \quad \forall t. \quad (3)$$

It means that, when the level of energy in the battery reaches the threshold E_K , the incoming energy is overflowed.

(d) The device can be in two modes: ON and OFF depending on the feeding process.

- In mode ON, the device is operational and powered by the battery. It consumes E_C energy units in each feeding time.
- In mode OFF, the device is turned off and the feeding process of the device stops.

(e) Let $E_B(T_n^-)$ be the stored energy in the battery right before the feeding time T_n for $n = 1, 2, \dots$

- If $E_B(T_n^-) - E_C \geq E_M$, the device is properly powered and a *feasible feeding task* is performed at time T_n .
- If $E_B(T_n^-) - E_C \leq E_M$, the feeding of the device starts but, when the system detects that the stored energy is going to drop below E_M , the feeding process stops. In this case, the device is not properly powered and it is called an *unfeasible feeding task*. Right after an unfeasible feeding task, the energy stored in the battery is equal to the threshold E_M and the device enters in OFF mode and the system keeps on harvesting energy. When the stored energy reaches the level E_0 , the device enters in ON mode and the feeding process starts again.

(f) Feeding times are periodic, that is, $T_i - T_{i-1} = T$ and $T > 0$ for $i = 1, 2, \dots$

Let $\{E_H^{(i)}(t), t \geq 0\}$ replicas of the process $\{E_H(t), t \geq 0\}$ for $i = 1, 2, \dots$. Starting from 0, the system evolves as follows

$$E_B(t) = E_H^{(1)}(t) + E_B(0), \quad 0 \leq t < T.$$

The harvested energy in $[0, T]$ is noted by $E_H^{(1)}(T)$ with density given by (1). Right before T , the energy stored in the battery is equal to

$$E_B(T^-) = \min \left(E_H^{(1)}(T) + E_B(0), E_K \right).$$

At time T ,

- If $E_B(T^-) - E_C \geq E_M$, the feeding task is performed properly and the device remains in mode ON. The

stored energy in the battery at time T^+ is equal to

$$E_B(T^+) = E_B(T^-) - E_C.$$

• If $E_B(T^-) - E_C < E_M$, the feeding task is not performed properly (*unfeasible*) and the device enters in mode ON. The stored energy in the battery at time T^+ is equal to

$$E_B(T^+) = E_M.$$

Hence, the stored energy in the battery at time T^+ is equal to

$$E_B(T^+) = \max(E_B(T^-) - E_C, E_M).$$

In a general setting, let $E_B(iT^-)$ be the stored energy in the battery right before a feeding time iT .

• If $E_B(iT^-) - E_C \geq E_M$, the device is properly powered and it is called a *feasible feeding task* and

$$E_B(iT^+) = E_B(iT^-) - E_C.$$

• If $E_B(iT^-) - E_C < E_M$, the feeding of the device starts but, when the system detects that the stored energy is going to drop below E_M , the feeding process is stopped and

$$E_B(iT^+) = E_M.$$

Hence

$$E_B(iT^+) = \max(E_B(iT^-) - E_C, E_M).$$

Notice that $E_M = 0$ implies that, as long as there is energy in the battery, the device remains in mode ON.

Let

$$V_i = \{E_B(iT^-) - E_C > E_M\}, \quad (4)$$

for $i = 1, 2, \dots$ that corresponds to a feasible feeding task at time iT .

Considering an finite horizon time T^* , the probability that the device is in mode ON from 0 to T^* is equal to

$$\mathbb{P}\left(\bigcap_{i=1}^{n^*} V_i\right), \quad \text{with } n^* = \lfloor \frac{T^*}{T} \rfloor,$$

where V_i are given by (4) Whenever (3) is fulfilled in $[0, T^*]$, the total energy consumed by the device due to the feeding tasks at time T^* , $E_D(T^*)$, is equal to

$$E_D(T^*) = \lfloor T^*/T \rfloor E_C.$$

Figure 1 shows the evolution of the energy stored in the battery. Blue arrows corresponds to feasible feeding tasks. Red arrows correspond to unfeasible feeding task, that is, the energy stored in the battery was not enough to feed the device under the safety constraint given by (3).

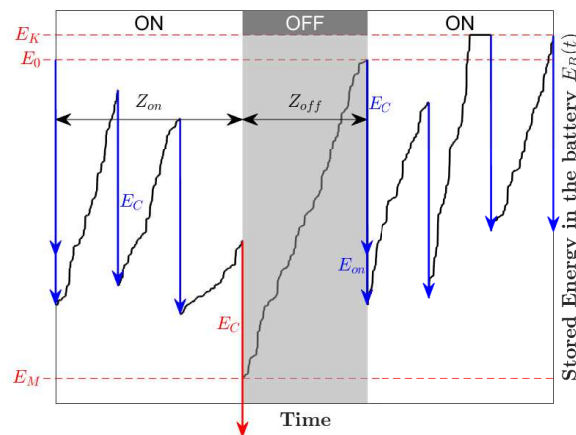


Figure1. Evolution of the energy stored in the battery due to the operation of the EHS

3. Management policy of the EHS

Renewal techniques are used in this work to describe management policies of this system. We recall that a renewal process is an arrival process in which the inter-arrival intervals are independent and identically distributed random variables. Applying the renewal process to this system, a renewal cycle takes place every time that the period OFF ends.

After the switching on at time $t = 0$, the energy stored in the battery is E_0 . The device remains in mode ON until the first unfeasible feeding time. After this unfeasible feeding time, the device enters in model OFF and the system keeps harvesting energy until the energy stored in the battery exceeds the level E_0 , time in which the system again enters in ON mode (see Figure 1).

Let $n_{uf}T$ be the first unfeasible feeding time, that is,

$$n_{uf} = \inf \{i \geq 0, E_B(iT^-) - E_c < E_M\}. \quad (5)$$

Starting from $t = 0$, at time $n_{uf}T$, the energy stored in the battery is equal to

$$\mathbb{E}[n_{uf}T^+] = E_M.$$

After this unfeasible feeding task, the device remains in OFF mode until the energy stored in the battery reaches the level E_0 . Then Z_{off} follows the same distribution as $\sigma_{E_0-E_M}$ where $\sigma_{E_0-E_M}$ denotes the first hitting time to the level $E_0 - E_M$ of an homogeneous gamma process with parameters α and β . That is,

$$\sigma_{E_0-E_M} = \inf \{t \geq 0, X(t) \geq E_0 - E_M\},$$

where $X(t)$ denotes an homogeneous gamma process with parameters α and β . The distribution function of Z_{off} is given by

$$\begin{aligned} F_{\sigma_{E_0-E_M}}(t) &= P[\sigma_{E_0-E_M} \leq t] \\ &= P[X(t) \geq E_0 - E_M] \\ &= \int_{E_0-E_M}^{\infty} f_{E_H(0,t)}(x) dx \\ &= \frac{\Gamma(\alpha t, (E_0 - E_M)\beta)}{\Gamma(\alpha t)}, \end{aligned}$$

where $\Gamma(\alpha t)$ is given by (2) and $\Gamma(\alpha t, (E_0 - E_M)\beta)$ denotes the incomplete gamma function given by

$$\Gamma(\alpha t, \beta(E_0 - E_M)) = \int_{\frac{\beta(E_0-E_M)}{\alpha}}^{\infty} z^{\alpha t-1} e^{-z} dz. \quad (6)$$

Due to the jump discontinuities of the sample paths of a gamma process, the energy stored in the battery in the period OFF fulfils

$$E_B(Z_{off}) \geq E_0 - E_M,$$

hence the overshoot

$$E_B(Z_{on} + Z_{off}) - E_0,$$

is positive almost surely. To avoid complicated mathematical technicality, in this paper this overshoot is disregarded what it implies that the energy stored in the battery right after the end of the period OFF is equal to E_0 , that is,

$$E_B(Z_{on} + Z_{off}) = E_0.$$

Let $R_1, R_2, \dots$, be the time between successive renewals of the system. Random variables R_i are independent and identically distributed with $R_i \equiv_d R$ where R is given by

$$R = Z_{on} + Z_{off} = n_{uf}T + Z_{off} \quad (7)$$

where $n_{uf}T$ is given by (5).

The optimal management policy corresponds to the value of T that maximises the expected number of

feasible feeding tasks per unit time in a renewal cycle. As objective function, the ratio

$$T_f^{-1} = \frac{\mathbb{E}[n_{uf}] - 1}{\mathbb{E}[R]},$$

where R and n_{uf} are given by (7) and (5) respectively is chosen. The problem of finding the optimal management policy is formulated as

$$\max_{T>0} \left(\frac{\mathbb{E}[n_{uf}] - 1}{\mathbb{E}[R]} \right).$$

4. Conclusions and final remarks

In this paper, the management policy for an EHS system is analyzed. The main novelty of this paper is the use of the gamma process to model the harvested energy in a system. Although complex stochastic processes have been used to describe the harvested energy, as far as we are concerned, it is the first time in which the gamma process is used in the context of harvesting energy. Due to the properties of the gamma process, the model is analytically tractable and useful to provide insights for solving some theoretical management policies. Although the technical justification of the use of gamma process has been described by Castro *et. al* (2018), the model has not been justified with empirical measurements.

For an EHS system with infinite capacity of storage (that is, $E_K = \infty$), the problem exposed in this paper can be seen as a system with degradation modeled using a gamma process and subject to periodic imperfect repairs. These imperfect repairs decreases the degradation by a deterministic and constant amount whenever the degradation of the system exceeds a fixed threshold.

Two important aspects considered in this paper is that the battery does not suffer degradation (that is, battery can save in a proper way the energy that receives) and also that the energy demand is periodic (that is, feeding times are periodic). However, future works can consider extensions of this work where the inefficiency of the storage can be considered and also the demand of the energy by the device can be continuous.

5. Acknowledgements

For the first author, this research was supported by Ministerio de Economía y Competitividad, Spain (Project MTM2015-63978-P), Gobierno de Extremadura, Spain (Project GR15106), and European Union (European Regional Development Funds). For the second and third author, this research was supported by Gobierno de Extremadura, European Union (European Regional Development Funds) and SferaOne (Project GlobalEnergy AA-16-0124-2) and by Ministerio de Economía y Competitividad, Spain (Project TEC2017-85376-C2-X-R). For the third author this research was supported also by Spanish Ministerio de Educación, Cultura y Deporte (FPU 00022/15).

References

- Pantelopoulos A, Pantelopoulos NG (2010) A survey on wearable sensor-based systems for health monitoring and prognosis. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(1), 1-12.
- Sudevalayam S, Kulkarni P (2011) Energy harvesting sensor nodes: survey and implications *IEEE Communications Surveys & Tutorials*, 13(3), 443-461.
- Ozel O, Ulukus S (2012) Achieving AWGN capacity under stochastic energy harvesting *IEEE Transactions on Information Theory*, 58(10) 6471–6483.
- Tutuncuoglu K, Aylin Y (2012) Optimum transmission policies for battery limited energy harvesting nodes *IEEE Transactions on Wireless Communications*, 11(3), 1180-1189.
- Yang J, Wu X, Wu J. (2016) Optimal online sensing scheduling for energy harvesting sensors with infinite and finite batteries *IEEE Journal on Selected Areas in Communications*, 34(5) (2016), 1578–1589.
- Khuzani MB, Mitran P (2014) On online energy harvesting in multiple access communication systems *IEEE Transactions on Information Theory*, 60(3) (2014), 1883–1898.

- Van Noortwijk JM (2009) A survey of the application of gamma processes in maintenance,. *Reliability Engineering & System Safety*, 94(1), 2-21.
- Gani J (1957) Problems in the probability theory of storage systems *Journal of the Royal Statistical Society. Series B (Methodological)*, 19(2), 181-206.
- Castro IT, Landesa L, Serna A. (2018) The role of the gamma process in RF harvesting systems Paper submitted to *IEEE Transactions on Systems, Man, and Cybernetics: Systems*.
- Huynh KT, Castro IT, Barros A, Bérenguer C. (2012) Modeling age-based maintenance strategies with minimal repairs for systems subject to competing failure modes due to degradation and shocks *European Journal of Operational Research*, 218, 140-151.

Multicriteria Model to Support A Definition of a Maintenance - Spare Parts Policy

P. P. L. Dantas*, T. F. O. Rodriguez and C. A. V. Cavalcante*****

Universidade Federal de Pernambuco - UFPE

Av. da Arquitetura - Cidade Universitária, Recife - PE, 50740-550- Recife – PE

*priscillapld@hotmail.com, ** tuliofor@gmail.com, ***cristianogesm@gmail.com

Abstract. The maintenance became one of the primaries factors in guaranteeing the proper functioning of a productive system. Despite this, some assumptions assumed by maintenance planning are not fully complied with, such as the availability of spare parts whenever its need. Hence, the integrated models of inventory maintenance and management came out, used to seek the levels that best fit for the two policies. Integration is usually based on the criteria of total costs, but not always this factor is the only interest of the decision maker. Therefore, the present work aims at the construction of a multicriteria model to support an application of an integrated model developed in a previously work. New criteria have been introduced as choice dimensions and an outranking method was developed using PROMETHEE II to generate an order of values that reflects the interests of the decision maker.

1. Introduction

Maintenance is an activity that seeks to ensure the full functioning of the components and, consequently, the achievement of quality. Therefore, through maintenance policies, it is sought to reduce the impact of machine breaks in production (Chen et. al. 2006).

Traditionally, maintenance policies only consider repair or replacement as a consequence of the failure occurrence. That is the reason why the existence of spare parts in stock is a premise (Yoo et al 2001). In practice, the planning and execution of the maintenance model adopted in the production unit directly affects the availability of spare parts in the inventory. Thus, if the level of spare parts was below the quantity requested to make the replacements, the reliability of the system would be affected. On the other hand, if the level was over dimensioned, the maintenance costs of the stock would be high (Rodriguez et al 2016).

For Horenbeek et al. (2013), the joining of maintenance and spare parts takes into account the trade-off between maintenance and stock policies. Authors such as Panagiotidou (2013) and Horenbeek et al. (2013), among others, has studied policies that integrate maintenance planning with the planning of spare parts. However, a fact observed when studying integrated maintenance and inventory policies is what the great majority takes into account only the cost criterion to justify such integration.

According to Jajimoggalla and Beela (2012), some criteria may interfere in decision making regarding spare parts, such as costs, availability and demand. Moreover, according to Triantaphy et al. (1997), maintenance can take into account criteria such as availability, reliability and maintainability. Thus, it is possible to perceive that cost is not necessarily the only criterion that should influence the integration of a maintenance policy with the management of spare parts, since the integration involves two plans that can already be influenced by several criteria.

For Xie and Wang (2008) there is no optimal policy if there is a separate or sequential treatment between dependent models. Therefore, conventional policies are not efficient when they are treated in isolation, being necessary, for a better representation of real problems, an approach that seeks integration between the areas of interest (Panagiotidou, 2013).

As a result of this, authors such as Sarker et al (2000), Armstrong and Atkins (1996), Kabir and Al-Olayan (1996), Elwany and Gebrael (2007), Xie and Wang (2008), among others, developed models with only one criterion to be achieved to justify the integration. Horenbeek et al. (2013) surveyed the main types of integrated models, which in essence vary according to the premises and conditions established.

Authors such as Panagiotidou (2013) and Rodriguez et al. (2016) integrated the maintenance planning with the management of spare parts defining as the only criterion to be reached the service cost and, thus, find its best alternative using mathematical methods, which are based on the search of the compromise solution of this criterion. However, there is a barrier in optimizing the problem and finding the best alternative when the

interest is in more than one criterion. An alternative to this approach is the use of multicriteria models, whose approach is used in some policies of maintenance and management of spares acting in isolated way, as can be seen in Ferreira et al. (2009), Almeida (2001), Wheel et al. (2014), Jajimoggala and Beela (2012), and Shafiee (2015), among others.

Therefore, since an integrated policy is the combination of several models, it can provides a large number of important criteria for the decision maker who wants a comprehensive view on the problem faced. In this sense, it is observed the relevance of the multicriteria approach to support the decision in the context of an integrated maintenance policy and stock of spare parts.

This work has the purpose of using multiple objectives to support the use of integrated policy. According to Almeida et al (2015) it is possible to affirm that multiobjective problems can be approached using MCDM (Multicriteria Decision Models) or Multiobjective Optimization. For this, multicriteria modeling by offering an interaction with the preferences of a decision maker and including its judgment of value was seen as more advantageous for the situation. Since such preferences are obtained from subjective evaluations from the decision maker and they are inherent to reality, thus approaching the modeling of the real problem in order to meet the following objectives: efficiently use resources, increase system availability and reliability, and increase productivity.

2. Structuring and application of the model

The construction of the multicriteria model was based on the integrated policy of maintenance and spare parts developed by Rodriguez (2016), which developed it with focus only as a criterion the total costs. The multicriteria model seeks to determine the best values of the decision variables that determine the option that best answers the preferences of the decision maker. In Figure 1 (Almeida et al. 2015) describes a methodology based on twelve steps presented for the creation of a multicriteria model, starting from the simplest, such as the definition of criteria, to defining the type of problem and method to be used, and the possibility of returning to an earlier stage as a form of refinement and improvement.

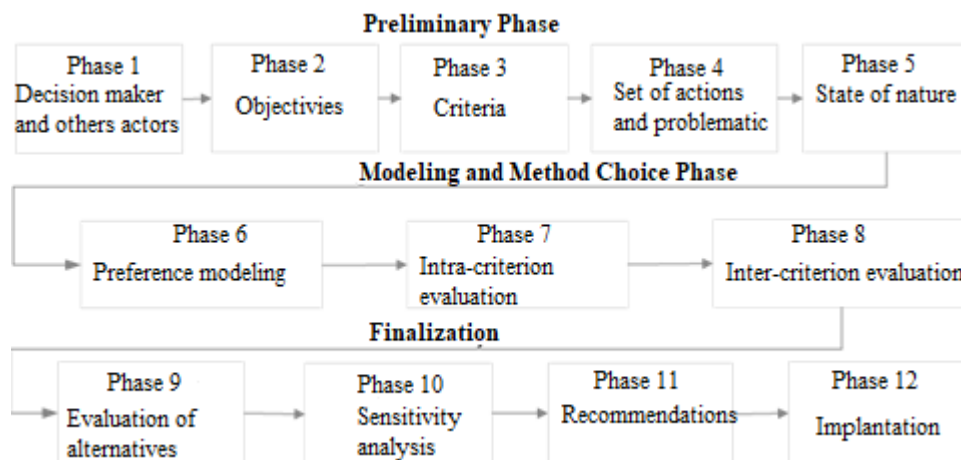


Figure 1: Methodology for the creation of the Multicriteria model (Adapted: Almeida et al 2015)

3. Multicriteria model

Based on the methodology presented by Almeida et al. (2015) for the determination of a multicriteria model, a proposal to support the use and definition of an integrated maintenance and inventory management policy was developed.

3.1. Preliminary phase

The core of the preliminary phase is the determination of possible objectives and evaluation criteria. In this case, the objectives to use the integrated model of the policies presented above can be translated as increasing the effectiveness and efficiency of the maintenance activity and the management of spare parts acting in a synchronous way. So that it is possible to obtain reduction of costs, attend the demand; reduce the time the system is without spare parts in stock, in addition to increasing preventive maintenance by maximizing the

useful life of the component. Finally, the decision maker seeks to increase the average life of the system to ensure greater system uptime and uninterrupted operation. From the definition of the objectives of the decision-maker, the criteria for measuring the performance of this objective are defined (Keeney 1992). With this, it is possible to associate nine criteria or attributes that can represent them (Table 3.1).

C ₁ , Cost-life relation	C ₄ , System vulnerability	C ₇ , Preventive maintenance rate
C ₂ , Loss due to unavailability	C ₅ , Average number of times the stock remains zero	C ₈ , Average system life
C ₃ , Average number of breaks due to lack of inventory	C ₆ , Corrective maintenance rate	C ₉ , Reliability of maintenance policy

Table 3.1: Criteria for the multicriteria model

The multicriteria problem, in this case, is to find values that best satisfy the needs of the decision maker. Therefore, because we believe that an order of possible alternatives must be made, in view of the presented options, the ordering problem is the one present in the multicriteria model in question. In addition, believing that there is no compensatory relationship between the criteria, that is, the performance of one criterion does not interfere or compensate for the other, since, despite being related, the impacts caused in the decision are not compensatory, the model developed is not-compensatory, being this characteristic conserved in the outranking methods.

3.1.1. Extension of the Integrated Model to Obtain Missing Criteria

The multicriteria method requires that alternatives be elaborated, and for this purpose were generated simulated values from the integrated model of Piggyback and inventory management by Rodriguez (2016). Thus, was developed an extension for the modeling and simulation of the integrated model using computational programming, configuring the integrated model to obtain the values to the alternatives of the criteria that the decision maker flagged as relevant.

Therefore, from the development of the integrated model presented, it was possible to simulate the values for the alternatives and criteria, creating the necessary environment for the use of the multicriteria method.

3.2. Modeling and method choice phase

The second phase, preference modeling and method choice, is where the definition of the multicriteria method is performed, which is the most flexible of all, and its three steps must be performed simultaneously to give a better understanding of the situation.

Therefore, because it is a non-compensatory relationship, the intra-criterion evaluation occurs in an ordinal way, that is, ordering values as the importance. The inter-criterion evaluation is given by the determination of weights for each criterion, as seen in Table 3.2. The sum of the values given to the weights must be equal to 1, and, in agreement with the decision maker.

C ₁	C ₂	C ₃	C ₄	C ₅	C ₆	C ₇	C ₈	C ₉
0,20	0,18	0,08	0,18	0,08	0,05	0,05	0,10	0,08

Table 3.2: Weight of the criteria

The predominant method for the development of the model is the PROMETHEE II, which is used for ordering problems, case presented here, and also establishes a complete pre-order, not assuming the relation of incomparability.

By obtaining the values of each criterion through the simulation it was possible to construct the consequence matrix, which presents the values of each criterion within each alternative. A set of 56 alternatives were established, which are presented in the format (T, s, S), which represents, respectively, the time interval between maintenance, minimum stock and maximum stock. From the consequence matrix and the weights it was possible to apply the PROMETHEE II multicriteria method to generate the order of alternatives.

3.3. Finalization

The finalization phase is when the model is effectively applied in the search of the result. Thus, it is possible to perform a sensitivity analysis and perhaps retrace steps for changes and improvements.

It was possible to ordinate the values that meet the criteria of the decision maker, from the net flow created by PROMETHEE II, as can be seen in Table 3.4. According to the ranking created to order the alternatives, the option presented in the first place is (75, 2, 4), that is, a maintenance interval of B of 75 thousand copies, with minimum inventory, equal to 2, and maximum stock, equal to 4. Being this the scenario that meets the objectives of the decision maker according to the degree of importance given to each criterion, that is, this is the best result.

Nº	Alternative	ϕ	ϕ^+	ϕ^-	Nº	Alternative	ϕ	ϕ^+	ϕ^-
1	(75, 2, 4)	0,6353	0,7340	0,0987	29	(45, 6, 8)	0,1198	0,4564	0,3365
2	(90, 2, 4)	0,6051	0,7558	0,1507	30	(45, 4, 6)	0,1038	0,4495	0,3456
3	(60, 2, 4)	0,4805	0,6940	0,2135	31	(45, 8, 10)	0,1020	0,4493	0,3473
4	(90, 8, 10)	0,4751	0,6347	0,1596	32	(45, 3, 5)	0,0758	0,4460	0,3702
5	(90, 5, 7)	0,4422	0,6144	0,1722	33	(30, 1, 3)	0,0635	0,4885	0,4251
6	(90, 1, 3)	0,4242	0,6742	0,2500	34	(45, 7, 9)	-0,0664	0,3776	0,4440
7	(90, 4, 5)	0,3822	0,5816	0,1995	35	(30, 8, 10)	-0,0813	0,3731	0,4544
8	(75, 6, 8)	0,3767	0,5758	0,1991	36	(30, 4, 6)	-0,0831	0,3684	0,4515
9	(45, 2, 4)	0,3718	0,6433	0,2715	37	(30, 6, 8)	-0,1200	0,3436	0,4636
10	(75, 8, 10)	0,3713	0,5822	0,2109	38	(30, 5, 7)	-0,1465	0,3267	0,4733
11	(75, 3, 5)	0,3691	0,5824	0,2133	39	(30, 3, 5)	-0,2265	0,3267	0,5533
12	(90, 6, 8)	0,3671	0,5745	0,2075	40	(30, 7, 9)	-0,2389	0,3184	0,5573
13	(90, 7, 9)	0,3644	0,5969	0,2325	41	(15, 1, 3)	-0,3429	0,2922	0,6351
14	(75, 4, 6)	0,3516	0,5718	0,2202	42	(15, 2, 4)	-0,4464	0,2573	0,7036
15	(75, 1, 3)	0,3476	0,6311	0,2835	43	(15, 5, 7)	-0,4664	0,2344	0,7007
16	(75, 7, 9)	0,3415	0,5578	0,2164	44	(15, 4, 6)	-0,5127	0,2118	0,7245
17	(60, 1, 3)	0,3029	0,6140	0,3111	45	(15, 3, 5)	-0,5129	0,2145	0,7275
18	(60, 7, 9)	0,3004	0,5471	0,2467	46	(15, 8, 10)	-0,5193	0,2075	0,7267
19	(90, 3, 5)	0,2969	0,5495	0,2525	47	(15, 6, 8)	-0,5247	0,2056	0,7304
20	(30, 2, 4)	0,2965	0,6118	0,3153	48	(15, 7, 9)	-0,5365	0,2025	0,7391
21	(75, 5, 7)	0,2869	0,5369	0,2500	49	(0, 4, 6)	-0,6080	0,1729	0,7809
22	(60, 4, 6)	0,2602	0,5235	0,2633	50	(0, 2, 4)	-0,6240	0,1753	0,7993
23	(45, 1, 3)	0,1995	0,5629	0,3635	51	(0, 8, 10)	-0,6262	0,1638	0,7900
24	(60, 6, 8)	0,1956	0,4884	0,2927	52	(0, 6, 8)	-0,6275	0,1645	0,7920
25	(60, 5, 7)	0,1800	0,5255	0,3455	53	(0, 1, 3)	-0,6296	0,1691	0,7987
26	(60, 3, 5)	0,1611	0,4722	0,3111	54	(0, 5, 7)	-0,6395	0,1567	0,7962
27	(45, 5, 7)	0,1587	0,4784	0,3196	55	(0, 3, 5)	-0,6618	0,1469	0,8087
28	(60, 8, 10)	0,1236	0,4667	0,3431	56	(0, 7, 9)	-0,6918	0,1373	0,8291

Table 3.4: Order obtained by the net flow of PROMETHEE II

From the application of the outranking method it was possible to verify that the values differ from the one suggested by the integrated model of Rodriguez (2016). Since in that case evaluated only one criterion was taken into account and a pre-established value was used for the maximum limit of stock, S. With this it was verified that the maintenance interval found is greater and the quantities of spare parts in stock also.

3.4. Sensitivity analysis

For the analysis of sensitivity or robustness IX scenarios were elaborated, in which the weights of the criteria was altered. In each scenario, the focus was an evaluation criterion, either in the increase or decrease of its value, and the values of the other criteria respected proportionality. It is worth noting that in all cases there

were few variations in the order of all alternatives, so it is possible to prove that the model converges to the result obtained initially, showing its robustness.

4. Conclusions

The present work dealt with a multicriteria model to support the use of an integrated policy of maintenance Piggyback and management of spares of continuous revision developed by Rodriguez (2016).

The determination of the values of each criterion was made from modeling and simulation in a computational software, as a search for an approximation of reality. It was based on the integrated model developed by (Rodriguez, 2016) to construct an algorithm that determined the values within certain intervals. With this activity, it was possible to construct an consequences matrix to obtain the multicriteria model and to find the values that best met the needs of the decision maker.

As a result, we obtained a configuration different from that found in the integrated model of Rodriguez (2016), showing that evaluating a problem from one perspective alone is not always enough to make the best decision when there is a need to meet other objectives. In the case studied here, an optimal solution was not found, but rather an order of the possible values that the variables can be assumed to meet all the desired objectives, some of them conflicting, for the decision maker.

It was also observed, from the construction of the values of each criterion through the simulation, that the maintenance interval in all the simulated scenarios was one of the most rigid, and the variables referring to the minimum and maximum stock were intermediate values, thus balancing the results objectified by the decision maker. It is worth noting that the present work represents an opportunity to view the integrated models in a different way, through the multi-criteria approach.

References

- ALMEIDA, A. T. Processo de decisão nas organizações: construindo modelos de decisão multicritério. - São Paulo: Atlas, 2013.
- ALMEIDA, A. T. Multicriteria decision making on maintenance: Spares and contracts planning. *European Journal of Operational Research* 129, 2001. – pp. 235 – 241.
- ALMEIDA, A. T.; CAVALCANTE, C. A. V.; ALENCAR, M. H.; FERREIRA, R. J. P.; ALMEIDA FILHO, A. T. de; GARCEZ, T. V. Multicriteria and Multiobjective Models for Risk, Reliability and Maintenance Decision Analysis; *International Series in Operations Research & Management Science* vol 231. 1. ed. New York: Springer, 2015. v. 1. 416p.
- ARMSTRONG, M. J.; ATKINS, D.A. Joint optimization of maintenance and inventory policies for a simple system. *IIE transactions* 28, 1996. – pp. 415–424.
- CHEN, M. C.; HSU, C. M.; CHEN, S. W. Optimizing joint maintenance and stock provisioning policy for a multi-echelon spare part logistics network. *Journal of the Chinese Institute of Industrial Engineers* 23 (4), 2006. – pp. 289–302.
- ELWANY, A.; GEBRAEEL, N., 2008. Sensor-driven prognostic models for equipment replacement and spare parts inventory. *IIE Transactions* 40 (7), 2008. – pp. 629–639.
- FERREIRA, R. J. P.; ALMEIDA, A. T.; CAVALCANTE, C. A. V. A multi-criteria decision model to determine inspection intervals of condition monitoring based on delay time analysis. *Reliability Engineering and System Safety* 94, 2009. – pp. 905–912.
- JAJIMOGGALA, S.; BEELA, K. R. S. Spare parts criticality evaluation using hybrid multiple criteria decision making technique. *Int. J. Information and Decision Sciences*. 2012. Vol 4, No. 4. – pp. 350–370.
- KABIR, A. B. M. Z., & AL-OLAYAN, A. S. A stocking policy for spare part provisioning under age based preventive replacement. *European Journal of Operational Research*, 1996. 90(1), - pp 171–181.
- PANAGIOTIDOU, S. Joint optimization of spare parts ordering and maintenance policies for multiple identical items subject to silent failures. *European Journal of Operational Research*, 2013.
- RODA, I.; MACCHI, M.; FUMAGALLI, L.; VIVEROS, P. A review of multi-criteria classification of spare parts. *Journal of Manufacturing Technology Management*, 2014. Vol. 25 Iss 4 pp. 528 – 549.
- RODRIGUEZ, T. O. Contribuições ao modelo de manutenção preventiva “Piggyback” pela abordagem multicritério e de gestão de sobressalentes. Recife, 2016. 98p. (Mestrado – Universidade Federal de Pernambuco/UFPE)
- RODRIGUEZ, T. F. O. ; RIBEIRO, L. F. A. ; Cavalcante, Cristiano A.V.. Joint optimization of maintenance preventive Piggyback and spare parts policies. In: 9th IMA International Conference on

Modelling in Industrial Maintenance and Reliability (MIMAR), 2016, London. 9th IMA International Conference on Modelling in Industrial Maintenance and Reliability (MIMAR), 2016.

SARKER, R.; HAQUE, A. Optimization of maintenance and spare provisioning policy using simulation. *Applied Mathematical Modelling* 24 (10), 2000. – pp 751–760.

SHAFIEE, M. Maintenance strategy selection problem: an MCDM overview. *Journal of Quality in Maintenance Engineering*, 2015. Vol. 21 Iss 4 pp. -

TRIANAPHYLLOU, E.; KOVALERCHUK, B.; MANN, L.; KNAPP, G. M. Determining the most important criteria in maintenance decision making, *Journal of Quality in Maintenance Engineering*, 1997. Vol. 3 Iss 1 pp. 16 - 28

VAN HORENBEEK. A.; BURÉ, J.; CATTRYSSSE, D.; PINTELON, L.; VANSTEENWEGEN, P. Joint maintenance and inventory optimization systems: A review. *International Journal Production Economics*. - 2013. - 2 : Vol. 143. - pp. 446-508.

XIE, J.; WANG, H. Joint Optimization of Condition-Based Preventive Maintenance and Spare Ordering Policy. Proceedings of the 4th International Conference on Wireless Communications, Networking and Mobile Computing. *IEEE*, 2008. - pp. 1–5.

YOO, Y.K.; KIM, K.-J., SEO, J. Optimal joint spare stocking and block replacement policy (cost modelling of spare stocking and block replacement). *International Journal of Advanced Manufacturing Technology* 18 (12), 2001. – pp. 906–909.

Remaining Useful Life estimation methods for predictive maintenance models: defining intervals and strategies for incomplete data

A. Delmas*, M. Sallak*, W. Schön* and L. Zhao**

* Laboratoire Heudiasyc, Sorbonne Universités, Université de Technologie de Compiègne, CNRS, Heudiasyc UMR 7253, CS 60319, 60203 Compiègne Cedex, France {adrien.delmas, mohamed.sallak, walter.schon}@utc.fr

** SECTOR group, 12 avenue du Québec BP 636 Villebon sur Yvette, 91965 Courtaboeuf cedex, France linda.zhao@sector-group.net

Abstract. Predictive maintenance allows the drawbacks of corrective and preventive maintenance to be overcome by estimating the Remaining Useful Life (RUL) of components. This value helps undertake required tasks at the right time: when the component is deteriorated but right before a failure occurs.

We propose an imprecise RUL estimation method with the use of Hidden Markov Models (HMM). HMMs have been used for prognosis in several studies. Here, we fit the training data by a polynomial and then we frame it with 2 other polynomials of the same degree. This results in 3 distinct trained HMMs: the first is used for the RUL estimation and the other 2 provide the lower and higher bounds. Finally, we propose a strategy that can be adopted when dealing with incomplete or imprecise data. This work has been applied to the IEEE 2008 PHM challenge which was a competition aiming at evaluating prognosis methods.

1. Introduction

Maintenance is an essential step in the life cycle of every complex system. The European standard BS EN 13306 (BSI, 2010) defines maintenance as the *combination of all technical, administrative and managerial actions during the life cycle of an item intended to retain it in, or restore it to, a state in which it can perform the required function*. In other words, maintenance allow ensuring the nominal operating state of the system in conditions that guarantee the safety of the user and the environment, along with its availability. However, it is usually pretty expensive and the system is not available during maintenance tasks. Moreover, current habits are hardly ever optimal:

- Corrective maintenance involves waiting for a failure before repairing or replacing the affected components
- Preventive maintenance involves replacing the components at regular intervals, according to a fixed planning

The former strategy usually induces additional costs (Lee, Ni, Djurdjanovic, Qiu, & Liao, 2006) and an important unavailability of the system (Palem, 2013). Besides, a failure can raise safety issues as the systems can no longer perform all or part of the functions it was designed for. Preventive maintenance reduces significantly the risk of unexpected failures as well as the downtimes but can cause high maintenance costs because each operation is performed without knowing the actual state of the system. Therefore, some components are replaced whereas they are still in a nominal operating state or slightly degraded (Lee, Ni, Djurdjanovic, Qiu, & Liao, 2006) (Le, 2016).

In order to overcome these drawbacks, a new form of maintenance has emerged: predictive maintenance (Palem, 2013). It consists of a regular surveillance of the organs of the system (Bartelds, et al., 2004), which allow evaluating their health state and their proper functioning. It is then possible to forecast the appearance of a system failure before it occurs (Lee, Ni, Djurdjanovic, Qiu, & Liao, 2006) (Le, 2016) (Jardine, Lin, & Banjevic, 2006). Consequently, the adequate maintenance tasks can be planned at just the right time (Mercier & Pham, 2012).

The interest in prognosis has been growing for a few decades and many solutions have been developed, tested and improved over the years. These methods can be classified into 3 main categories (Le, 2016) (Vachtsevanos, Lewis, Roemer, Hess, & Wu, 2006) (Soualhi, Razik, Clerc, & Doan, 2014) (Roemer, Byington, Kacprzynski, & Vachtsevanos, 2006):

- *Physical model-based prognosis* that rely on mathematical or physical models of the degradation phenomenon.

- *Evolutionary or trending models* that are based on temporal data sets and indicators that can describe the evolution of the health state.
- *Experience-based prognosis* which is based on the experience and the expert knowledge learned, for example, on the history of failures or degradation of the system

Data driven models allow performing prognosis from data sets produced by means of surveillance. In particular, stochastic models are of interest for RUL estimation because their implementation is quite simple and understandable. Among them, Markov process are dynamic and multi-state models that can model a degradation pretty faithfully. Moreover, hidden Markov models (HMM) take into account the fact that it is not possible to measure directly the degradation of a component. Besides, these models have proven to be effective in various domains and have been used for predictive maintenance purposes in several studies (Sikorska, Hodkiewicz, & Ma, 2011) (Tobon-Mejia, Medjaher, Zerhouni, & Tripot, 2011) (Geramifard, Xu, Zhou, & Li, 2012) (Tang, Makis, Jafari, & Yu, 2015). They are based on Markov chains that represent the health state of the studied component.

However, these studies consider simplified cases, in particular, perfect data sets, whereas in real applications the data can be altered. For example, a sensor failure might result in some imprecise or even missing measurements, or several sensors can give conflicting information (Krüger, Measures of Conflicting Evidence in Bayesian Networks for Classification, 2013) (Krüger, Detection of Failing Sensors by Conflicting Evidence in Bayesian Classification, 2014).

This paper proposes a general RUL prediction method with the use of HMM. This RUL is provided with an interval thanks to a new methodology. Finally, we propose a strategy that can be adopted when dealing with incomplete or imprecise data.

The paper is organized as follow: a second section presents HMM and their use for RUL estimation. In the third section, the method is applied to the 2008 IEEE PHM challenge which was a competition aiming at evaluating the performance of prognosis models. The fourth section presents our method for determining the interval. Finally, the last section proposes our strategy in case of uncertain data.

2. HMM for RUL estimation

An HMM can model the degradation of a component, each state of the model representing a health state of the component. Such a model is composed of a Markov process with states that are not directly observable but can be revealed thanks to collected data called observations.

Knowing the model and the probabilities of transitions, it is possible to calculate the mean time of sojourn in each state of the model. Therefore, once the health state of the component has been estimated, we can deduce its RUL (Sikorska, Hodkiewicz, & Ma, 2011).

An HMM is described by the following parameters (Rabiner, 1989) (Degirmenci, 2014):

- The state transition probability distribution $A = \{a_{ij}\}$, where a_{ij} is the transition probability from state number i to state number j . In other words, it is the probability that the system be in the state S_j at the time $t + 1$ knowing that it is in the state S_i at the time t .
- The observation probability distribution of each state $B = \{b_j(k)\}$, where $b_j(k)$ is the probability of producing the observation v_k knowing that the system is in the state S_j .
- The initial distribution $\pi = \{\pi_i\}$, where π_i is the probability that the system is in the state S_i at the initial time.

A HMM is usually defined by the 3 matrices and is written $\lambda = (A, B, \pi)$.

The estimation of a RUL requires the estimation of the current health state of the component: we have an observation sequence and a trained HMM that represents the component and we would like to know the current state the model is in. Among the existing methods, the Viterbi algorithm gives the most optimal sequence of states according to the maximum likelihood criterion. The last state of this sequence corresponds to the current health state.

Once the current health state known, we can estimate the RUL of the component. In most cases, the final state of the HMM is terminal, meaning that once in this state, the model has a 100% chance of staying in this state.

This means that the transition probability matrix has the following form: $A = \begin{pmatrix} Q & R \\ 0 & 1 \end{pmatrix}$, where Q is a square matrix of size $N - 1$, R is a matrix of size $(N - 1) * 1$ and 0 is a zero matrix of size $1 * (N - 1)$.

By recurrence, $A^n = \begin{pmatrix} Q^n & [I + Q + \dots + Q^{n-1}]R \\ 0 & 1 \end{pmatrix}$

We have $\lim_{n \rightarrow \infty} Q^n = 0$, so $\lim_{n \rightarrow \infty} A^n = \begin{pmatrix} 0 & FR \\ 0 & 1 \end{pmatrix}$, where $F = (I - Q)^{-1}$ is called the fundamental matrix of the model. Each element f_{ij} of F is the mean time of the transition from state number i to state number j .

Then, from any state i , the mean time before entering the failed state (state N) is given by: $E_i = \sum_{j=1}^N f_{ij}$

Therefore, the RUL can be obtained by subtracting the time spent in the current state from the mean time before entering the failed state.

3. Application to the IEEE 2008 PHM challenge

This method has been applied to the IEEE 2008 PHM challenge which is a competition sponsored by IEEE in order to evaluate prognostic models (Le, 2016). It consists of 3 data sets: one for model training, one for assessing and improving the model, and the last one for evaluating and comparing the models. The data was built by simulation with NASA's C-MAPSS model (Commercial Modular Aero-Propulsion System Simulation). This model simulates an airplane engine. It takes 3 parameters as inputs and gives 21 sensor measurements as outputs (Saxena, Goebel, Simon, & Eklund, 2008). Figure 1 presents the engine as simulated by the model.

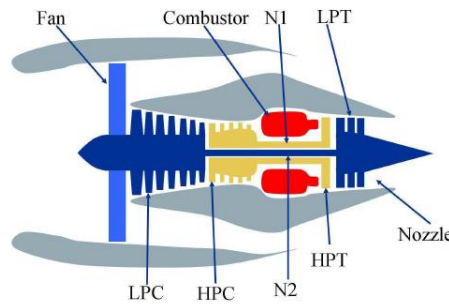


Figure 1. Airplane engine as simulated for the 2008 PHM challenge (Saxena, Goebel, Simon, & Eklund, 2008)

In order to create a HMM describing the degradation of the engine, we need an indicator of this degradation. In this study, we use the same indicator as (Le, 2016) which was originally proposed by (Le Son, Fouladirad, Barros, Levrat, & Iung, 2013). It is constructed from a selection of 7 of the 21 sensors. Figure 2 show this indicator for one of the 218 units that make the training data set.

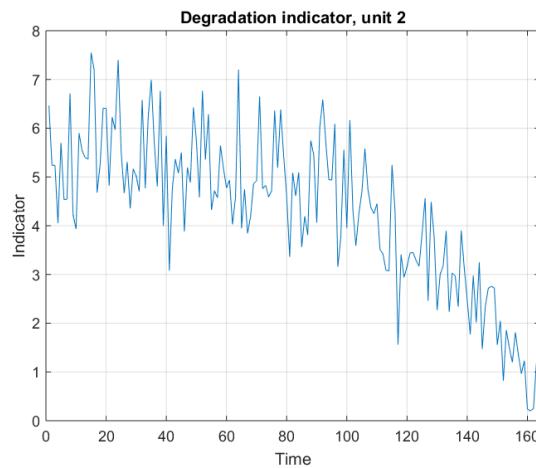


Figure 2. Degradation indicator for the unit number 2

Once the indicator defined, the next step is to build the HMM that will model the engines. To do so, we need to determine the health state and the observation emitted at each time step. Let $S_i(t)$ and $V_i(t)$ be respectively the state of unit number i and the observation emitted by unit i at time t . To determine $S_i(t)$ we proceed as follow:

- We fit the curve of the health indicator by a 3rd degree polynomial P_i
- We define $P_{max}(i) = \max_{1 \leq t \leq T_{failure}^i} (P_i(t))$, where $T_{failure}^i$ is the failure time of unit i . Let P_{MAX} be the 95th percentile of P_{max}
- The value given to $S_i(t)$ depends on the value of $P_i(t)$ compared to some thresholds:
 - If $P_i(t) \geq \frac{N-2}{N-1} P_{MAX}$ then $S_i(t) = 1$ (nominal state)
 - If $\frac{N-k}{N-1} P_{MAX} > P_i(t) \geq \frac{N-k-1}{N-1} P_{MAX}$, $2 \leq k \leq N-2$ then $S_i(t) = k$
 - If $P_i(t) \geq \frac{1}{N-1} P_{MAX}$ then $S_i(t) = N-1$
 - $S_i(T_{failure}^i) = N$

Figure 3 (left) shows how the states are defined for a 6 states HMM

For the observation values, the process is quite similar:

- Let $D_i(t)$ be the indicator made by unit number i at time t , and $D_{max}(i) = \max_{1 \leq t \leq T_{failure}^i} D_i(t)$
- Let D_{MAX} be the 95th percentile of D_{max}
 - If $D_i(t) \geq \frac{N-1}{N} D_{MAX}$ then $V_i(t) = 1$
 - If $\frac{N-k}{N} D_{MAX} > D_i(t) \geq \frac{N-k-1}{N} D_{MAX}$, $1 \leq k \leq N-1$ then $V_i(t) = k+1$

Figure 3 (right) shows how the observation symbols are defined for a 6 states HMM

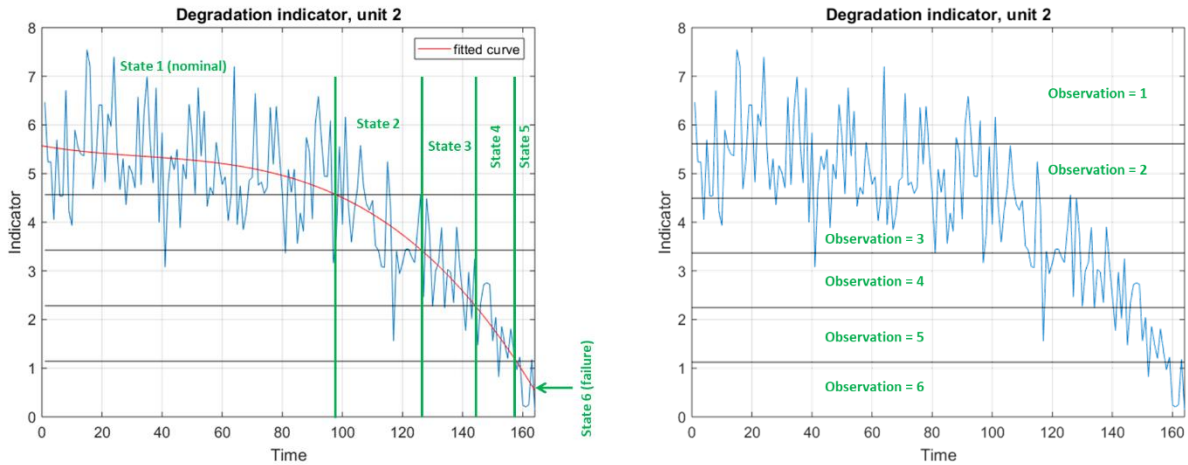


Figure 3. Definition of the states (left) and the observation symbols (right) for each time step for unit number 2

These steps result in the creation of a data set we can actually use for model training. Therefore, the next step is to train our HMM with this data thanks to the Baum-Welch algorithm.

We now have a trained model of the engines and can begin estimating the RULs with the 2nd data set. As before, we construct the health indicator from the 7 chosen sensors and we follow the same procedure to define the observation symbols. Then, we use the Viterbi algorithm to find the most likely state sequence and we take the latest state of this sequence: it is the current health state of the studied engine. Finally, we calculate the RUL thanks to the method described earlier.

4. Defining an interval for the RUL estimation

Like every estimation, the RUL obtained as described in the previous section is a value subject to uncertainties. Therefore, it is important to evaluate the scale of this uncertainty. Consequently, in this section we propose a method to determine an interval for the RUL values. The idea is to create 2 other HMMs that will give an upper and a lower bound for the RUL. The procedure is as follow:

- We calculate the error between the indicator D_i and the fitted polynomial P_i at each time step: $dist_i(t) = P_i(t) - D_i(t)$, $1 \leq i \leq N_{units}$, $1 \leq t \leq T_{failure}^i$

- This error is split in 2: $dist_{sup}^i$ and $dist_{inf}^i$ that store respectively the positive and negative values of $dist_i$
- We calculate the mean values of $dist_{sup}^i$ and $dist_{inf}^i$ over a sliding window. Let's call them $dist_{sup}^i$ and $dist_{inf}^i$
- We create 2 new polynomials:
$$\begin{cases} P_{sup}^i(t) = P_i(t) + \overline{dist_{sup}^i} \\ P_{inf}^i(t) = P_i(t) + \overline{dist_{inf}^i} \end{cases}$$
- These new polynomials result in 2 new sets of state sequence for each unit in the training data set. We then train the model as earlier and obtain 2 new HMMs. One of them will give an upper bound and the other will give a lower bound for our estimated RUL.

Figure 4 (left) shows the fitted curve P_i along with the upper and lower bounds P_{sup} and P_{inf} for one particular units in the training data set.

Knowing the RUL along with its interval, we can plot the survival function of the studied engine with the upper and lower bounds. Figure 4 (right) shows this survival function and the 2 bounds for one particular unit from the test data set. We chose to present the results as survival functions because it is easier to interpret and it provides enough information for maintenance policy implementation.

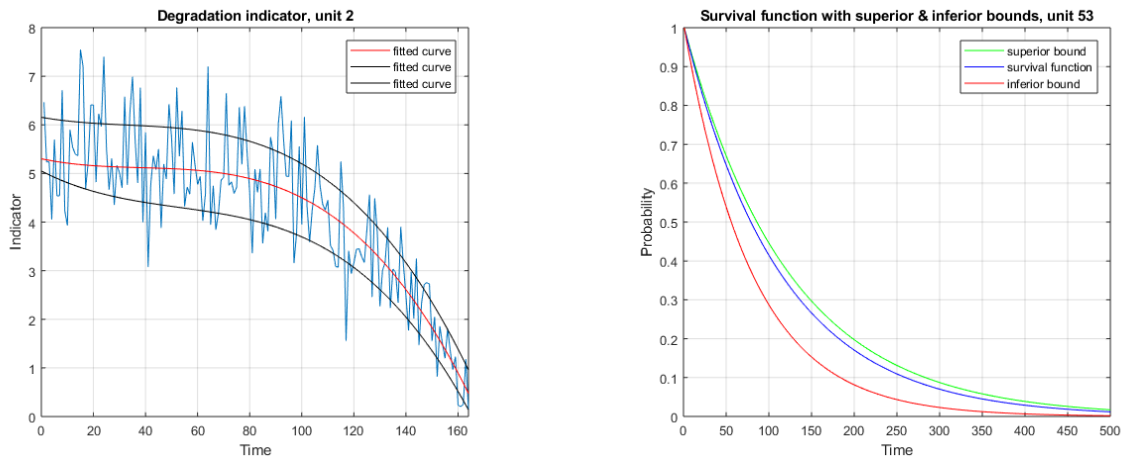


Figure 4. Upper and lower bounds of the polynomial P_i (left). Survival function obtained for unit number 53 with the upper and lower bounds (right)

5. Dealing with uncertain data

Our RUL estimation method is based on a data set that can describe the health state of the component. Therefore, it relies highly on the quality of the collected data and on the confidence we can grant it. However, part of this data can be altered for various reasons and this may have an impact on our results. A HMM being memoryless by definition, if the uncertain data is in the beginning of the observation sequence it will not influence the RUL estimation. But if this data is quite recent, it can impact the result. Consequently, we propose several strategies to overcome this issue:

- Safe attitude: we consider that every piece of uncertain data is an observation of the maximum value (6 in the example presented above). In other words, we consider the worst-case scenario and maximize the safety
- Unsafe attitude: we consider that every piece of uncertain data is an observation of value 1. In other words, we consider the most optimistic scenario
- Medium attitude: we consider that every piece of uncertain data is an observation of mean value (3 or 4 in our example).
- Blind attitude: we make no assumption about the uncertain data. We know each piece of data belongs to the set $\{1, 2, \dots, M\}$ (in our example $M = 6$) so we consider every case and search the maximal and minimal values of the RUL of the component.

These strategies were applied to the 2008 PHM challenge. We artificially altered the data of some units from the test data set. Let T_f^i be the time of the latest piece of data available for unit i . We consider that all the data collected between $T_f^i - 10$ and $T_f^i - 5$ included is damaged and we apply our strategies. Figure 5 shows the survival function obtained by the different attitudes described earlier. The blind attitude considers every possible case, in particular the worst and the most optimistic ones. Therefore, the maximal and minimal values given by this attitude are the same as the unsafe and safe attitudes respectively.

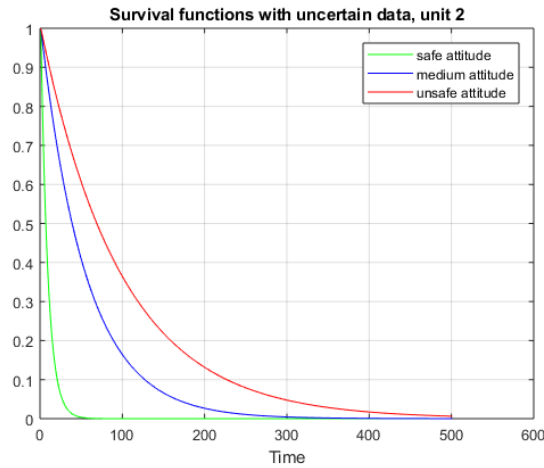


Figure 5. Survival functions obtained with our strategies when dealing with uncertain data

This figure highlights the benefits of such strategies for decision making. Indeed, it helps choose the right time for the maintenance tasks. As an example, we consider that the maintenance should be done when the component has a 0.5 chance to fail. Then, the safe attitude will result in the planification of the maintenance in a very close future while the unsafe attitude will recommend to do the maintenance within the next 70 time steps. In practice, whether one strategy or the other is adopted will depend on several parameters such as the price of the component or whether it is safety related or not.

On the other hand, this figure also shows the limits of the method: as the number of uncertain pieces of data increases, the result of the safe strategy tends toward 0. It can be seen in Figure 5 that the survival function for the safe attitude decreases very quickly towards 0. It shows that the result of the RUL estimation is nearly useless because it gives very little time to plan the required maintenance tasks.

Conclusion

In this paper, we have shown how to calculate components RUL from their degradation state thanks to HMMs. Then, we proposed a solution to give an interval for this estimated RUL. Finally, we proposed a strategy that can be adopted when dealing with uncertain data. These methods have been tested on the IEEE 2008 PHM challenge. The results show they can be significant decision-making support tools thanks to the additional knowledge they provide. However, this application has also underlined some limits.

Future works will try to apply these methods to a real-life case. We will also try to improve them and overcome the drawbacks we discussed in this paper.

References

- Bartelds, G., Biemans, C., Boller, C., Breidne, M., Claesson, Å., Delebarre, C., . . . Worden, K. (2004). *Health Monitoring of Aerospace Structures – Smart Sensor Technologies and Signal Processing*. Southern Gate, Chichester, West Sussex PO19 8SQ, England: John Wiley & Sons Ltd.
- BSI. (2010). Maintenance - Maintenance terminology. Standard, British Standard Institution.
- Degirmenci, A. (2014). Introduction to Hidden Markov Models. Harvard University, 1-5.
- Geramifard, O., Xu, J., Zhou, J., & Li, X. (2012). A Physically Segmented Hidden Markov Model Approach for Continuous Tool Condition Monitoring: Diagnostics and Prognostics. *IEEE Transactions on Industrial Informatics*, 8(4), 964 - 973.

A framework for maintenance 4.0

Erkki Jantunen*, Giovanni Di Orio**, Felix Larrinaga***, Martin Becker****, Michele Albano*****, and Pedro Maló**

*VTT Technical Research Centre of Finland Ltd, P.O. Box 1000, 02044 VTT, Finland, email: erkki.jantunen@vtt.fi

**Dep De Eng. Eletrotécnica, FCT-UNL, UNINOVA-CTS, Lisboa, Portugal

***Mondragon University, Loramendi 4, 20500 Mondragon, Spain

****Fraunhofer IESE, Fraunhofer Platz 1, 67663 Kaiserslautern, Germany

*****CISTER, ISEP/INESC-TEC, Polytechnic Institute of Porto, 4249-015 Porto, Portugal

Abstract. The rapid development of new low-cost sensors of reasonable processing power have led to the introduction of Cyber-Physical Systems (CPS) to support maintenance, which in turn together with a scalable, two level data processing architecture has taken maintenance as such to a new level i.e. Maintenance 4.0. The potential of CPS to support maintenance is explained, and it is related to the requirements set upon the sensors, such as robustness, communication capabilities, intelligence, small size, etc. A new framework and architecture with support for Health management, Prognostics and Collaborative decision-making functionalities that takes the full advantage of the above introduced new technologies is described both theoretically and in the light of some industrial use case examples. A comparative between the approaches implemented in the use cases brings light into the maintenance possibilities to address in such a framework.

1. Introduction

The combination of new emerging technologies and paradigms such as IoT, Machine-to-Machine (M2M), Cloud/Edge computing, service-oriented architecture (SOA), service-oriented computing (SOC) together with the wider dissemination of information technology (IT) and embedded technologies are driving powerful and, above all, new business opportunities for enterprises. As explained by Di Orio et al. (2015) current technological trends are radically changing the way systems in different domains like manufacturing, healthcare, transportation, consumer electronics, etc. are designed and deployed, as well as, monitored and controlled. In particular, to capitalize on the key markets opportunities and winning the competition for markets share, manufacturing companies are caught between the growing needs for: a) implementing more and more exclusive, efficient and sustainable production systems to assure a more efficient and effective management of the resources and to produce innovative and appellative customized products as quickly as possible with reduced costs while preserving product quality; and b) creating new sources of value by providing new integrated product-service solutions to the customer (Cavaliere et. al. 2012). Future internet technologies – such as Internet of Things (IoT) – supported by the wider dissemination of Cyber-Physical Systems (CPS) facilitate the deployment of advanced solutions in plant floor, as well as, applications and services while promoting the meshing of virtual and physical devices and the interconnection of products, processes, infrastructures and more closely connected customers, suppliers and manufacturers within the manufacturing value chain. As stated in Acatech (2011), three main converging trends are pushing manufacturing enterprise into the world of “smartness & connectivity”, namely: 1) Smart embedded systems, mobile services, and pervasive computing (IoT paradigm); 2) Internet based business processes; and 3) social network and communities. In particular, smart embedded systems, mobile services and pervasive computing are contributing to the establishment of a cyber space where data coming from asset and resources can be potentially used to provide some added value. As a parallel to the emerging “Industry 4.0”, which is the term coined for this industrial (r)evolution, a number of subfield have been created, between them “Maintenance 4.0”, which comprises all topics related to the application of Industry 4.0’s techniques and technologies to maintenance. This paper is grounded upon one of these European research initiatives the ECSEL-MANTIS project and describes how it aims at revolutionising maintenance by means of CPS, to attain Maintenance 4.0.

2. Components of Cyber-Physical Systems

Nowadays, the research stream on CPS is extremely active and vibrant as also confirmed by the number of research activities on the topic. As a matter of fact, there is extensive literature dealing with the materialization of the CPS vision and the related challenges technical, social and educational as confirmed in Bytschkow et

al. (2014). As stated in Leitão et al. (2016), the term “Cyber-Physical Systems” was coined in 2006, but several definitions of the term can be found in the literature. According to Baheti et al. (2011) CPS can be defined as transformative technologies that allow to manage interconnected systems between physical assets and computational capabilities. The definitions in Steering Committee (2013) highlight the concept of collaboration and service provisioning. A working definition for CPS has been offered By Cengarle et al. (2013), where a CPS is defined as a system consisting of computational, communication and control components combined with physical processes. Regardless to the specific definition of CPS it is possible to identify their core elements and/or characteristics, extended from Sanislav et al. (2012): 1) Enhancement of physical entities with Cyber capabilities; 2) Networked at multiple and extreme scale; 3) Dynamic behaviour (plug and unplug during operation); 4) High degrees of automation, the control loops are typically closed; 5) High degree of autonomy and collaboration to achieve a higher goal; and 6) tight integration between devices, processes, machines, humans and other software applications. Connected to the term CPS several other terms have been created for better express and defining groups and networks of connected CPSs.

3. Service Oriented Architecture for Cyber Physical Systems

The Device Profile for Web Services (DPWS) has been designed and developed to introduce web services protocols for device networking (Cândido, 2013). It is a standard supported by the OASIS Web Services Discovery and Web Services Devices Profile Technical Committee, since June 2009. DPWS is a stack of web-based protocols and profile for devices, which defines two fundamental elements: the device and its hosted services. Devices perform the discovery and metadata exchange protocols, and hosted services (developed by end-users) provide the functional behaviour of the device DPWS specifies also a set of infrastructure services: a) Discovery services (WS-Discovery): used by a device connected to a network to advertise itself and to discover other devices (Oasis, 2009). WS-Discovery uses User Datagram Protocol (UDP) and a multicast address to broadcast and listen to the discovery messages. b) Metadata exchange services (WS-Metadata Exchange): provide dynamic access to de- vices hosted services and to their metadata, such as WSDL or XML Schema definitions. c) Event publish/subscribe services (WS-Eventing): allow other devices to subscribe to asynchronous messages (events) produced by a given vendor-defined service. DPWS is built on top of the SOAP 1.2 standard, and relies on additional WS specifications, such as WS-Addressing and WS-Policy, to further constrain the SOAP messaging model. At the highest level, the messages correspond to vendor-specific actions and events. Messages are delivered using HTTP, Transmission Control Protocol (TCP) and UDP transport protocols. More details about DPWS can be found in Rajaraajeswari et al. (2016). Due to its popularity, considerable open and not open implementations of DPWS can be found, the most relevant are: WS4D Java Multi Edition DPWS Stack (JMEDS) or the Microsoft Web Service on Device API (WSDAPI). The OPC UA (Unified Architecture) is the new version of the vastly deployed OPC architecture originally designed by the OPC Foundation to connect industrial devices to control and supervision applications as explained by Hadlich (2006). The focus of OPC is on getting access to large amounts of real-time data while ensuring performance constraints without disrupting the normal operation of the devices. The original OPC specifications were based on Microsoft COM/DCOM. The main visible transformation when comparing OPC with and its newer architecture OPC-UA is the shift to cross-platform capable web services (Candido et al. 2010). Moreover, OPC-UA includes support for secure communication, support for the publish-subscribe communication paradigm, unification of several OPC data models as a single set of services and the extension to other domains such as manufacturing (shop-floor), production, maintenance, and business application to implement the total integrated manufacturing. OPC-UA can be mapped using widespread Web standards, including XML, WSDL, SOAP and other WS-* specifications, along with other OPC UA defined specifications. To facilitate and spread as much as possible the usage of OPC-UA technology and strategy/approach the OPC-UA community is collaborating with major industry standards organizations as well as to the possibility to move the information models provided by these organizations to the end user community. Finally, OPC UA provides a homogeneous and generic meta-model and defines a set of web services interfaces to represent and access both structure information and state information in a wide range of devices. The Arrowhead Framework (Delsing et al. 2017) is the result of a large European effort that aimed at normalizing the interaction between industrial IoT applications by means of Service Oriented Architectures.

The approach targeted a number of application domains comprising industrial production, smart buildings, electro mobility, and energy production. The services are exposed and consumed by (software) systems, which are executed on devices. The devices are grouped into local automation clouds, which are self-contained, geographically co-located, independent from each other, and protected from external access through security measures. Services are considered either application services when implementing a use case, or core services when they provide support actions such as service discovery, security, service orchestration, and Quality of Service. The approach considers the interoperability of the involved devices as one of the most important issues, and thus trust all the core services to a common Arrowhead Framework, which also provides registries containing formal descriptions of devices, systems and services. Particular support is given to the orchestration of services, to allow the creation of distributed applications that build over existing ones, for example to create demotics applications that connect to the Virtual Market of Energy that Arrowhead provides, or that interact through the publish/subscribe paradigm extended to the service-oriented world (Albano et al. 2018).

4. Architecture for health management, prognostics and collaborative decision making

The contemporary presence of Ethernet-based solution and distributed computational power in the industrial automation domain is pushing for the productively implementation and deployment of CPS, as confirmed by the several initiatives – such as Reference Architecture for Industry 4.0 (RAMI4.0) and Industrial Internet Consortium – Industrial Internet Reference Architecture (IIC-IIRA) – that have been created to support and enforce the usage/introduction of new internet-based technologies for improving manufacturing companies' flexibility, adaptability, efficiency and effectiveness. In particular, in the context of the MANTIS project, the RAMI4.0 and the IIC-IIRA have been analysed and studied with the objective of creating, developing a delivering reference models from the maintenance perspective i.e. for facilitating the implementation of predictive and proactive maintenance strategies. To do that, MANTIS focuses on the deep usage of IoT sensors and additional devices (Albano et al. 2018) attached to machine controllers to create an ecosystem at the shop floor level where physical assets and resources are virtualized and available as shop floor services in the cyber space. Moreover, edge and cloud computing design, architectural patterns and technical principles are used to create two level data processing architecture where: a) Low level (edge tier): where the extraction, transforming, loading and analysis functionalities are deployed near to the physical assets and resources to model and understand their behaviour while ensuring short-term decisions. b) High level (platform tier): where the extraction, transforming, loading and analysis functionalities are deployed on the cloud to model and understand the global behaviour of physical resources and assets typically using the results of the low-level data processing. The MANTIS approach is in line with the slow but inexorable transformation of the manufacturing systems towards complex environment characterized by aggregations of systems and organized according to a System of Systems (SoS) paradigm and where a large scale of collaboration take place at micro and macro level (Colombo et al. 2014)(Colombo et al. 2009) . In this context, micro level means collaboration between entities at edge level/tier on the contrary macro level collaboration relies collaboration between entities at system and/or platform level/tier. One of the tier is based on the Edge Computing paradigm, which is on the verge of rapid breakthrough. As cloud-based computing is particularly well suited for processing of large amount of information, the delay it introduces may not acceptable for applications requiring time-critical decision in the sub-milliseconds range. The main idea of Edge Computing is to process data as closely to the source as possible, i.e. locally. This processing can be as simple as filtering relevant data, but can also include virtualized services, like machine learning application and real-time visualizations. *****The advantages of edge can be distinguished in three main topics: a) Latency, if reaction speed is critical, latency is very important. An example might be the analysis on in-process sensors which need to be able to trigger a valve in sub-milliseconds. b) Security and privacy. Edge devices with respective computational power and hardware features can be used to set-up hardware encrypted data connections to the cloud to enable better protected pathways to the cloud. But it also enables to process data locally, preventing the requirement to send all data to the cloud, either solving privacy regulation or the potential risk of losing operation-critical data. c) Reliability of connection to the cloud is critical. An unreliable connection to the cloud would mean that (business critical) data processing halts or would be unreliable at best. Where to do the processing in the Edge. Make a decision about where to perform a specific processing is nothing obvious, there are several aspects to

be taken into consideration, such as the device/product where the deployment will be performed, characteristics of the data at edge level, type of analytics to apply, the life-cycle of analytics, the functional perspective and design considerations. Edge data has specific characteristics. It is generated at the network edge and includes machine running data, environment data, and information system data. It features high throughput, fast flow movement, diversity, strong correlation, and high requirements on real-time analysis and processing. Compared with business big data scenarios such as Internet, smart analysis of edge data has the following characteristics and differences: a) Causal relationship vs. association relationship: Edge data is mainly targeted at smart assets that generally run with explicit input and output causal relationships, whereas business big data focuses on data association relationships. b) High reliability vs. low reliability: Industries such as manufacturing have high requirements on the accuracy and reliability of models. If the accuracy or reliability is low, property loss or even personal injury may occur. In contrast, business big data analysis generally has low requirements on reliability. It is required that the edge data analysis result be explainable. Therefore, black-box deep learning is restricted in some application scenarios. The combination of the traditional mechanism models and data analysis methods is the innovation and application direction of smart analysis. c) Small data vs. big data: Assets such as machine tools and vehicles are designed and manufactured by people. Most data in their running process is predictable, and only data generated in abnormal or critical conditions is truly valuable. On the other hand, business big data analysis requires mass data. Nor should one forget that analytics has its own life cycle. Deployment of analytics typically consists of three steps: a) train a (predictive) analytics model, b) test and validate the model on previously unseen data, and c) deploy the model to make predictions on real (streaming) data. Analytic models can be deployed at the edge on heterogeneous computing elements, such as a CPU, GPU or FPGA. For example, FPGAs are ideal deployment targets for very low latency applications. From a functional point of view, the entire data analytics process consists of a series of specific modules that can, or will not, be deployed on the same node. Where to deploy data preprocessing, data analysis, data distribution and policy execution, and visualization and storage functionalities is an architecture-significant decision to make, which is costly to revise. Hence, it demands for careful consideration of design alternatives and trade-offs.

5. Use cases

This section presents three of the near a dozen uses cases implemented in Mantis using the reference architecture presented in the previous section. Each example presents a summary of the objectives, the system implemented (sensors and platform), the data analysis techniques performed at platform level and the results.

5.1 Use case example Proactive Maintenance of Clutch Brake for Press Machines

Objective: This use case (Larrinaga et al. 2018) is concerned with analysing a clutch brake system and its components in press machines to detect the most important failure sources and be able to perform predictive maintenance in those press machines. The overall objective is to early detect internal wear of a clutch-brake. To do that, the static part of the clutch-brake is sensorized. The following conditions were monitored; Friction material slippage, Brake spring degradation, Friction material wear and Piston chamber air leakage. **Sensors & Platform:** Several sensors have been place to measure; friction slippage (encoder sensor velocity and acceleration signals), line pressure, clutch brake input port pressure and chamber pressure (pressure sensors), friction material wear (pressure sensors) and piston chamber air leakage (air mass flowmeter). The architecture implementation agrees with the three-tier architecture presented in section 4 and in (Hegedus et al. 2018). Part of the data is collected and analysed locally (edge tier) in an industrial PLC. Information is also sent to a cloud platform where further analysis is conducted. **Data Analysis Techniques:** To support predictive maintenance of clutch-brake assets data analysis techniques and algorithms are implemented in the platform. These technologies are (1) Root Cause Analysis powered by Attribute Oriented Induction Clustering and (2) Remaining Useful Life powered by Time Series Forecasting. **Results:** At the edge tier, changes on the velocity and acceleration in the clutch-brake process have been detected when slippage situations occur. Additionally the pressure sensors and the model to represent the brake springs stiffness perform with an error less than a 5%. The soft sensor is able to estimate the wear level of the friction material attached to the friction discs, establishing thresholds for different friction wear levels. The air mass flowmeter is able to quantify the air leakage that the whole system is having while the clutch brake is clutched. At platform level and by applying

the Remaining Useful Life techniques, normal behaviour of the system has been modelled and anomalies have been detected (further research is required at this level).

5.2 Use case example Proactive Maintenance of flue gas blowers in power plants

Objective: This use cases addresses optimization of maintenance of components in a power plant. The main optimization maintenance goal is to minimize the total lifecycle maintenance costs and maximize the total lifecycle availability of those components. The use case in Mantis consisted of instrumenting a flue gas recirculation blower at a power plant with several different types of vibration sensors and creating a data collection and data storage system for condition base monitoring of rotating machines. Data collection from sensors is complemented with data from the plant process. All data is stored in the same data storage. Then the goal for data analysis and failure prediction study of the pilot focusses on rotating machines. **Sensors & Platform:** The blower consists of three parts; the engine, bearing and impeller inside the impeller housing. Most common faults for any rotating machines are bearing failures, imbalance and misalignment. All of these causes have distinct vibration patterns that can be identified with proper vibration instrumentation. The instrumentation consisted of vibration sensors installed to the blower bearing and a tachometer for measuring the rotation frequency of the blower. The architecture of the platform agrees with the three-tier architecture presented in section 4 and in (Hegedus et al. 2018) but in this case the implementation used a Microsoft Azure based MIMOSA deployment as the information exchange platform. The data storage follows the Mimosa data structure presented in the OSA-CBM and OSA-EAI standards. **Data Analysis Techniques:** The success/failure of lifecycle optimization is measured by total plant availability and maintenance costs. However power plants are designed to operate for 25 – 40 years making lifecycle optimization difficult. Modern data collection, remote monitoring and analysis tools such as neural networks, statistical analytics or physical models allow cost effective implementation of more advanced maintenance methods on less critical components. **Results:** The bearing fault was detected about half a year before the change of the bearing based on manual expert analysis of the measured vibration signals. During the project the vibration was measured with manual method and two on-line detection systems which provided data to the cloud. For the development of the automatic prediction of the remaining useful life of the bearing a simulator that can provide vibration signals was developed i.e. due to the very limited amount of real data also simulated has been made available for the development of the prognosis tools.

6. Conclusion

The paper discussed how CPS can revolutionize maintenance, and presented architectures that were developed to support them, and real-life use cases where the CPS have provided dramatic improvement. The use cases show that different requirements can be implemented over the same reference architecture and that data analysis and decision making algorithms can and should be implemented both at the edge tier (for fast response) and at platform tier for combination of parameters and conditions from different sensors, components and systems. Future work will consider analysing these implemented systems of systems, to both provide quantitative measures of their benefit to business cases, and comparison between them in real life.

Acknowledgement

This work has been developed with the support of funds made available provided by the European Commission in the scope of ECSEL/H2020 MANTIS Research and Innovation Action (Project ID: 662189), by the Portuguese Fundação para a Ciência e a Tecnologia (FCT, I.P.) in the framework of project UID/EEA/00066/2013 PEST (Strategic Plan for Science and Technology) for the Centre of Technology and Systems (CTS), by the Finnish Funding Agency for Innovation Tekes, and by Ministerio de Industria, Energía y Turismo (Spain).

References

Acatech (2011) *Cyber-Physical Systems - Driving force for innovations in mobility*. Springer.
Albano M., Lino Ferreira F., Di Orio G., Maló P., Webers G., Jantunen E., Gabilondo I., Viguera M., Papa G., and Novak F., et al. (2018) Sensors: the Enablers for Proactive Maintenance in the Real World, 5th

- International Conference on Control, Decision and Information Technologies (CoDIT 2018)*. Thessaloniki, Greece. 10 to 13, April, 2018
- Albano M, Lino Ferreira L, Sousa J, (2016) Extending publish/subscribe mechanisms to SOA applications, *12th IEEE World Conference on Factory Communication Systems (WFCS 2016)*, May 3-6, 2016, Aveiro, Portugal. pp. 1-4 DOI: 10.1109/WFCS.2016.7496528
- Baheti R, Gill H, (2011) Cyber-physical systems, *The Impact Control Technology*, T. Samad and A.M. Annaswamy (eds.), pp. 161–166
- Bytschkow D, Campetelli A, Cengarle MV, Irlbeck M, Schorp K (2014) *Reference Framework for the Engineering of Cyber-Physical Systems: A First Approach*
- Cândido GM (2013) Service-oriented architecture for device lifecycle support in industrial automation
- Candido G, Jammes F, de Oliveira JB, Colombo AW (2010) SOA at device level in the industrial domain: Assessment of OPC UA and DPWS specifications, *8th IEEE International Conference on Industrial Informatics (INDIN)*, pp. 598–603.
- Cavalieri S, Pezzotta G (2012) Product–Service Systems Engineering: State of the art and research challenges, *Computers in Industry* vol. 63, no. 4, pp. 278–288
- Cengarle M, Bensalem S, McDermid J, Passerone R, Sangiovanni-Vincetelli A, Torngren M, (2013) Characteristics, capabilities, potential applications of Cyber-Physical Systems: a preliminary analysis, D2.1
- Colombo, A., Bangemann, T., Karnouskos, S., Delsing, J., Stluka, P., Harrison, R., Jammes, F., Lastra, J.L. (Eds.), 2014. *Industrial Cloud-Based Cyber-Physical Systems: The IMC-AESOP Approach*, 2014 edition. ed. Springer, New York.
- Colombo, A.W., Karnouskos, S., 2009. Towards the factory of the future: A service-oriented cross-layer infrastructure, *ICT Shaping the World: A Scientific View*, European Telecommunications Standards Institute (ETSI) John Wiley Sons 65, 81.
- Delsing, J (2017) The Arrowhead Framework architecture, *IoT Automation: Arrowhead Framework*, CRC Press, pp 45-91.
- Di Orio G, Cândido G, Barata J (2015) The Adapter module: A building block for Self-Learning Production Systems, *Robotics and Computer-Integrated Manufacturing*, vol. 36, pp. 25–35
- Hadlich T (2006) Providing device integration with OPC UA, in *Industrial Informatics, IEEE International Conference on Industrial Informatics*, pp. 263–268.
- Hegedus C., Varga P. and Moldovan I. «The MANTIS architecture for proactive maintenance.» 5th International Conference on Control, Decision and Information Technologies CODIT 2018 Thessaloniki (Greece), 2018.
- Larrinaga F., Fernandez-Anakabe J., Zugasti E., Garitano I., Zurutuza U., Anasagasti M., Mondragon M. «Implementation of a Reference Architecture for Cyber Physical Systems to support Condition Based Maintenance.» 5th International Conference on Control, Decision and Information Technologies CODIT Thessaloniki (Greece), 2018.
- Leitão P, Colombo AW, Karnouskos S (2016) Industrial automation based on cyber-physical systems technologies: Prototype implementations and challenges, *Computers in Industry*, vol. 81, pp. 11–25
- OASIS (2009) OASIS Web Services Dynamic Discovery (WS-Discovery) Version 1.1 [Online], Available: <http://docs.oasis-open.org/ws-dd/ns/discovery/2009/01>.
- Rajaraajeswari S, Selvarani R, Raj P (2016) Integration Approaches for the Internet of Things (IoT) Era in *Connectivity Frameworks for Smart Devices*, Z. Mahmood, Ed. Springer International Publishing, pp. 117–146.
- Sanislav T, Miclea L (2012) Cyber-Physical Systems - Concept, Challenges and Research Areas, *Journal of Control Engineering and Applied Informatics*, vol. 14, no. 2, pp. 28–33
- Steering Committee (2013) *Strategic R&D Opportunities for 21st Century Cyber-Physical Systems, Connecting computer and Information systems with physical world*.

Modeling spare part demand data using transmuted or exponentiated Poisson distribution

R. Jiang * and Z. Chen **

* Faculty of Automotive and Mechanical Engineering Changsha University of Science and Technology Changsha, Hunan 410114, China. Email address: jiang@csust.edu.cn.

** School of Mathematics and Statistics Changsha University of Science and Technology Changsha, Hunan 410114, China. Email address: 734316364@qq.com.

Abstract. Traditionally, it is assumed that the lead time spare part demand follows the Poisson or normal distribution. It is well-known that the mean of Poisson distribution equals to its variance, which is called Poisson dispersion. Actual demand data scarcely meet the Poisson dispersion. To overcome this problem, several two-parameter variants (e.g., zero-inflated Poisson and hurdle shifted Poisson) of the Poisson distribution have been used for modelling irregular spare part demand data. In this paper, we propose two new two-parameter alternatives, which are transmuted and exponentiated Poisson distributions. The proposed distributions can be over- or under-Poisson dispersive, and hence provide better flexibility without losing the simplicity of the Poisson distribution too much. The appropriateness of the proposed models is illustrated through three real-world examples. Based on the idea that a fitted distribution should provide a good goodness-of-fit to the right tail of the empirical demand distribution so as to achieve good inventory performance, a weighted least square method is also proposed to estimate the model parameters. The appropriateness of the proposed approach is also illustrated.

Keywords. Spare parts demand, transmuted Poisson distribution, exponentiated Poisson distribution, weighted least square method

1. Introduction

The determination of the optimal stock level in an inventory management problem needs to appropriately use a demand distribution (Costantino et al. 2017). The standard assumption is that lead time demand follows a normal or Poisson distribution. This is often not true, particularly for high irregular demand patterns (e.g., see Turrini and Meissner 2017). The departures from the assumed distribution may lead to very unsatisfactory inventory performance, e.g., high inventory costs (Nenes et al. 2010, Turrini and Meissner 2017).

A number of different distribution models have been used in the literature, including the stuttering Poisson (or geometric Poisson e.g., see Chen 2012), zero-inflated Poisson (ZIP), hurdle shifted Poisson (HSP), negative binomial, normal and gamma distributions (e.g., see Snyder et al. 2012, Costantino et al. 2017, Turrini and Meissner 2017). Among these distributions, the stuttering Poisson is computationally intensive, ZIP, HSP and negative binomial distribution are only applicable for over-Poisson dispersive cases and the normal and gamma distributions are continuous distributions.

In this paper, we present two two-parameter extended Poisson distributions: exponentiated Poisson and transmuted Poisson distributions as alternative distributions for modelling spare parts demand. The exponentiated Poisson was previously used in modelling bus-motor failure data (Jiang 2010). In this paper, we apply it to model spare parts demand data. We will show that it provides very good goodness-of-fit to the spare parts demand data. The transmuted Poisson distribution is a new model. We study its properties and examine its appropriateness in modelling spare parts demand data.

According to Turrini and Meissner (2017), most inventory applications involve calculating an upper percentile (e.g., 98%) of the forecasted demand distribution, implying that the right tail of the distribution provides more information for inventory management. Based on this observation, we propose a weighted least square method (LSM) for fitting a distribution to a set of demand data in order to achieve good inventory performance. We use a normal distribution function as the weight function, whose parameters depend on the sample mean. The fitted model obtained from the weighted method is expected to improve the inventory performance.

The paper is organized as follows. In Section 2, we summarize two-parameter extensions of the Poisson distribution and study the properties of the transmuted Poisson distribution. The weighted LSM is presented in Section 3. The appropriateness and usefulness of the proposed models and approach are illustrated in Section 4. The paper is concluded in Section 5.

2. Proposed model and its properties

2.1 Earlier extensions of the Poisson distribution

There are several extended Poisson distributions and those extensions with two parameters are outlined as follows.

The first extension is the stuttering Poisson distribution, which is obtained by compounding the Poisson distribution with a geometric distribution (Chen 2012). Its probability mass function (pmf) is given by

$$f(x) = \begin{cases} e^{-\lambda}, & \text{for } x = 0 \\ \sum_{i=1}^x C_{x-1}^{i-1} (1-\rho)^i \rho^{x-i} \frac{\lambda^i e^{-\lambda}}{i!}, & \text{for } x > 0 \end{cases}, \lambda > 0, \rho \in (0, 1). \quad (1)$$

The mean and variance are given by

$$\mu = \lambda / (1 - \rho), \sigma^2 = \mu(1 + \rho) / (1 - \rho). \quad (2)$$

From equation (2), we have $\sigma^2 > \mu$, implying that it is always over-Poisson dispersive. Clearly, it is computationally intensive and hence its application is limited.

The second extension is the zero-inflated Poisson distribution. It is developed for modeling the excess zeros (e.g., see Lambert 1992, Majeske and Herrin 1995, and Costantino et al. 2017). The pmf of the ZIP model is given by

$$f(x) = \begin{cases} 1 - \alpha + \alpha e^{-\lambda}, & \text{for } x = 0 \\ \alpha \frac{\lambda^x e^{-\lambda}}{x!}, & \text{for } x > 0 \end{cases}, \lambda > 0, \rho \in (0, 1). \quad (3)$$

The mean and variance are given by

$$\mu = \alpha \lambda, \sigma^2 = \mu(1 + \lambda - \alpha \lambda). \quad (4)$$

Clearly, we have $\sigma^2 > \mu$, implying that it is always over-Poisson dispersive.

The third extension is the hurdle shifted Poisson distribution (SPD), which is a mixture of a deterministic (or degenerate) distribution and the shifted Poisson distribution (Snyder et al. 2012). Its pms is given by

$$f(x) = \begin{cases} q, & \text{for } x = 0 \\ (1-q) \frac{\lambda^{x-1} e^{-\lambda}}{(x-1)!}, & \text{for } x > 0 \end{cases}. \quad (5)$$

The mean and variance are given by

$$\mu = (1-q)(\lambda+1), \sigma^2 / \mu = q(1+\lambda) + \lambda / (1+\lambda). \quad (6)$$

Clearly, it can be over-Poisson dispersion (when λ or q is large) or under-Poisson dispersion (when λ and q are small).

2.2 Proposed extensions of the Poisson distribution

The first author proposes an exponentiated Poisson distribution (EPD) for modelling bus-motor failure data (Jiang 2010). The EPD is defined by

$$F(x) = e^{-\lambda \gamma} \left(\sum_{i=0}^x \frac{\lambda^i}{i!} \right)^\gamma, \lambda, \gamma > 0. \quad (7)$$

To examine its Poisson dispersion, we randomly generate a set of combinations of λ and γ . For each combination, we calculate the values of μ and σ^2 . Figure 1 shows the plot of μ / σ^2 vs. γ , and Figure 2 shows the plot of μ vs. λ . As can be seen from these two figures, the EPD can be either over-Poisson dispersion (when $\gamma < 1$) or under-Poisson dispersion (when $\gamma > 1$); and the mean demand (μ) is highly correlated with the value of λ .

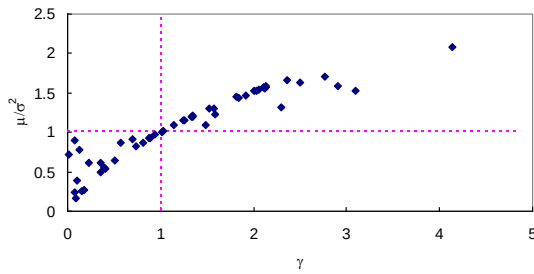


Figure 1 Plot of μ/σ^2 vs. γ

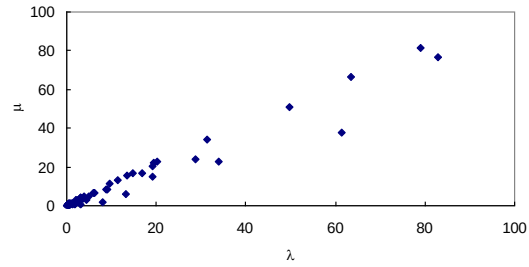


Figure 2 Plot of μ vs. λ

2.3 Transmuted Poisson distribution

For an arbitrary distribution function $G(x)$, Shaw and Buckley (2009) define a transmuted distribution as

$$F(x) = (1 + \alpha)G(x) - \alpha G^2(x), \quad |\alpha| \leq 1. \quad (8)$$

When $G(x)$ is the Poisson distribution, we obtain the transmuted Poisson distribution (TPD). In this paper we examine the appropriateness of TPD as an alternative distribution to model the demand data.

To examine its Poisson dispersion, we randomly generate a set of combinations of λ and α . For each combination, we calculate the values of μ and σ^2 . Figure 3 shows the plot of μ/σ^2 vs. α ; and Figure 4 shows the plot of μ vs. λ . As can be seen from these two figures, TPD can be slightly over-Poisson dispersive (when $\alpha > 0$ and λ is relatively small) or under-Poisson dispersion; and the mean demand (μ) is highly correlated with the value of λ .

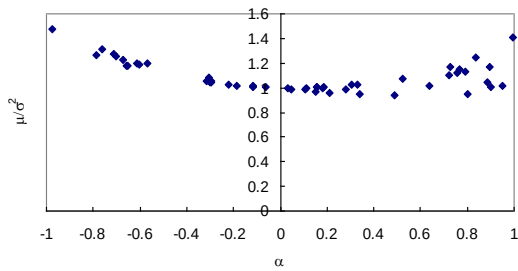


Figure 3 Plot of μ/σ^2 vs. α

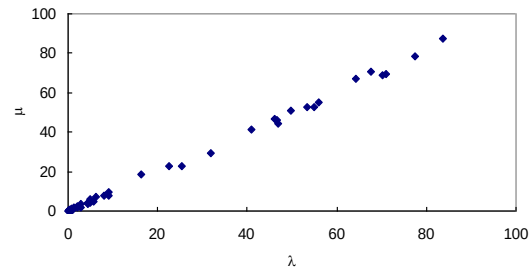


Figure 4 Plot of μ vs. λ

3. Weighted least square method

Suppose that we have a set of demand data given by

$$(n_i, 0 \leq i \leq m). \quad (9)$$

The empirical distribution function is given by

$$F_i = \sum_{j=0}^i n_j / \sum_{j=0}^m n_j, \quad 0 \leq i \leq m; \quad F_i = 1, i > m. \quad (10)$$

Consider a parametric model $F(i; \theta)$, where θ is the parameter set to be estimated. To stress the observations with large value of i , we introduce a weight function $w(i)$, which is positive and increases with i . In this paper, we define the weight function as

$$w(i) = \Phi(i; \mu_s, \sqrt{\mu_s}) \quad (11)$$

where $\Phi(i; \mu_s, \sqrt{\mu_s})$ is the normal distribution with mean and variance being the sample average μ_s . The parameters are estimated through minimizing the sum of weighted square errors (SWSE)

$$SWSE = \sum_{i=0}^N w(i) [F(i; \theta) - F_i]^2, \quad N > m. \quad (12)$$

We use $1 - F_{i_{0.98}}$ to evaluate the performance of the fitted model, where $i_{0.98}$ is the 98 percentile of the fitted distribution. A good fit implies:

$$1 - F_{i_{0.98}} \approx 0.02. \quad (13)$$

4. Illustrations

In this section, we present two real-world examples to illustrate the appropriateness of the proposed models and weighted approach.

4.1 Example 1

The data shown in Table 1 come from Kogan and Rind (2011) and deal with seven years historical consumption of 25 kV distribution transformers.

Table 1 Demand data for Example 1

	2001	2002	2003	2004	2005	2006	2007
1	168	30	8	5	7	10	4
2	264	12	12	0	29	8	2
3	106	0	12	13	6	8	0
4	65	0	20	9	18	0	1
5	213	62	23	7	28	0	19
6	179	49	13	20	20	13	16
7	156	30	29	14	21	4	1
8	162	65	17	19	21	2	20
9	67	22	7	3	0	0	27
10	102	5	10	6	0	1	3
11	7	24	13	0	11	0	28
12	23	14	4	2	15	0	1

It is noted that the spare parts consumed in 2001 is much more than the spare parts consumed in 2002. Kogan and Rind (2011) think that the cause is due to the impact of extreme weather events. In fact, the cause may also be due to earlier failures. In this paper, we exclude the data in 2001.

The lead-time is 8 months. We consider all the sums in any successive eight months. Table 2 shows the lead time demand data. The sample mean and variance are 98.64 and 3724.55, respectively. Clearly, the data are highly over-Poisson dispersive.

Table 2 Lead time demand data

248	271	117	134	116	61	87	71	71	150	116	52	45	20	8	63	115
240	217	118	133	98	61	85	71	96	143	98	52	35	24	27	86	
233	180	126	131	85	71	91	80	110	114	86	65	28	13	42	87	
257	162	131	132	69	72	78	72	131	119	73	58	20	9	43	115	

Table 3 shows the parameters estimated from the weighted least square method (WLS) and corresponding performances in terms of $1-F_{i_{0.98}}$. To illustrate the advantage of the proposed parameter estimation method, we also show the parameter and corresponding performance of the Poisson distribution (in the second row), obtained from the moment method (MM). From the table, we have the following observations:

- According to the results in the second and third rows, the model fitted from the WLS method outperforms the model fitted from the moment method in terms of $1-F_{i_{0.98}}$.
- TPD offers the best goodness-of-fit in terms of SWSE and $1-F_{i_{0.98}}$.
- The performance of TPD is almost the same as the performance of the Poisson distribution. This is because the data is highly over-Poisson dispersive.

Table 3 Estimated parameters and model performances for Example 1

Model	Method	λ	α, γ, q	SWSE	$i_{0.98}$	$1-F_{i_{0.98}}$
Poisson	MM	98.6		3.192	120	0.2615
Poisson	WLS	104.7		2.939	126	0.2462
ZIP	WLS	139.3	0.3539	0.9269	158	0.1231
SPD	WLS	138.3	0.6462	0.9284	158	0.1231
EPD	WLS	172.4	0.0301	0.7092	173	0.1077
TPD	WLS	105.3	0.1067	2.934	126	0.2462

4.2 Example 2

The data shown in Table 4 come from Guo et al. (2017) and deal with the annual total spare parts turnover numbers (which reflect the actual demand of a repairable unit) of Parts 1 and 2. The sample mean and variance for Part 1 [Part 2] are 5.4 and 3.14 [9.3 and 12.70], respectively, implying under-Poisson [over-Poisson] dispersion.

Table 5 shows the parameters estimated from the MM (for the Poisson distribution) and WLS method as well as the corresponding performances. From the table, we have the following observations:

- According to the results in the 3rd and 4th rows as well as the 11th and 12th rows, the model fitted from WLS method outperforms the model fitted from MM in terms of SWSE, $i_{0.98}$ and $1-F_{i_{0.98}}$.

- For the demand data of Parts 1 and 2, EPD offers the best goodness-of-fit.
- For the demand data of Part 1, TPD is the second best model. This is because this data set is under-Poisson dispersive.
- For the demand data of Part 2, TPD is the third best model (slightly poorer than the ZIP). This is because this data set is over-Poisson dispersive.

This illustrates the appropriateness of the proposed models and estimation approach.

Table 4 Demand data for Example 2

Year	Part 1	Part 2	Year	Part 1	Part 2	Year	Part 1	Part 2	Year	Part 1	Part 2
1	7	11	14	6	10	27	6	11	40	3	11
2	6	9	15	8	9	28	5	6	41	4	8
3	8	10	16	9	16	29	4	4	42	4	7
4	6	8	17	5	18	30	4	5	43	3	6
5	5	8	18	8	10	31	5	5	44	4	5
6	10	9	19	7	18	32	3	12	45	4	7
7	7	12	20	4	12	33	3	7	46	3	5
8	6	7	21	6	10	34	5	9	47	5	13
9	5	5	22	5	8	35	7	8	48	3	8
10	5	6	23	7	9	36	8	15	49	4	9
11	6	6	24	5	7	37	4	17	50	4	7
12	4	13	25	4	7	38	7	9			
13	4	8	26	9	8	39	6	17			

Table 5 Estimated parameters and model performances for Example 2

Model	Method	λ	α, γ, q	SWSE, 10^{-3}	$i_{0.98}$	$1 - F_{i_{0.98}}$
Part 1						
Poisson	MM	5.400		7.074	11	0
Poisson	WLS	5.193		4.525	10	0
ZIP	WLS	n/a ($\square \square > 1$)				
SPD	WLS	n/a ($q < 0$)				
EPD	WLS	4.330	1.818	0.9993	10	0
TPD	WLS	6.704	0.9866	1.428	10	0
Part 2						
Poisson	MM	9.300		23.93	16	0.08
Poisson	WLS	9.112		22.58	16	0.08
ZIP	WLS	9.557	0.8844	20.19	16	0.08
SPD	WLS	8.712	0.1443	22.56	16	0.08
EPD	WLS	11.42	0.4368	12.67	17	0.04
TPD	WLS	9.660	0.3004	20.90	16	0.08

5. Conclusions

In this paper we have proposed to use the transmuted and exponentiated Poisson distributions as alternative models for modelling spare parts demand data. We have fitted these two models, along with the Poisson distribution as well as zero-inflated and hurdle shifted Poisson distributions, to three sets of demand data and found that the exponentiated Poisson distribution offers the best goodness-of-fit and the transmuted Poisson distribution is the second best model for under-Poisson dispersive demand data. The main advantages of the exponentiated and transmuted Poisson distributions are their flexibility and relative simplicity.

To improve the inventory performance, we have proposed a weighting approach to estimate the parameters of a distribution model. Its appropriateness has been confirmed.

Topics for future study include

- To carry out more case studies to further check their appropriateness for modelling spare part demand data,
- To examine the inventory performance when the proposed models are used to model the lead time demand, as conducted by Turrini and Meissner (2018),
- To compare the proposed models with other models such as the negative binomial distribution,
- To optimize the parameters of the weight function, and
- To apply the proposed models to other areas.

Acknowledgment

The research was supported by the National Natural Science Foundation of China (No. 71771029).

References

- Chen CW (2012) Using geometric Poisson exponentially weighted moving average control schemes in a compound Poisson production environment. *Computers and Industrial Engineering*, 63 (2), pp. 374-381.
- Costantino F, Gravio GD, Patriarca R, Petrella L (2017) Spare parts management for irregular demand items. *Omega*, in press, Available online 28 September 2017.
- Guo F, Diao J, Zhao QH, Wang DX, Sun Q (2017) A double-level combination approach for demand forecasting of repairable airplane spare parts based on turnover data. *Computers and Industrial Engineering* 110, pp. 92-108.
- Jiang R (2010) Discrete competing risk model with application to modeling bus-motor failure data, *Reliability Engineering and System Safety*, pp. 981-988.
- Kogan V, Rind T (2011) Determining critical power equipment inventory using extreme value approach and an auxiliary Poisson model. *Computers and Industrial Engineering*, 60(1), pp. 25-33.
- Lambert D (1992) Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34, pp. 1-14.
- Majeske KD, Herrin GD (1995) Assessing mixture-model goodness-of-fit with an application to automobile warranty data. In: *Proceedings of Annual Reliability and Maintainability Symposium*, pp. 378-383.
- Nenes G, Panagiotidou S, Tagaras G (2010) Inventory management of multiple items with irregular demand: A case study. *European Journal of Operational Research*, 205(2), pp. 313-324.
- Shaw WT, Buckley IR (2009) The alchemy of probability distributions: beyond Gram-Charlier expansions, and a skew-kurtotic normal distribution from a rank transmutation map.
- Snyder RD, Ord JK, Beaumont A (2012) Forecasting the intermittent demand for slow-moving inventories: A modelling approach. *International Journal of Forecasting* 28, pp. 485-496.
- Turrini L, Joern Meissner JM (2017) Spare parts inventory management: new evidence from distribution fitting. *European Journal of Operational Research*, in press, Available online 29 September 2017.

Optimal Sequential Resource Allocation under Error-prone Success Assessment

K. Kalyanam¹, M. Pachter², D. Casbeer³

¹InfoSciTex Corporation, a DCS Company, Dayton, OH 45431, USA: krishnak@ucla.edu

²Air Force Institute of Technology, Wright-Patterson A.F.B., OH 45433, USA: meir.pachter@afit.edu

³Air Force Research Laboratory, Wright-Patterson A.F.B., OH 45433, USA: david.casbeer@us.af.mil

Abstract We formulate a stochastic optimization problem, wherein finite resources are sequentially allocated to incoming tasks so as to maximize a cumulative reward. At each decision stage, the incoming task has known value. However, upon allocating a resource, a task is completed with probability less than one. Furthermore, the decision maker is informed about the task completion status via an error-prone feedback mechanism that is subject to both type I and type II classification errors. The decision maker may choose to reallocate a resource to the current task or move on to the next task in the sequence. We show that a resource is allocated if and only if the expected reward from a task exceeds a threshold. Furthermore, we establish a lower bound on the task completion probability under which the threshold is monotonic decreasing in the number of remaining resources.

1. Introduction

The operational scenario is the following. A Decision Maker (DM) with finite resources sequentially engages incoming tasks. Should the DM decide to allocate a resource to a task, it will be completed with probability $p < 1$. We assume that a single resource allocated to an incomplete task either leads to completion or it does not; there is no partial or cumulative task completion. Upon successful completion, the DM receives a known reward. However after allocating a resource, the DM only has access to an error-prone status report on the task. Indeed, the feedback process is an error prone classifier in that a completed task may be reported as being incomplete and vice-versa. Upon receiving the status report, the DM can either re-assign a resource to the current task or move on to the next task in the sequence. This decision obviously relies on the probability that the task is still not complete given the history of observations and the number of tasks and resources left. There is no turning back i.e., the DM is not allowed to revisit tasks that had been previously serviced. We are interested in the optimal feedback policy that results in the maximal expected cumulative reward. In particular, we are interested in the structural property, if any, of the optimal policy.

1.1. Prior Work

The framework we consider is applicable in military weapon target engagements, e.g., see Mastran and Thomas (1973), Kisi (1976), Aviv and Kress (1997), shooting problems, e.g., see Sato (1997a), Glazebrook and Washburn (2004), cybersecurity, e.g., see Hu et al. (2017), Gao et al. (2013), attacker-defender games, e.g., see Levitin and Hausken (2012) and more broadly to stochastic sequential resource allocation, e.g., see Sato (1996), Sato (1997b). In sequential assignment problems of the kind considered herein, the decision rule usually takes the form, *service the task iff its value is no less than a certain threshold c* , where the optimal c is to be determined. Moreover, one would expect the optimal threshold to be monotonic decreasing in the number of remaining resources. If the probability of successful task completion, $p = 1$, there is no need for repeated allocation to the same task and the resulting problem bears similarity to Revenue Management (RM) - wherein, the threshold monotonicity property is known to hold, e.g., see van Ryzin and Talluri (2005), Aydin et al. (2009). However, it is not obvious that the result holds for $p < 1$. To clearly distinguish this paper from our earlier work, we note that:

- 1) In Krishnamoorthy et al. (2016), a perfect information scenario is considered, where the DM is informed if the task was completed or not and the threshold monotonicity result has been established for this case.
- 2) In Krishnamoorthy et al. (2017), the rewards are considered to be random (as in RM) but known prior to decision time and the task status is perfectly known to the decision maker; the threshold monotonicity result has been established for this case as well. In addition, for this case, we also showed that the threshold is monotonic non-decreasing in the number of remaining tasks.

In this article, we extend the monotonicity result in Krishnamoorthy et al. (2016) to the case of an error-prone feedback mechanism. We also note that the result in Krishnamoorthy et al. (2016) does not follow directly from the result established in this work - see Remark 2 for alternate analysis.

In military parlance, our model embraces a Shoot-Look-Shoot (SLS) approach, in that homogenous resources are assigned one at a time with observations in between that assess the success or failure of the previous allocation. The analysis of SLS strategies in the presence of partial information was first reported in Aviv and Kress (1997). Gaver and Jacobs (1997) consider a conventional classifier model with type I and type II errors and provide effectiveness measures for different shooting tactics. A survey of the state of the art in SLS methods and the optimal use of information is provided in Glazebrook and Washburn (2004). It is also known that if additional complicating factors such as a search cost for finding a task or a scenario wherein resources can be replenished are considered, the monotonicity property breaks down, e.g., see Sato (1996), Sato (1997b).

In this article, we present a low resolution model of the SLS scenario where an imperfect binary classifier is employed. Our emphasis is on providing an analytical relationship between the probabilities of correctly classifying a completed task and misclassifying an incomplete task and the task completion probability under which the threshold monotonicity property holds.

2. Dynamic Model

Let the two possible observations after a resource is allocated to a task be given by $y = 1, 2$, where 1 indicates that the task is reported to be incomplete and 2 indicates that the task is reported to be complete. Furthermore, let the conditional probability that $y = 1$ given that the task is incomplete be given by q_1 and the conditional probability that $y = 1$ given that the task is complete be given by q_2 .

2.1. Monotonicity condition

Let P_t be the prior probability that task t is incomplete. Now, suppose a resource is allocated to it. After expending the resource, the DM receives feedback/ observation: $y = 1, 2$, indicating that the task is reported to be incomplete and complete respectively. Let $P_t^+(y)$ indicate the a posteriori probability that task t is incomplete after receiving the report y . We have the conditional probabilities: $Prob(y = 1|incomplete) = q_1$ and $Prob(y = 1|complete) = q_2$.

THEOREM 1. *If $q_2 < q_1 \leq \frac{1}{1-p}q_2$, then $P_t^+(y) < P_t$, for all y .*

Proof. If $q_1 > q_2$, the probability of correctly identifying a completed task is more than the probability of mis-classifying an incomplete task. One would expect a reliable and “truthful” classifier to satisfy this condition. Let us refer to the probability that the task is incomplete prior to the observation as $\alpha_t = P_t(1 - p)$. To compute the a posteriori probability, we use Bayes’ theorem and get:

$$P_t^+(y = 1) = \frac{q_1 \alpha_t}{q_1 \alpha_t + q_2 (1 - \alpha_t)} \text{ and} \quad (1)$$

$$P_t^+(y = 2) = \frac{(1 - q_1) \alpha_t}{(1 - q_1) \alpha_t + (1 - q_2) (1 - \alpha_t)}. \quad (2)$$

If $q_1 > q_2$, we have:

$$\begin{aligned} \beta &= q_1 - q_2 > 0, \\ \Rightarrow \alpha_t \beta &> \alpha_t^2 \beta, \text{ since } \alpha_t < 1, \\ \Rightarrow \alpha_t \beta + \alpha_t q_2 &> \alpha_t^2 \beta + \alpha_t q_2, \\ \Rightarrow P_t^+(y = 1) &= \frac{\alpha_t q_1}{\alpha_t \beta + q_2} > \alpha_t. \end{aligned} \quad (3)$$

A similar argument shows that $P_t^+(y = 2) < \alpha_t$. So, we have:

$$P_t^+(y = 2) < \alpha_t < P_t^+(y = 1). \quad (4)$$

In other words, given that the classifier is a reliable indicator, a negative report that the *task is incomplete* increases the posterior probability (compared to the prior) and a positive report has the opposite effect. In

addition, for monotonicity, we require:

$$\begin{aligned}
 P_t^+(y=1) &\leq P_t, \\
 &\Rightarrow \frac{q_1(1-p)}{q_1P_t(1-p) + q_2(1-P_t(1-p))} \leq 1, \\
 &\Rightarrow q_1(1-p)(1-P_t) \leq q_2(1-P_t + (1-p)), \\
 &\Rightarrow (q_1 - q_2)(1-p)(1-P_t) \leq q_2p.
 \end{aligned} \tag{5}$$

Since $P_t \leq 1$, a sufficient condition that guarantees monotonicity is given by:

$$p \geq 1 - \frac{q_2}{q_1}. \tag{6}$$

To summarize, if $q_1 > q_2$ and the successful task completion probability p is sufficiently high (6), the a posteriori probabilities satisfy: $P_t^+(y) \leq P_t$, $y = 1, 2$. □

REMARK 1. Note that $q_1 = q_2$ is a degenerate case, since the classifier is redundant and renders no useful information about the outcome of the allocation. Indeed, Bayes' update tells us that, for this case, $P_t^+(y = 1, 2) = \alpha_t$ i.e., the posterior and prior probabilities are the same. Also, $q_1 < q_2$ is the strange case of a counter-indicator (or a reliable "liar") and so the conditions for monotonicity are the same as established before, except that we switch q_1 for q_2 and vice-versa.

2.2. Stochastic Dynamic Programming

Let there be N tasks waiting to be serviced. We use $t = 1, \dots, N$ to indicate the current task index. If task t is completed, the DM will receive a running reward of r_t . Suppose the DM is currently facing task t with k resources in hand. Further suppose that the probability that task t is incomplete be given by $P_t \leq 1$. If the DM just received task t , we would of course have $P_t = 1$. Let $V(t, k|P_t)$ indicate the optimal expected cumulative reward i.e., "payoff to go", that can be achieved thereafter. It follows that $V(t, k|P)$ must satisfy the Bellman recursion:

$$V(t, k|P_t) = \max_{u=0,1} \left\{ u \left(pP_t r_t + \sum_y \text{Prob}(y) V(t, k-1|P_t^+(y)) \right) + (1-u)V(t+1, k|1) \right\}, \tag{7}$$

where the control action, $u = 1$ indicates that an additional resource is to be spent on the current task and $u = 0$ indicates that the DM will move on to the next task in the sequence. If the decision $u = 1$ is taken, the immediate expected reward of $pP_t r_t$ is gained. After each allocation, the DM receives the status report, y . In (7), $\text{Prob}(y)$ is the probability that report y is generated when $u = 1$ is chosen. Indeed, we have:

$$\text{Prob}(y=1) = q_1P_t(1-p) + q_2(1-P_t(1-p)) \text{ and} \tag{8}$$

$$\text{Prob}(y=2) = (1-q_1)P_t(1-p) + (1-q_2)(1-P_t(1-p)). \tag{9}$$

The a posteriori probability that the task is still incomplete after receiving y , is given by $P_t^+(y)$ and is computed, as before, using Bayes' theorem:

$$P_t^+(y=1) = \frac{q_1P_t(1-p)}{\text{Prob}(y=1)} \text{ and} \tag{10}$$

$$P_t^+(y=2) = \frac{(1-q_1)P_t(1-p)}{\text{Prob}(y=2)}. \tag{11}$$

If the DM faces the last task and still has $k > 0$ resources at hand, the expected reward is given by:

$$V(N, k|1) = r_N(1 - q^k), \tag{12}$$

where $q = 1 - p$. In other words, q^k represents the probability that the last task is not completed by any of the k resources. So, $1 - q^k$ is the probability that it gets completed; thereby yielding the reward (12).

The optimal feedback policy is therefore given by:

$$\mu(t, k|P_t) = \arg \max_{u=0,1} \left\{ u \left(pP_t r_t + \sum_y \text{Prob}(y) V(t, k-1|P_t^+(y)) \right) + (1-u) V(t+1, k|1) \right\}. \quad (13)$$

The solution entails solving the Stochastic Dynamic Program (SDP) given by (7). Here, the stage $t = 1, \dots, N$ and the state is (k, P_t) , where k is the number of resources in hand at stage t .

2.3. Dynamic Programming Solution

We note that the Dynamic Programming (DP) equation (7) is unconventional, in that the variable P_t takes continuous values in $(0, 1]$. So, standard solution methods such as value/ policy iteration do not directly apply. However, we point out that for any finite instance of the problem with say M resources and N tasks, one can pre-compute all possible values that the posterior probability, $P_t^+(y_1, \dots, y_n)$ can take as a function of the observation history $y_1, \dots, y_n$ over $n \leq M$ resources allocated to task t . Indeed, let the one-step Bayes update (10) and (11) be denoted by $P_t^+(y=1) = \mathcal{F}_1(P_t)$ and $P_t^+(y=2) = \mathcal{F}_2(P_t)$ respectively. Let us define the indexed set,

$$\begin{aligned} \mathcal{P}_n &= \cup_{i=1,2} \cup_{P \in \mathcal{P}_{n-1}} \mathcal{F}_i(P), \quad n = 1, \dots, (M-1), \\ \mathcal{P}_0 &= \{1\}. \end{aligned} \quad (14)$$

It is clear that the cardinality of $\mathcal{P}_n$ is 2^n . Having precomputed the sets $\mathcal{P}_n$, $\forall n < M$, the DP equation (7) can be restated as follows:

$$\begin{aligned} V(t, k|P) &= \max_{u=0,1} \left\{ u \left(pP r_t + \sum_{i=1}^2 \text{Prob}(i) V(t, k-1|\mathcal{F}_i(P)) \right) + (1-u) V(t+1, k|1) \right\}, \\ k &= 1, \dots, M, \forall P \in \cup_{n=0}^{M-k} \mathcal{P}_n. \end{aligned} \quad (15)$$

So, $V(t, k|P)$ can be computed so long as $V(t, k-1|\mathcal{F}_i(P))$ and $V(t+1, k|1)$ are available. However, $V(t, 1|P)$ depends only on $V(t+1, 1|1)$ since $V(t, 0|\mathcal{F}_i(P)) = 0$. For $k > 1$, we have, $\mathcal{F}_i(P) \in \cup_{n=1}^{M-k+1} \mathcal{P}_n$ since $P \in \cup_{n=0}^{M-k} \mathcal{P}_n$. So, one could solve (15) using backward Dynamic Programming. Indeed, we first compute:

$$V(N, k|1) = r_N(1 - q^k), \quad k = 1, \dots, M. \quad (16)$$

We then proceed to compute $V(t, k|P)$, $\forall P \in \mathcal{P}_{M-k}$ in the decreasing order of t from $N-1$ to 1 and increasing order of k from 1 to M using (15). In essence, the state component P is quantized and as such the DP algorithm deals with a finite discrete state space. However, scalability is an issue since the total number of states is $\mathcal{O}(N \times M \times 2^M)$.

Although interesting in its own right, we shall not dwell any longer on the computational aspects of the DP algorithm. Instead, our focus here is to characterize the optimal solution. As mentioned earlier, under fairly reasonable assumptions, the optimal policy can be shown to have a special structure. Indeed, a resource from an available inventory of k resources is allocated to a task of value r_t iff $pP r_t$ exceeds a threshold. In the next section, we prove that a threshold policy is optimal and furthermore, the threshold is monotonic decreasing in the inventory level.

3. Monotone Threshold Policy

Let $\Delta_{t+1}(k) := V(t+1, k|1) - V(t+1, k-1|1)$ indicate the expected marginal reward yielded by assigning an additional resource over and above $k-1$ resources to the future tasks $t+1$ to N .

PROPOSITION 1. $\Delta_t(k)$ is a monotonically decreasing function of k .

We shall prove the above proposition later. Notice however that the marginal reward yielded by the last task in the sequence,

$$\Delta_N(k) = V(N, k|1) - V(N, k-1|1) = p r_N q^{k-1}, \quad (17)$$

is clearly a decreasing function of k given that $q < 1$.

Suppose Proposition 1 is true i.e., $\Delta_{t+1}(k)$ is a monotonically decreasing function of k . For any $P \in (0, 1]$, we

can define $\kappa_t(P) = \min_{k=1,2,\dots} k$ such that $pPr_t \geq \Delta_{t+1}(k)$. Indeed, $\kappa_t(P)$ is the smallest positive integer such that pPr_t is no smaller than the marginal reward $\Delta_{t+1}(k)$. The following result shows that a thresholding policy is optimal.

LEMMA 1. *If $\Delta_{t+1}(k)$ is a monotonically decreasing function of k , the optimal policy:*

$$\mu(t, k|P_t) = \begin{cases} 0, & k < \kappa_t(P_t), \\ 1, & \text{otherwise.} \end{cases}$$

Proof. From the Bellman recursion (7), we have: $V(t, k|P_t) \geq V(t+1, k|1)$. It follows that:

$$\begin{aligned} pPr_t + \sum_y \text{Prob}(y)V(t, k-1|P_t^+(y)) &\geq pPr_t + V(t+1, k-1|1) \\ &\geq V(t+1, k|1), \forall k \geq \kappa_t(P_t), \end{aligned} \quad (18)$$

where (18) follows from the definition of $\kappa_t(P_t)$. Recall the Bellman recursion (7):

$$\begin{aligned} V(t, k|P_t) &= \max_{u=0,1} \left\{ u \left(pPr_t + \sum_y \text{Prob}(y)V(t, k-1|P_t^+(y)) \right) + (1-u)V(t+1, k|1) \right\}, \\ \Rightarrow V(t, k|P_t) &= pPr_t + \sum_y \text{Prob}(y)V(t, k-1|P_t^+(y)), \quad k \geq \kappa_t(P_t). \end{aligned} \quad (19)$$

$$\Rightarrow \mu(t, k|P_t) = 1, \quad \forall k \geq \kappa_t(P_t). \quad (20)$$

We shall prove the second part of the result, i.e., $\mu(t, k|P_t) = 0, k < \kappa_t(P_t)$ by induction on k . Recall the definition of $\kappa_t(P_t)$, which gives us:

$$pPr_t + V(t+1, k-1|1) < V(t+1, k|1), \quad k < \kappa_t(P_t). \quad (21)$$

If $\kappa_t(P_t) = 1$, there is nothing left to prove. So, suppose $\kappa_t(P_t) > 1$. From the Bellman recursion (7), we have:

$$\begin{aligned} V(t, 1|P_t) &= \max_{u=0,1} \{ upPr_t + (1-u)V(t+1, 1|1) \} \\ &= V(t+1, 1|1), \end{aligned} \quad (22)$$

where (22) follows from (21) since $\kappa_t(P_t) > 1$. Suppose $V(t, h-1|P_t) = V(t+1, h-1|1)$ for some $h < \kappa_t(P_t)$. We have the following monotonicity result (for proof, see Sec 2.1):

$$\begin{aligned} q_2 < q_1 &\leq \frac{1}{1-p} q_2 \\ \Rightarrow P_t^+(y) &\leq P_t, \quad y = 1, 2. \end{aligned} \quad (23)$$

Now $V(t, h-1|P_t) = V(t+1, h-1|1)$ implies that it is optimal to engage the next task in the sequence if the probability that task t is incomplete is P_t and the inventory level is $h-1$. It follows that if the probability that t is incomplete is even lower, the DM must move on to the next task in the sequence and so, $V(t, h-1|P_t^+(y)) = V(t+1, h-1|1), y = 1, 2$. Using the induction argument, the Bellman recursion (7) yields:

$$\begin{aligned} V(t, h|P_t) &= \max_{u=0,1} \left\{ u \left(pPr_t + \sum_y \text{Prob}(y)V(t, h-1|P_t^+(y)) \right) + (1-u)V(t+1, h|1) \right\}, \\ &= \max_{u=0,1} \{ u(pPr_t + V(t+1, h-1|1)) + (1-u)V(t+1, h|1) \}, \\ &= V(t+1, h|1), \end{aligned} \quad (24)$$

where (24) follows from (21) since $h < \kappa_t(P_t)$. Combining (22) and (24), we have:

$$\begin{aligned} V(t, k|P) &= V(t+1, k|1), \quad \forall k < \kappa_t(P), \\ \Rightarrow \mu(t, k|P) &= \begin{cases} 0, & \forall k < \kappa_t(P). \\ 1, & \text{otherwise.} \end{cases} \end{aligned} \quad (25)$$

□

The above result tells us that 1 out of the current inventory of k resources is deployed on the current task iff the immediate expected reward, $pP_t r_t$ is no less than the marginal reward obtained by assigning an additional resource over and above $k - 1$ resources to future tasks. So, a threshold policy is optimal, with the threshold value given by $\Delta_{t+1}(k)$.

REMARK 2. The special case of perfect feedback information is recovered by setting $q_1 = 1$ and $q_2 = 0$. However, Lemma 1 relies on the condition $(1 - p)q_1 \leq q_2$ and therefore the above result does not hold for the perfect information case. A separate analysis is required for this case - see Krishnamoorthy et al. (2016).

THEOREM 2. $\Delta_t(k)$ is monotonically decreasing in the inventory level k .

Proof. We prove the result by induction on the task index, t . From (17), we know that $\Delta_N(k)$ is monotonic decreasing in k . Let us suppose that $\Delta_{t+1}(k)$ is also a decreasing function of k . Combining (19) and (25), the optimal expected cumulative reward takes the form:

$$V(t, k|P_t) = \begin{cases} V(t+1, k|1), & k < \kappa_t(P_t) \\ pP_t r_t + V(t+1, k-1|1), & k = \kappa_t(P_t) \\ pP_t r_t + \sum_y \text{Prob}(y) V(t, k-1|P_t^+(y)), & k > \kappa_t(P_t). \end{cases} \quad (26)$$

By repeated application of (26), we have for $k > \kappa_t(P_t)$:

$$\begin{aligned} V(t, k|P_t) &= pP_t r_t + \sum_y \text{Prob}(y) \left[pP_t^+(y) r_t + \sum_y \dots \right] \\ &= pP_t r_t + p r_t \sum_y \text{Prob}(y) P_t^+(y) + \sum_y \text{Prob}(y) \left[\sum_y \dots \right]. \end{aligned} \quad (27)$$

However, from the Bayes' update (10) and (11), we note that

$$\sum_y \text{Prob}(y) P_t^+(y) = P_t(1 - p) = P_t q. \quad (28)$$

So, we can write:

$$\begin{aligned} V(t, k|P_t) &= pP_t r_t + pP_t r_t q + \sum_y \text{Prob}(y) \left[\sum_y \dots \right], \\ \Rightarrow V(t, k|P_t) &= pP_t r_t \sum_{i=0}^{k-\kappa_t(P_t)} q^i + V(t+1, \kappa_t(P_t) - 1|1), \quad k \geq \kappa_t(P_t). \end{aligned} \quad (29)$$

So, we have the marginal reward given by:

$$\begin{aligned} \Delta_t(k) &= V(t, k|1) - V(t, k-1|1) \\ &= \begin{cases} \Delta_{t+1}(k), & k < \kappa_t(1), \\ p r_t q^{k-\kappa_t(1)}, & k \geq \kappa_t(1). \end{cases} \end{aligned} \quad (30)$$

We proceed to show that $\Delta_t(k)$ as prescribed by (30) is a decreasing function of k . By our induction argument, $\Delta_t(k)$ decreases as k goes from 0 to $\kappa_t(1) - 1$. From the definition of $\kappa_t(1)$, we have:

$$p r_t < \Delta_{t+1}(\kappa_t(1) - 1).$$

For $k \geq \kappa_t(1)$, $\Delta_t(k) = p r_t q^{k-\kappa_t(1)}$ is a decreasing function of k since $q < 1$. □

REMARK 3. Note that, in addition, if the tasks arrive in the order of increasing value, i.e., $r_1 \leq r_2 \leq \dots \leq r_N$, it can be shown that the threshold is also monotonic non-increasing in t . In particular, if all rewards are the same, at least one resource will be allocated to each task.

4. Conclusion

We considered a dynamic resource allocation problem, wherein arriving tasks are sequentially serviced by a DM equipped with homogenous resources. Feedback on the task completion status upon allocation is available

in the form of an error-prone status report. Stochastic Dynamic Programming yields the optimal policy which specifies that a resource is allocated to a task iff the immediate expected reward exceeds a threshold. Under mild conditions on the feedback classifier error probabilities, we show that the threshold is monotonic decreasing in the number of resources. The work is readily applicable to a dynamic weapon target assignment scenario, where an error prone Battle Damage Assessment (BDA) report is available.

References

- Aviv, Y., Kress, M., (1997). Evaluating the effectiveness of shoot-look-shoot tactics in the presence of incomplete damage information. *Military Operations Research* 3 (1), 79–89.
- Aydin, S., Akçay, Y., Karaesmen, F., (2009). On the structural properties of a discrete-time single product revenue management problem. *Operations Research Letters* 37 (4), 273–279.
- Gao, X., Zhong, W., Mei, S., (2013). A game-theory approach to configuration of detection software with decision errors. *Reliability Engineering and System Safety* 119, 35–43.
- Gaver, D. P., Jacobs, P. A., (1997). Probability models for battle damage assessment (simple shoot-look-shoot and beyond). Tech. Report NPS-OR-97-014, Naval Postgraduate School, Monterey, CA.
- Glazebrook, K. D., Washburn, A., (2004). Shoot-look-shoot: A review and extension. *Operations Research* 52 (3), 454–463.
- Hu, X., Xu, M., Xu, S., Zhao, P., (2017). Multiple cyber attacks against a target with observation errors and dependent outcomes: Characterization and optimization. *Reliability Engineering and System Safety* 159, 119–133.
- Kisi, T., (1976). Suboptimal decision rule for attacking targets of opportunity. *Naval Research Logistics* 23 (3), 525–533.
- Krishnamoorthy, K., Casbeer, D., Pachter, M., (2017). Monotone optimal threshold feedback policy for sequential weapon target assignment. *AIAA Journal of Aerospace Information Systems* 14 (1), 68–72.
- Krishnamoorthy, K., Rathinam, S., Casbeer, D., Pachter, M., (2016). Optimal threshold policy for sequential weapon target assignment. In *IFAC Symposium on Automatic Control in Aerospace*. Vol. 49 of IFAC-PapersOnLine. Sherbrooke, Quebec, Canada, pp. 7–10.
- Levitin, G., Hausken, K., (2012). Resource distribution in multiple attacks with imperfect detection of the attack outcome. *Risk Analysis* 32 (2), 304–318.
- Mastran, D. V., Thomas, C. J., (1973). Decision rules for attacking targets of opportunity. *Naval Research Logistics* 20 (4), 661–672.
- Sato, M., (1996). A sequential allocation problem with search cost where the shoot-look-shoot policy is employed. *Journal of the Operations Research Society of Japan* 39 (3), 435–454.
- Sato, M., (1997a). On optimal ammunition usage when hunting fleeing targets. *Probability in the Engineering and Informational Sciences* 11 (1), 49–64.
- Sato, M., (1997b). A stochastic sequential allocation problem where the resources can be replenished. *Journal of the Operations Research Society of Japan* 40 (2), 206–219.
- van Ryzin, G. J., Talluri, K. T., (2005). An Introduction to Revenue Management. *INFORMS TutORials in Operations Research*. Springer Berlin, Ch. 6, pp. 142–194.

Practical uses of the Weibull distribution for decision making

G. Knight B.Eng MSc C.Eng MIMechE

Rosen(UK) Ltd., Floor 2, Q5, Quorum Business Park, Newcastle-Upon-Tyne, NE12 8BS, gknight@rosen-group.com

Abstract. Weibull analysis is often seen as a complicated technique employed by reliability engineering enthusiasts and statistical analysis aficionados, with un-relatable features such as “beta parameters” and “characteristic life”. The array of information that is made available from a quick web search doesn’t do much to convince the casual browser of anything different. What does exist is a dearth of information or guidance on just how useful the Weibull distribution is, therefore no compelling case is made for the casually interested party to look further into the use of this as part of an analysis technique. Analysis using the Weibull distribution include determining an appropriate improvement activity or asset management tactic, assessing test data for product life or supporting investment decisions. This paper gives an overview of the Weibull distribution in plain language, explaining the terminology in relatable terms to the general population. Also covered is the way in which each input variable can be calculated or estimated. Applications for Weibull analysis are explored, with proposals of action based upon the results. The different ways in which the data can be presented is also shown, with key points to consider when proposing asset management activities.

1. A brief introduction to the Weibull distribution

The Weibull distribution is one of the most commonly used statistical models for determining failure patterns. The distribution is useful in terms of describing the behaviour of an item under study i.e. the probability of survival to a certain life, or inversely the likelihood of failure in future.

The Weibull distribution is a continuous distribution and is defined with three parameters. It is acknowledged that these parameters can have different notations depending upon the information source viewed but nonetheless serve the same function. The parameters described in terms of their use in reliability engineering analysis are as follows:

1. shape parameter (β): gives a rough indication of the failure behaviour that can be expected (see Section **Erreur ! Source du renvoi introuvable.**).
2. scale parameter (η): or more descriptively the characteristic life, which is broadly related to the mean time between failures (MTBF) but more accurately it is the time at which 63.2% of the items under study have failed (see Section 2.2).
3. location parameter (μ): used when some guaranteed life is expected.

The parameters are all expressed as a function of a variable t , which typically represents time; although other units may be used, for example number of cycles or loading to failure.

Continuous random variables can be defined using a Probability Density Function (PDF) and a Cumulative Distribution Function (CDF). The PDF can be expressed using all three parameters. However, if the location parameter is taken as zero, this means that the time, t , is the start of the failure prediction and that failures cannot occur before this time. This reduces the PDF and CDF to a two parameter definition given in Equation (1) and Equation (2), respectively:

$$f(t) = (\beta/\eta) \cdot (t/\eta)^{\beta-1} e^{-(t/\eta)^\beta} \quad (1)$$

$$F(t) = 1 - e^{-(t/\eta)^\beta} \quad (2)$$

2. Weibull distribution parameters

The Weibull distribution is used to understand the lifetime characteristics of an item for use in reliability engineering studies, be it an individual component or a complex machine, and hence determining asset management activities. The spread and shape of the distribution gives a useful prediction of the reliability, or failure rate, of an item at a given point in its life and how the susceptibility to failure changes during the lifecycle.

Linear regression is used to establish the Weibull parameters for the item under study, using data collected either through testing or from observations; e.g. from a computerised maintenance management system (CMMS). The two-parameter CDF of a Weibull distribution (Equation (2)) can be rearranged into the format of an equation for a straight line:

$$\ln \left[\ln \left\{ \frac{1}{1-F(t)} \right\} \right] = \beta \ln t - \beta \ln \eta. \quad (3)$$

In plotting $\ln \left[\ln \left\{ \frac{1}{1-F(t)} \right\} \right]$ against $\ln t$ a straight line graph is produced with gradient, β , and intercept, $-\beta \ln \eta$. For example, using a sample, n , of tests of an item to failure, the data can be given a rank, r , in order from shortest to longest time to failure. Then, $F(t)$ is estimated using the median rank method, given by:

$$F(t) = (r - 0.3)/(n + 0.4). \quad (4)$$

A plot, such as that given in Figure 1, can be derived and a straight line fit to the data to establish the gradient and intercept. The shape and scale parameters of the Weibull distribution can then be determined.

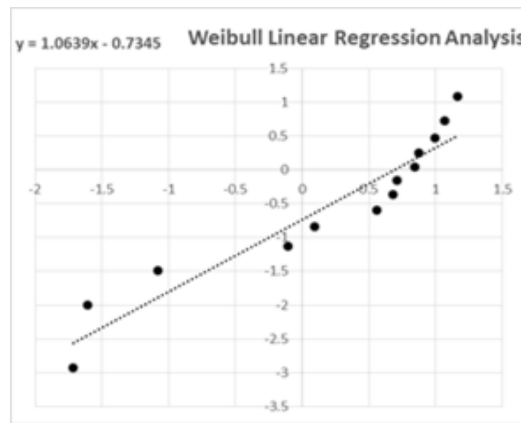


Figure 1. Weibull Linear Regression plot

2.1 The shape parameter as an indicator of failure behaviour

The shape parameter, β , represents the behaviour of the item under study and indicates if the failure rate is increasing, constant or decreasing. This is useful to categorise items into early life (or infant mortality), failure and wear out (or end of life) failure patterns, where:

$\beta < 1$ represents early life failure; $\beta = 1$ represents random failure; and, $\beta > 1$ represents wear out behaviour.

An example of the PDF and CDF plots obtained for differing shape parameters, given a consistent scale parameter, is shown in Figure 2.

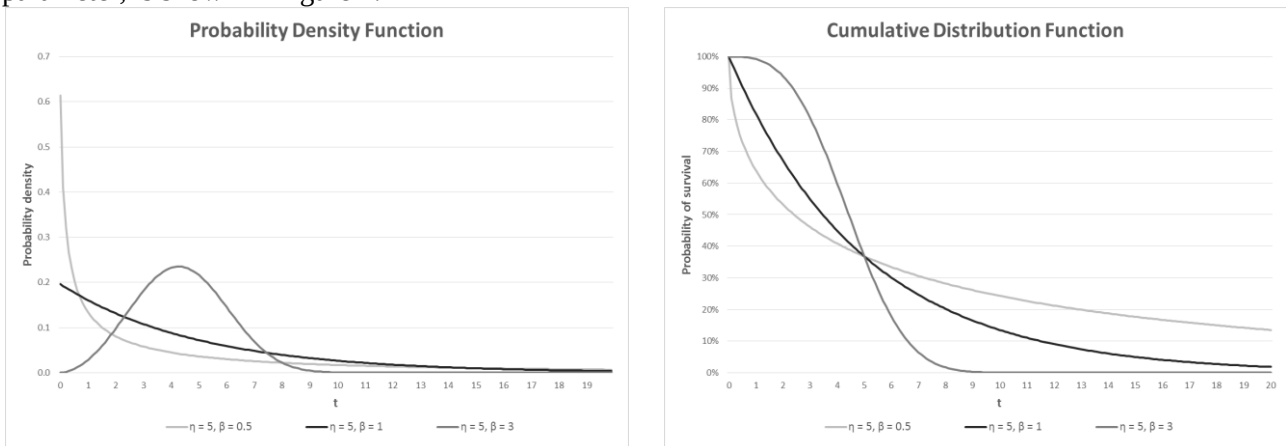


Figure 2. PDF and CDF for $\beta = 0.5, 1$ and 3

The linear regression plots for varying β values can be likened to the failure patterns proposed by the Reliability Centred Maintenance (RCM) methodology (Moubray), as shown in Figure 3.

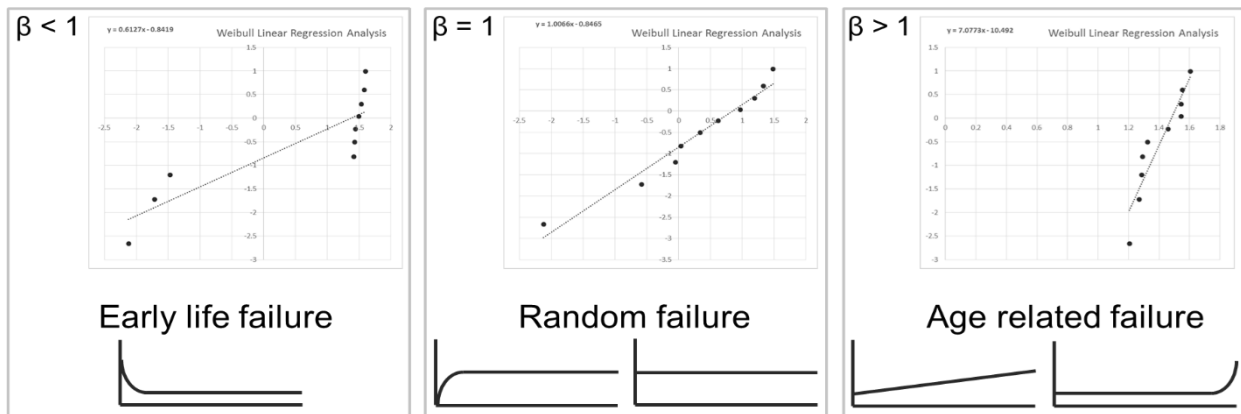


Figure 3 – Weibull distribution β parameters related to RCM failure patterns

Another failure pattern which may be observed is known as the ‘bathtub curve’, whereby three distinct slopes are evident in the linear regression plot, as shown in Figure 4.

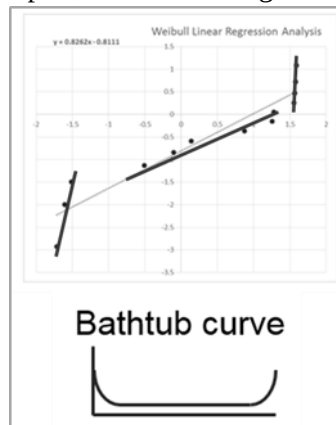


Figure 4 – Linear regression plot showing the bathtub curve failure pattern

The bathtub failure pattern presents a unique challenge in analysis as the failure behaviour would likely be represented with $\beta \approx 1$. If only the value of the shape parameter is considered, this may lead the analyst to believe that random failure is apparent when in truth the failure pattern may actually be pseudo-random. Therefore, it is important to review the linear regression plot to help understand the type of failure behaviour.

For example, consider the failure information which may be available related to a pump. Just having the times to failure tells us that the overall performance of the pump offers the opportunity to be improved. However, if data on the exact failure modes is reviewed this can help with identifying specific performance improvement activities. Taking into consideration each section on the bathtub curve, it was discovered that the failures with the lowest time to failure were due to misalignment, mechanical looseness in the mounting and a faulty mechanical seal. All of these infer problems with build quality which is a typical early life failure root cause. Improvements which may be considered include enhanced training and quality control measures. The failures on the random section of the linear regression plot were found to be poor lubrication practices and start up / shut down practices. These can be addressed through improving operational discipline. The end of life failures were discovered to be due to impeller wear, which might be acceptable from a design life point of view but could offer an opportunity for a design upgrade to extend the pump life.

2.2 The scale parameter as an indicator of characteristic life

The scale parameter, η , is used in reliability analysis to determine the intersection whereby 63.2% of failures have occurred. This allows an assessment of performance between different items given a similar β value. The significance of 63.2% on the Weibull distribution is this is the point at which the characteristic life is truly equal to MTBF when $\beta = 1$ (Warwick.ac.uk).

3 Available data to derive input parameters

As with any analysis the quality and availability of data is key. In reliability analysis, as much relevant information should be gathered regarding the failure instance, such as the reason for failure (the failure mechanism), the configuration / style / model / material of the item, the date / time of failure, ambient conditions, location etc. Input data which is validated either from testing or from other data gathering methods, such as product warranty or a well-managed CMMS, is considered to provide the most robust assessments.

3.1 Maintenance Management Systems

The way in which assets are identified and structured within any maintenance management system is probably the most critical to the ability to analyse in-service items. Within a CMMS, it must be ensured that asset tags are consistent and that fields are used to identify equipment by type (i.e. pump, compressor), duty, location, configuration etc. All failures should be logged within the system, as should any other maintenance activities that may “reset the clock”, such as replacement of components subject to wear or fatigue. This will allow the analyst to extract the data and carry out at least the most basic of failure assessments.

The majority of modern CMMSs allow the entry of fault and cause codes when a fault is logged against an equipment. Consistent use of this functionality will allow the analyst to draw deeper conclusions.

There are a number of published standards which offer guidance on how to structure systems and data to allow the capture and interrogation for maintenance purposes. One such standard is ISO 14224, developed for the petrochemical industry, but which offers a good structure that might be applied to most situations.

3.2 Highly accelerated life testing (HALT)

HALT is a widely used technique for demonstrating an items resistance to predicted environmental stresses such as loading, vibration, temperature, voltage and moisture. The application of the technique is dependent upon the item type and the performance standard required.

A company developing items and using HALT techniques presents the ideal opportunity for the collection of data and analysis using the Weibull distribution. Engineers who are involved in developing and deploying HALT are well placed to get the maximum benefit from the use of a Weibull based analysis to present either detailed information to designers on how to modify the items, or to end users who can incorporate the knowledge into their own Reliability, Availability, Maintenance (RAM) simulations (see Section 4.4 for further details on RAM) .

3.3 Alternative Data Sources

Where data from testing or say a CMMS is not available, does not exist or cannot be fully trusted, alternative data sources are required. The experience of experts may be used and by asking the right questions, this may not only lead to an average life being suggested (which can be approximated to η) but also a failure pattern. An example conversation may go:

“So how often do these pumps fail?”

“I would say about every two years give or take a few months”

“ Every time a pump is installed, does it always make it to two years?”

“Generally yes, but over the fifteen years’ experience I have had here a handful have failed in the first few months”

The failure behaviour suggested by this response could lead the analyst to use a β of between 0.5 and 0.7 (some evidence of early life failures) and a η of somewhere between 1.75 and 2.25 years (for an average time to failure of around two years). The usefulness of the estimation would depend upon the need for the data to incorporate into the analysis. For example, if the purpose of the analysis were to determine whether effort should be expended upon improving maintenance practices (reducing early hour failures) or to replace with a more robust item (improving characteristic life) then “what if” analyses can be carried out. Such analyses help determine the most appropriate course of action by comparing the cost of the improvements vs the expected return, either in improved uptime, reduced maintenance costs or a combination of the two.

Next in the hierarchy, vendor data may also be available. The validity of this data must be assessed as it may not be representative of the operational environment of the item. Vendors may provide data based upon warranty performance, or empirically derived data from similar components. The ability to have a discussion

with a vendor as to the performance of their items, backed with data and analysis, is a very powerful way to ensure that operating needs are met.

Finally, published data sources are available such as the Offshore and Onshore Reliability Data (OREDA) handbook for the oil and gas industry and the FIDES (Latin: Trust) guide for electronic systems and components. It is unlikely that these data sources will be truly representative of the unique operating context of the item to be assessed and there may also only be limited population that the data was drawn from. Therefore, thought must be given as to why the data is being used and the manner in which it is being applied. Results from studies using published data should be used with caution, as there is increased uncertainty in this as an input data source.

4. Using the analysis to determine appropriate actions

If quality data is collected, then the analyst will be able to draw better conclusions and propose appropriate actions from the assessment. A common technique deployed to investigate and propose corrective actions is Root Cause Analysis (RCA). RCA using the Weibull distribution is an extremely useful tool, as it can be used to validate any proposed failure mechanisms as part of the investigation. Understanding the exact reason for failure allows control and mitigation measures to be implemented that target the true cause therefore preventing recurrence of the problem.

4.1 Quality assurance and control

Early life failures are typically due to quality assurance and control issues. Technical and performance standards may include maintenance or operational tasks which were not completed effectively which led to a component that has failed unexpectedly or material that does not conform to specification.

Where there has been a failing in task execution, consideration may be given to reviewing procedures and processes. Robust procedures that are followed are a cornerstone for reliability whereby operational and maintenance interventions are inevitable.

Should a component not conform to standard, then the RCA will likely head into the supply chain. This includes a review of the steps an organisation takes to ensure the quality of incoming goods such as inspections, vendor audits and factory acceptance tests. Whereby components may fail unexpectedly during early life, but then conform to a random failure pattern, Highly Accelerated Stress Screening (HASS) may be an applicable technique. This testing technique is designed to subject items to higher than normal operational stresses (but which are still within the components design capability) in order to weed out any defective items prior to supply to the customer. Attention should also be paid to component shelf life, as the reason for failure may actually be deterioration during storage as opposed to true early life failure.

Material failures may head even further into the supply chain, looking into the raw materials and their processing methods. In principle, each step of the supply chain should be assuring the as-supplied material, but the organisation who actually suffers the effects of a failing may be several steps removed. For highly critical components, such as those used within the nuclear and aerospace industry, such far reaching traceability of components may be absolutely necessary for critical items.

4.2 Replacement / remediation

Whereby a definite end of life behaviour is exhibited, then an ideal opportunity is presented to carry out planned replacement or remediation maintenance activities. The time (or life) at which items should be replaced can be determined from the Weibull distribution CDF. In order to determine the intervention time, an appropriate failure probability limit would be assigned. For example, should there be a desire to accept no more than 0.1% probability (or once every thousand years) that a failure would occur then the intervention activity would be undertaken at a time before this limit is reached.

The timeframe for intervention may be read straight from the CDF plot at the desired survival probability level, 99.9% in this case. An example is shown in Figure 5, which would result in a replacement task within two years. Consideration must be given to how steeply the probability of failure increases over time. Whereby a failure may lead to severe safety concerns then the probability of failure using the Weibull distribution analysis is used in risk assessment.

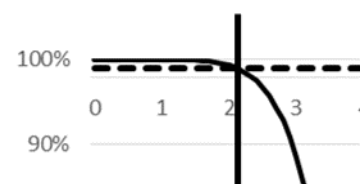


Figure 5. - 99.9% reliability

4.3 Condition monitoring

Pure random failure behaviour lends itself to condition monitoring, as there is no benefit to improving initial installation practices or assigning a retirement / decommissioning age. However, where limited windows for maintenance are available, tackling items that exhibit random failure on such an ad hoc basis may not be desirable. Understanding the failure characteristics allows residual risk to be expressed. The analyst may recommend the replacement of such components if the risk is deemed unacceptable. Very few items in truth conform to pure random failure, many more fall into the infant mortality pattern. The best decision may be to leave well alone; acknowledge the risk and plan for a failure should it occur.

4.4 Design - Reliability, Availability & Maintainability (RAM) Simulation

RAM simulation is used during the design of complex processes and systems in order to determine the overall performance over time. Its greatest use is as a comparator of competing designs, allowing decisions to be made on an optimum configuration.

A typical RAM simulation starts with the development of a base case, which is the base system design represented using a reliability block diagram to illustrate the reliability dependencies of the system. Each block in the diagram represents an item, which may be a component, a piece of equipment, a subsystem or even (during concept stage for mega-projects) entire systems or plants. The anticipated performance of each item in terms of a failure rate and a repair time can be expressed in terms of Weibull distribution parameters. Sensitivity cases can be developed, changing system configurations and/or the base data to allow an optimum design to be determined. The β and η parameters can be adjusted and the simulation run to determine if there will be any detrimental effect upon lifecycle performance. The validity of the output of RAM simulation is heavily dependent upon the data used. By virtue of the fact that the item under study, or process/system in design, is early in the lifecycle then it is highly likely that failure data will not currently exist to demonstrate behaviour. Generally, published data sources are used unless a company has its own data set.

Summary

The practical deployment of the Weibull analysis technique has been presented, along with advice that anyone working within an engineering environment can use to improve performance. Gaining an understanding of failure behaviour leads to robust engineering solutions in a much faster timeframe than by, say, experimenting and hoping for the best. The author encourages anyone involved in reliability (and indeed any field of) engineering to become familiar with the technique and try it – you will undoubtedly learn at least something.

Acknowledgements

Many thanks to Katherine Taylor for keeping me correct with the mathematics, and contributing in an editorial and technical capacity to this paper.

References

- Moubray, John (1997). *Reliability Centred Maintenance, Second Edition*. Butterworth-Heinemann (Elsevier)
- Warwick Manufacturing Group (2004). *The Use of Weibull In Defect Data Analysis* [online] Available at: <https://warwick.ac.uk/fac/sci/wmg/ftmsc/modules/modulelist/rdd/temp/weibull.pdf> [Accessed 15/04/2018].
- FIDES Group (2010). *Reliability Methodology for Electronic Systems* [online] Available at: http://www.fides-reliability.org/files/UTE_Guide_FIDES_2009_Ed_A_EN.pdf [Accessed 15/04/2018].
- International Standards Organization (2016). *ISO 14224: Petroleum, petrochemical and natural gas industries -- Collection and exchange of reliability and maintenance data for equipment*. Available at: <https://www.iso.org> [Accessed: 15/04/2018].

A pitch control system deterioration model considering wind speed turbulence intensity

Jinrui MA*, Mitra FOULADIRAD*, Antoine GRALL*

* ICD-LM2S, universit  de technologie de Troyes, 12 rue Marie Curie 10004 Troyes France

email address: jinrui.ma@utt.fr; mitra.fouladirad@utt.fr; antoine.grall@utt.fr

Abstract As a consequence of high failure rate occurring to wind turbines (WTs), more and more asset owners begin to pay attention to WT's maintenance planning. To implement an efficient condition-based maintenance policy, the deterioration modeling is of primary importance.

The inherence character of wind makes WT's operation full of uncertainties, hence, the pitch actuator movement is random and it can be modeled by a stochastic process. Meanwhile, the deterioration of the pitch control system related to the usage of the actuator is random. When the wind speed is higher than the rated wind speed (a design WT parameter, when wind speed is higher than the rated wind speed, the pitch system begin to operate) and the wind speed turbulence intensity (TI) is very important, the deterioration rate of the pitch control system increases significantly. For instance the pitch system has more failures in a windy season. Therefore, in order to propose an appropriate maintenance policy for pitch control system, it is more sensible to take into account the influence of wind speed in deterioration modelling.

In this paper, we focus on the deterioration of hydraulic pitch control system. After studying the factors influencing the deterioration, a stochastic process considering wind speed as covariates is used to model the deterioration. A maintenance policy based on an alarm threshold is considered.

1. Introduction

A WT is expected to operate at least 20 years. In reality, divers subsystem failures cause undesirable downtime of WT (P rez et al. (2013)). Moreover, pitch system has a higher failure rate compared with other subsystems. For a material, a component or a system, evolves from the healthy state to the failure. For example, the oil deterioration in a hydraulic system is a long-term process which associates with gradually accumulated metal particles, increased air/oil ratio and entrained water. In a hydraulic control system, physical deterioration brings unsatisfactory control response leading to loss efficiency. With the bigger scale of WT, off-shore location of wind farms and uncertainty operation environment, Prognostics and Health Management (PHM) of WT attract many research interests. To optimize the operation cost, to predict reliable deterioration level and to provide low-cost maintenance policy, studying the deterioration of WT components is necessary and can meet the wind power market requirements. Besides, the Supervisory Control And Data Acquisition (SCADA) system installed on each WT and used for Condition Based Maintenance (CBM), collects a massive source of WT operation data which provide the practicability of component's deterioration estimation and its Remain Useful Life (RUL) prognosis.

The aim of this paper is to study the pitch system deterioration by using the operation data. The wind influence on the deterioration is another interest of this paper. A brief introduction about the pitch controller and the hydraulic pitch actuator is arranged in section 2; Section 3 discusses the wind influence on the deterioration of pitch system followed by the deterioration model in section 4. Section 5 briefly introduces the RUL prediction and the maintenance policy. The conclusion is organized in the end.

2. Pitch controller and hydraulic pitch actuator

2.1. Pitch controller

The PI controller shown in Figure 1 is realized in SIMULINK environment. The error between the rated generator speed and the filtered generator speed allows the PI pitch controller to work out a satisfactory pitch angle implemented by hydraulic actuators. The gain factor is dependent on the blade pitch angle. The pitch rate is limited to 8  /s in absolute value, the minimum and maximum pitch settings are 0  and 90 , respectively. These values' setting is based on General Electric (GE) Wind's long-blade test program. Table 1 lists the pitch control properties. This pitch controller is used as collective pitch control.

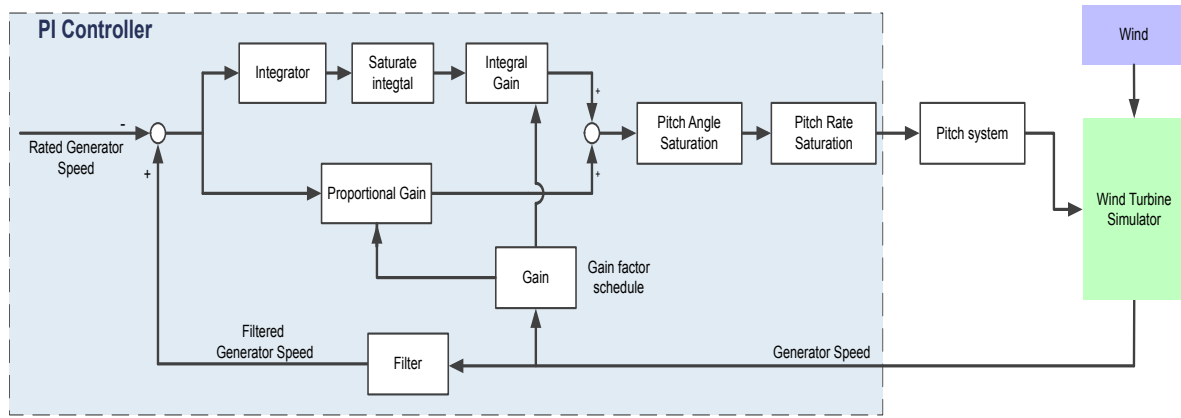


Figure1. Wind turbine simulator

Blade-Pitch Minimum Setting	0°
Blade-Pitch Maximum Setting	90 °
Maximum Absolute Blade Pitch Rate	8 °/s
Proportional Gain at Blade-Pitch Minimum Setting	0.01882681 s
Integral Gain at Blade-Pitch Minimum Setting	0.008068634
Rated Rotor Rotational Speed	12.1 rpm
Rated Generator Rotational Speed	1173.7 rpm

Table 1. Parameters of the pitch control system (Source: Jonkman et al. (2009))

2.2. Hydraulic pitch system

Hydraulic pitch system has the advantages letting it widely used in the commercial market, such as simple condition monitoring, emergency stop without electrical power, good electrical insulative and stable performance.

The considered collective pitch control scheme has a single reference pitch angle input acting on the pitch hydraulic system to implement the desired pitch angle. In principle, the fault-free hydraulic pitch system is a piston servo system which can be modeled by a second-order dynamic equation as follow

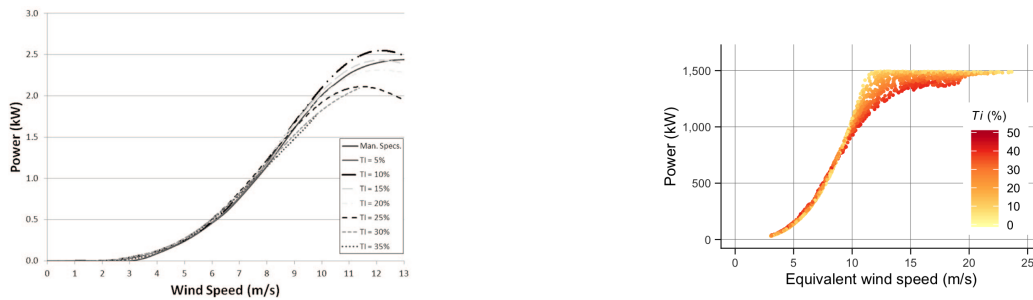
$$\ddot{\beta} + 2\zeta\omega_n\dot{\beta} + \omega_n^2\beta = \omega_n^2\beta_r \quad (1)$$

where β is the blade-pitch angle measurement, β_r is the reference blade-pitch angle from pitch control system, ω_n is the natural frequency, and ζ is the damping ratio. In the case of no deterioration occurring to the actuator, the following parameters are used: $\zeta = 0.6$, $\omega_n = 11.11 \text{ rad/s}$ (Odgaard et al. (2013)).

3. Wind influence

SCADA system notes 10 minutes average wind speed value, and it neglects wind speed turbulence. However, the same average wind speed with different turbulence intensity affects differently the WT out-put power (Clifton and Wagner (2014); Ward and Stewart (2015)). The essential reason is that with the same average wind speed, pitch operations are different relating to the various turbulence intensity at full load region.

Tavner et al. (2006) concerned with the influence of wind speed on the reliability of WT. This research quantifies the wind speed data as Wind Energy Index (WEI) which is defined as



(a) Power curves adjusted to different TI based on experimentally observed data (Source :Ward and Stewart (2015)) Clifton and Wagner (2014)) (b) Power curve of the WindPACT 1500 kW WT (Source: Clifton and Wagner (2014))

Figure2. Variation of out-put power with different turbulence intensity

$$WEI = \frac{\text{Actual monthly energy production from a collection of wind turbines}}{\text{Long term expected monthly energy production from those turbines in the presence of average weather}}$$

Figure 3 shows the relationship between the failure rate and WEI. From this figure and the research result of Tavner et al. (2013), we can conclude that weather and wind speed have significant influence on WT components deterioration and failure. Moreover, Tavner et al. (2006) points out that some WT components are more affected by wind speed than others, such as hydraulic system, generator, yaw control and mechanical brake (shown in Figure 4). In the opinion of the author, the reason is that these components are not designed with the rapidly changing effects of the wind speed variation. A Chinese report about WT pitch failure affirms that most pitch system failure occur in windy season as a consequence of high wind speed variation Lin (2013).

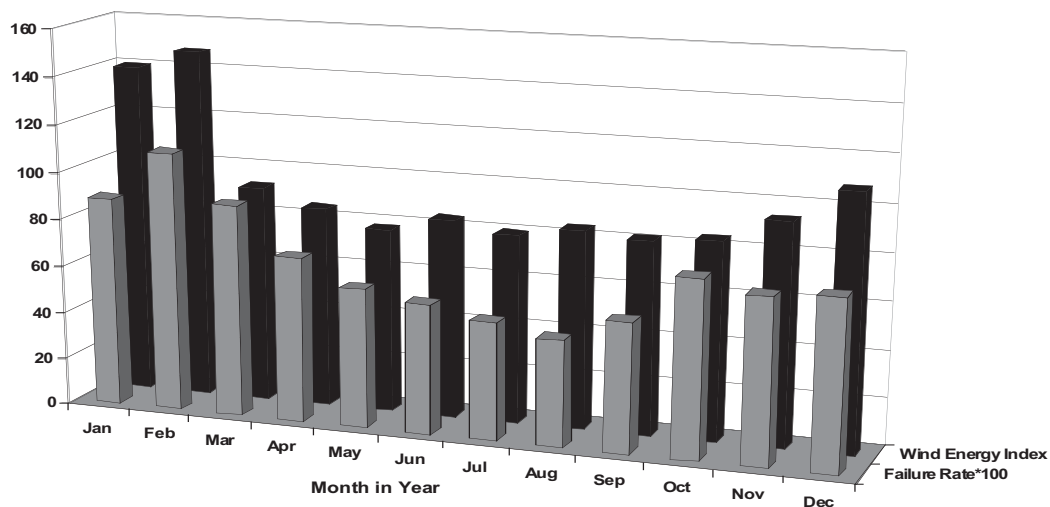


Figure3. Average monthly Failure Rate and WEI for each of the 12 months over the Survey period 1994-2004(source: Tavner et al. (2006))

In this research, we mainly consider the influence of wind speed turbulence. When wind speed is higher than the rated wind speed, small velocity changes from wind lead to the sensible action of pitch system. Combining the fact of high pitch system's failure rate in windy season, one can deduce, the variable wind accelerates the deterioration of pitch system. In other words, wind speed turbulence is the main factor causing actuator deterioration. Frequent usage of the actuator in a windy season is the reason for its failure. Therefore, we assume that for a pitch system frequent pitch actions caused by high wind speed turbulence accelerate the former's deterioration and that the pitch system only deteriorates when it implements the pitch requirements. This opinion has double meanings, firstly, WT's operation in partial load region doesn't cause the pitch actuator's deterioration

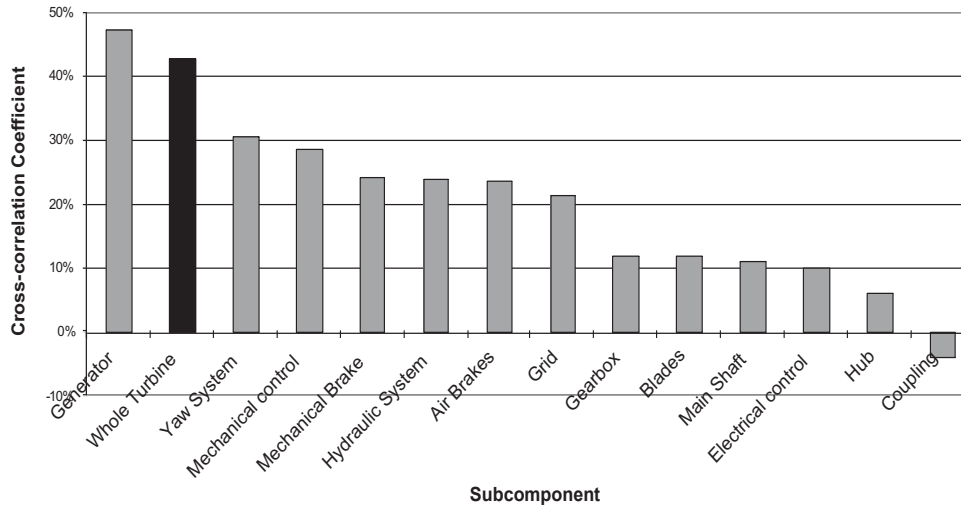


Figure4. Summary of cross-correlograms of subcomponent failure rates to WEI, 1994-2004 (source: Tavner et al. (2006))

as the pitch system is not in service. Secondly, in full load region different wind speed turbulence causing different deterioration. However, wind speed is random, how to measure the deterioration in a specific period?

4. Pitch system deterioration model considering the wind influence

4.1. Deterioration model

The available operation data relating to pitch system are pitch angle and pitch rate. To quickly adjust the generator rotational speed and the one of the rotor over the change of wind speed, current pitch rate can reflect the current wind speed turbulence. Hence we introduce the Accumulate Pitch Rate (APR) during one minute as a parameter used to represent the wind influence.

$$APR_i = \int_{t=i*60-60}^{i*60} B_t dt \quad (i = 1, 2, \dots, n) \quad (2)$$

APR_i is the i_{th} minute APR, B_t is the pitch rate at time t , the integral is carried out in the second time scale, please pay attention to the units (second, minute).

The usage of pitch actuator, which is quantified as APR_i , is considered slowly increasing the air/oil ratio of hydraulic pitch system. And the latter is assumed as monotonously increasing before maintenance. The increment of air content at time t_{i+1} only depends on the one at time t_i . As the intrinsic randomness of wind, the operation of pitch system is stochastic. Hence, for a given time period — one minute, APR_i is a random value and $APR_i \in \mathbb{R}^*$ ($i = 1, 2, \dots, n$) (an example is shown in Figure 5). The deterioration increment distribution at time t could be defined as follows:

$$f(x|g(APR)t, b) = \frac{1}{b^{g(APR)t} \Gamma(g(APR)t)} x^{g(APR)t-1} e^{-\frac{1}{b}x} \quad (3)$$

where, $g(APR)$ is a function related to APR, the value is $APR_{\lfloor t/60 \rfloor}$, b is constant.

4.2. Health indicator

Since the reference pitch angle (β_{ref}) and measured pitch angle (β_{mes}) can be easily achieved from SCADA, they are used to estimate the parameters of the transfer function, the estimated value of natural frequency ($\hat{\omega}$) is the indicator for the hydraulic pitch system deterioration.

Taking into account of the SCADA data acquisition frequency and the actual long-term deterioration process, $\hat{\omega}$ is estimated for every ten minutes. Hence, the estimated natural frequencies constitute a time series $\hat{\omega}\{i\}$ ($i = 1, 2, \dots, n$). If the latest $\hat{\omega}$ is smaller than 3.42 rad/s, the hydraulic pitch system is considered as failed.

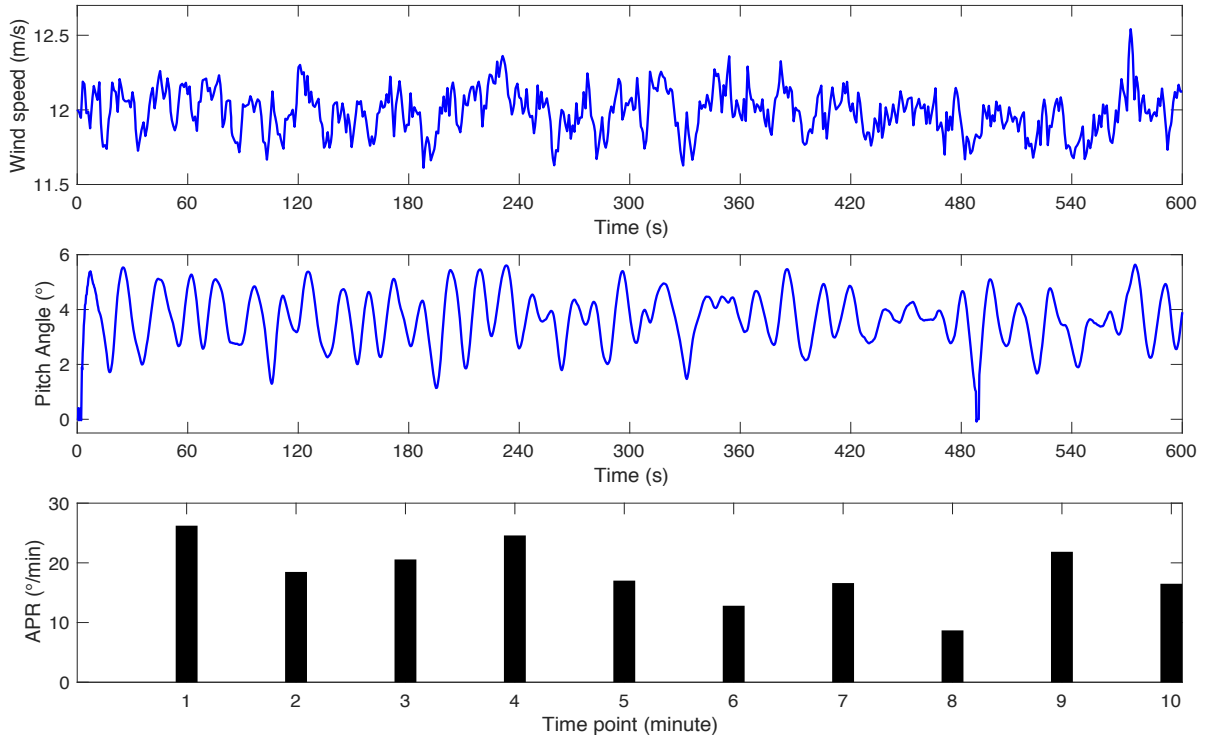


Figure5. A sample for APR value

5. RUL prediction and maintenance

The failure time T_F is the time when deterioration process reaches at the failure threshold L . $RUL(T_i)$ is the remaining useful life of the hydraulic pitch system at inspection time T_i . However, the $RUL(T_i)$ is a random value, more precisely, at time T_i , it follows a probability distribution defined as below:

$$\begin{aligned}
 F_{RUL(T_i)}(\tau) &= \mathbb{P}(RUL(T_i) < \tau) = \mathbb{P}(\hat{\omega}(T_i + \tau) < L | \hat{\omega}(T_i) = \omega_i) \\
 &= \mathbb{P}(\hat{\omega}(T_i + \tau) - \hat{\omega}(T_i) < L - \hat{\omega}(T_i)) \\
 &= \frac{\Gamma(a(T_i + \tau) - a(T_i), (L - \omega_i)\beta)}{\Gamma(a(T_i + \tau) - a(T_i))}
 \end{aligned} \tag{4}$$

where $a(x) = \alpha x$ and $\Gamma(m, n) = \int_{z=n}^{\infty} z^{m-1} \exp^{-z} dz$ is the incomplete gamma function.

The RUL estimation method can refer to (Moore (1982)).

The maintenance decision is based on the RUL estimation. For the role played by the pitch system in a WT, predictive maintenance is a good choice (Langeron et al. (2016), Le Son et al. (2016)). Le Son et al. (2016) propose an estimated-RUL-based maintenance policy which takes the estimation RUL into account. This policy defines a RUL_{min} as a fixed threshold, comparing the RUL_{min} with the estimated RUL $E(RUL(i))$ at each inspection time T_i , ω_i is the estimated natural frequency at time T_i

- if $\omega_i < L$ and $E(RUL(i)) \leq RUL_{min}$, the pitch system is preventively repaired.
- if $\omega_i < L$ and $E(RUL(i)) > RUL_{min}$, the maintenance decision is postponed until next inspection T_{i+1} .
- if $\omega_i \geq L$, the pitch system has failed, a corrective repair should be carry out.

6. Conclusion

A deterioration model of a hydraulic pitch system considering the wind influence is discussed in this paper. Future research will focus on the relationship between the pitch system deterioration and different wind profiles. The online RUL prediction and the maintenance policy will be deeply study.

References

- Clifton, A. & R. Wagner (2014). Accounting for the effect of turbulence on wind turbine power curves. In *Journal of Physics: Conference Series*, Volume 524, pp. 012109. IOP Publishing.
- Jonkman, J., S. Butterfield, W. Musial, & G. Scott (2009). Definition of a 5-mw reference wind turbine for offshore system development. Technical report, National Renewable Energy Lab.(NREL), Golden, CO (United States).
- Langeron, Y., M. Fouladirad, & A. Grall (2016). Controlled systems, failure prediction and maintenance. *IFAC-PapersOnLine* 49(12), 805–808.
- Le Son, K., M. Fouladirad, & A. Barros (2016). Remaining useful lifetime estimation and noisy gamma deterioration process. *Reliability Engineering & System Safety* 149, 76–87.
- Lin, L. (2013). S11500 series wind turbine pitch system failure analysis. *China Wind Power Production and Operation Management (2013) (in Chinese)*.
- Moore, R. (1982). Algorithm as 187: Derivatives of the incomplete gamma integral. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 31(3), 330–335.
- Odgaard, P. F., J. Stoustrup, & M. Kinnaert (2013). Fault-tolerant control of wind turbines: A benchmark model. *IEEE Transactions on Control Systems Technology* 21(4), 1168–1182.
- Pérez, J. M. P., F. P. G. Márquez, A. Tobias, & M. Papaelias (2013). Wind turbine reliability analysis. *Renewable and Sustainable Energy Reviews* 23, 463–472.
- Tavner, P., C. Edwards, A. Brinkman, & F. Spinato (2006). Influence of wind speed on wind turbine reliability. *Wind Engineering* 30(1), 55–72.
- Tavner, P., D. Greenwood, M. Whittle, R. Gindele, S. Faulstich, & B. Hahn (2013). Study of weather and location effects on wind turbine failure rates. *Wind Energy* 16(2), 175–187.
- Ward, N. J. & S. W. Stewart (2015). A turbulence intensity similarity distribution for evaluating the performance of a small wind turbine in turbulent wind regimes. *Wind Engineering* 39(6), 661–673.

Optimising remedial outcomes for Gas Turbines through large scale data analysis

Jason McGinty*, Stefanos Kollias**, Georgios Leontidis***

*Machine Learning Research Group, School of Computer Science, University of Lincoln, Brayford Pool, LN6 7TS, Lincoln, UK, * 16654927@students.lincoln.ac.uk / Jason.McGinty@siemens.com, ** skollias@lincoln.ac.uk, *** gleontidis@lincoln.ac.uk*

Abstract. An investigation into approaches to model and predict the costs, risks and outcomes, relating to a common failure mode within a large population of remotely monitored engines will be presented. This investigation will cover the relevant aspects of lifecycle management, customer operating cycle, visibility of classified issues remotely using low sample rate remote monitoring systems and identification of the most appropriate repair regime. Unlike previously identified studies, the datasets used are larger comprising sensor readings from over a 120 gas turbine units (distributed across the world) which total 100 million observations taken across 117 parameters. These are cross referenced with relevant service events (preventative, corrective, fault investigation and observational reports) and service execution data (spares and manpower required) gathered over a period of up-to 11 years. We have now succeeded in identifying behaviours in a particular problem domain using techniques such as K-means, that indicate non-optimal decisions may have been made (with the benefit of hindsight). This indicates that the intended outcome of the overall research to produce of quantitative models of the different strategies and applying them to optimize for the best outcome both in terms of and customer and supplier objectives in a justifiable form is possible.

1. Introduction

In the world of industrial power generation (and other industrial sectors) there has been a shift in business models from Original Equipment Manufacturers (OEM) with the ongoing service, corrective and upgrades portions of the business being seen as the biggest drivers of growth rather than the traditional staple of new unit sales (Drugugbo, 2016).

This has required some re-alignment by affected OEM's to change their product mix and find ways to competitively service their customer base. One of the ways OEM's have hoped to achieve this is through the use of remote diagnostic systems and services in order to have visibility of issues that occur at customers sites. Traditionally these services have been usually simply for post-mortem or for the use identification of individual issues.

Remote monitoring of Gas Turbines is still in many respects an art form, rather than a true science. Due to the complexity of the system, relatively high cost (in capital, maintenance, running and potential loss of revenue) and serious consequences of a wrong decision (such as loss of life), is still very manually controlled.

We have however a large quantity of information that has been collected over the years, and now with the gradual digitalization of other processes it become possible to automatically add context to much of the information that we have been collecting. This leads to us now being able to review historic data in a new light that was not possible potentially at the time, and analysis the effectiveness of the repair strategies used the processes of labelling and annotating a large dataset (comprising a total of 11 years of operating data from up to 120 engines, not all engines we in operation or connected for the duration of that period), this is taken) is still ongoing, and involved. The data primarily contains information provided by customers who are under the various long- term agreements, this means there is considerably more contextual information available in the form of site visit reports, parts orders and maintenance history to annotate the collected sensor data with.

A number of papers (such as those looking at machine fault detection (Zhang , 2014)), or signal reconstruction to improve availability (Yang, 2013) have been produced by various research groups in relation to the same core dataset, however their purview has been quite narrow in terms of the scope of the project regarding analysis of individual failures modes in a very specific (and often novel way), without looking at the context of the surrounding remedial actions that would need to be taking place.

This paper presents an overview of the direction in which the ongoing research is taking, illustrated by a sample of the data currently available. This for reasons of commercial sensitivity the data and the user, product and its technical features has been obfuscated.

2. Research Goal

With the rapidly increasing availability of usable information would it be possible to create a self-optimizing decision support system that would improve the consistency and quality of our support in major events domain of faults which are addressing are the areas where there are two (potential competing) disruptive and relatively costly remedies that could result in a fix for an issue, but which the chances of success are not guaranteed. These issues generally occur when there are major subcomponents that can be changed out in lieu of a full engine exchange, but could trade off longer term reliability or disruption with the customer if they were to fail.

It is also hoped that the work will allow for the integration of approaches to deal with No Fault Found conditions occurring which are become a major focus of cost reduction across affected industries (Khan, 2016), and in this domain results in large number of resources being used to investigate and replace components potentially unnecessarily.

3. Current Approach and Methods

The current core focus of the research at this point is to create the pipeline and infrastructure shown in Figure 1. This illustrates the high-level structure of the system and the data-sources that are available for the decision-making algorithms. It is novel due to the extent of cross-use of information between the parties. This is analogous to the system that was presented at the preceding conference in this series by Animah (Animah, 2016) in their description of a novel system for managing life extension . however it is focused on automatic feature extraction rather than expert driven constraints collection.

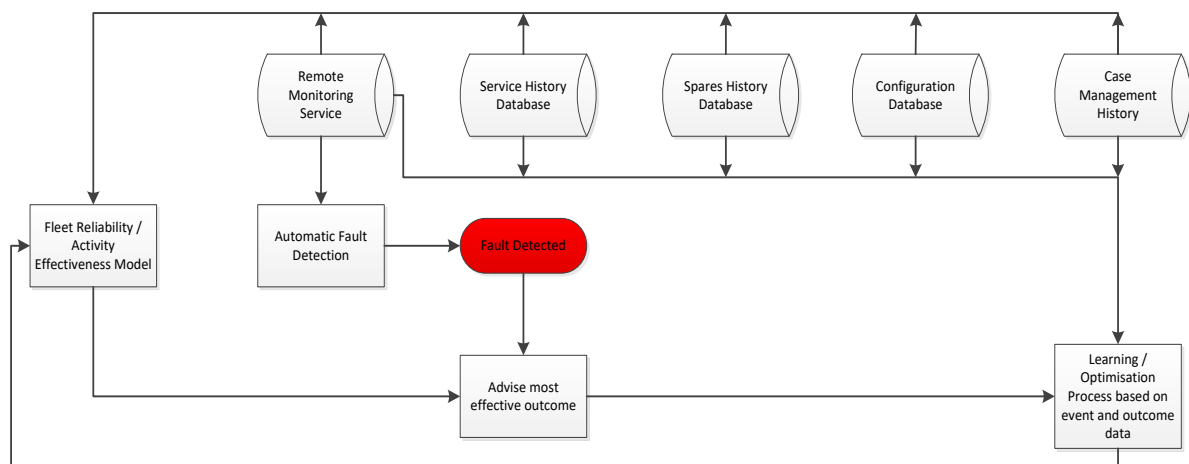


Figure 1: High-level overview of proposed system.

4. The initial sample problem

The fault which has been selected as the basis for the initial study has been selected from a set of common issues that have a high impact upon the availability of the engine and which can be seen remotely via telemetry. For the purposes of this investigation the problem selected is a condition which can occur in which the, there are two potential options for remediation, the first is that Gas Turbine can be repaired at site for a moderate cost in terms of material and manpower, but more time in operation and with an unknown chance of success (Option 1) or the Gas Turbine can be replaced wholesale at a much greater cost, but with much less downtime for the customer (Option 2). There have been a number of occasions however where option 1 (the parts replacement) has been undertaken only to require a full replacement shortly thereafter and also cases where a full replacement has taken place where a partial replacement may have proved adequately corrective.

5. Assessing the health of the engine

There are many variables that are taken into account when making a decision about an engine. Currently the majority of these are taken in a subjective fashion using a number of heuristics, which are believed to have a better than average chance of success. By having the service visit histories we are now able to identify the engines actual given point in its life including the parts and actions undertaken which may have resulted in a change in operation. Such as those seen in Figure 4 following the engine changes. Due to their complexity and the bespoke manor in which such engines are built they each have a unique character, but should ultimately within a small tolerance provide the output power within an acceptable efficiency. The extent of this variability is shown in Figure 2. Which also on the right hand panel shows the changes that occur have occurred on a single sample engine through its life.

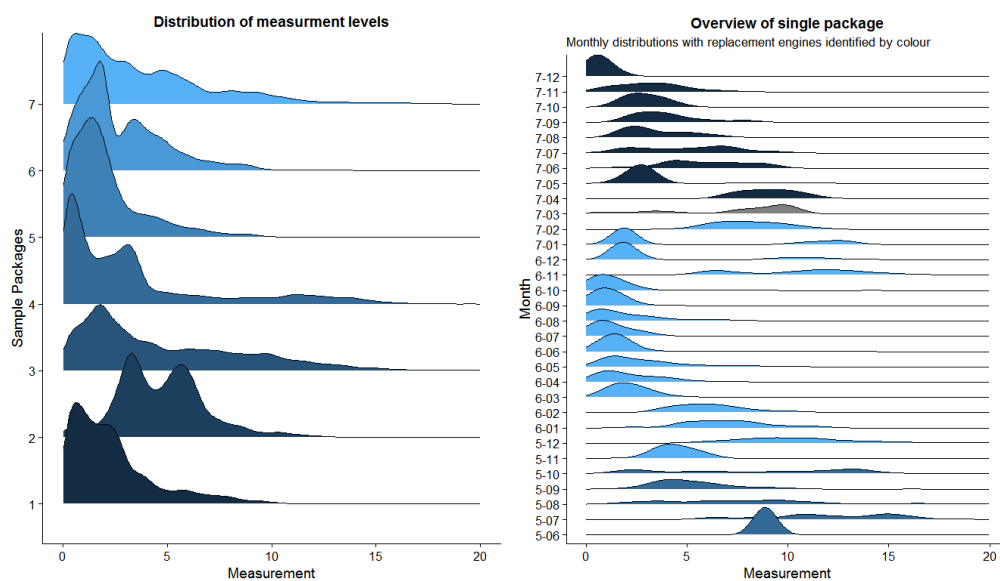


Figure 2: Relative distribution of an observable measurement across sample engines related to the sample problem.

An initial K-Means analysis was undertaken in the R environment (R Core Team, 2017) (which provides the workplace in which this research is currently being undertaken). The clustering was conducted using the common Hartigan-Wong (Hartigan, 1979) method and involved a number of parameters (such as distance between measured variables and the life of the engine at the point of measurement).

As can be seen from Figures 3, The K-means clustering appears to contradict what has been the traditional heuristic levels imposed for that measurements plotted (and as defined by the staged boxes on the plot). This indicates that there are other features which have been included within the clustering that are not generally considered which have an effect on the likelihood of the engine being replaced.

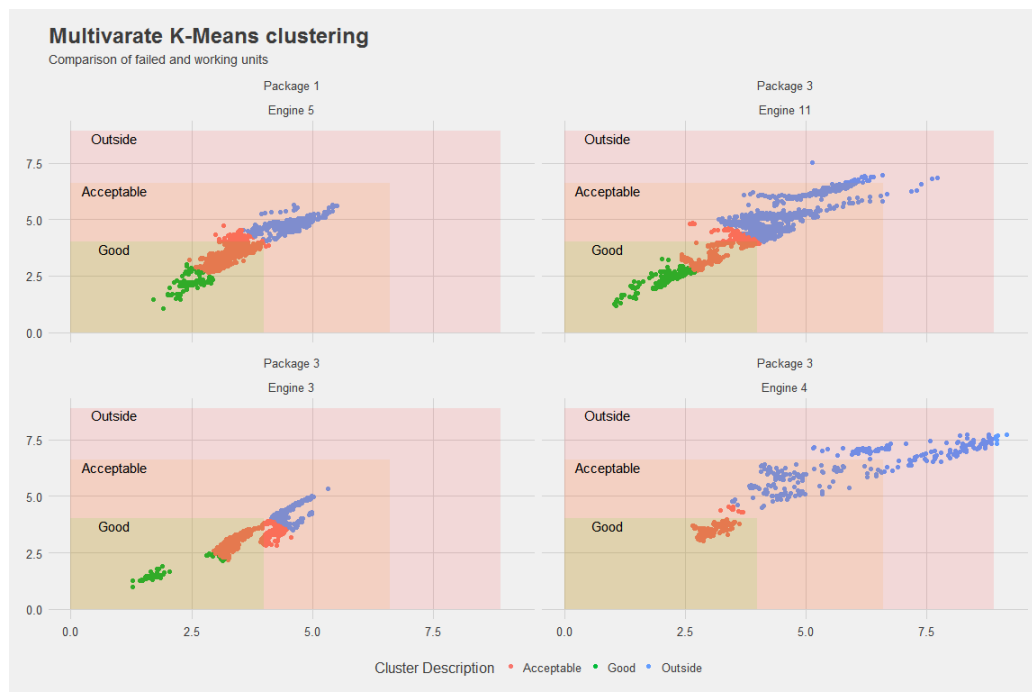


Figure 3: K-means distribution of multiple sites. The left side shows engines where no fault has occurred, right hand side shows two engines which have failed before reaching end of life. Cluster Description describes the predicted membership group using the multivariate approach as shown by the difference in colour. The boxes plotted are the heuristic limits identified for the measurement variables for traditional assessment of health.

As can be seen from the break out of the Package 3 in figure 4, all three engines which have been installed at Package 3 (a package is a full full turbine installation, including supporting equipment and a driven unit) appear to ultimately have issues (although 1 has been identified as not changed early that is simply because it was still in operation at the end of the available data, and had not a that point exceeded its intended life). The option 1 repair which has been attempted three times, appears to have little if any long-term effect. More worrisome is also that even the two engines (in this case the engine is the actual gas turbine itself) that were deemed not to be replaced early in Figure 3, also appear to be showing early signs of issues forming.

In order to identify whether it potentially simply wasteful to undertake the option 1 replacement as you may simply be adding additional costs we have begun considering applying an approach similar to the Number Needed to Treat approach described by Schechtman (Schechtman 2002). There is an assumption that an inconsistent approach has been taken in the previous handling of occurrences of the studied problems which should lead to us being able to identify cases when previous assumptions and heuristics in dealing with them have been misplaced, and through which we can identify a more optimum solution. This is likely to be due to the breadth of considerations that are required to be taken into account outside of the raw operational data from the engine (such as contractual or supply chain issues).

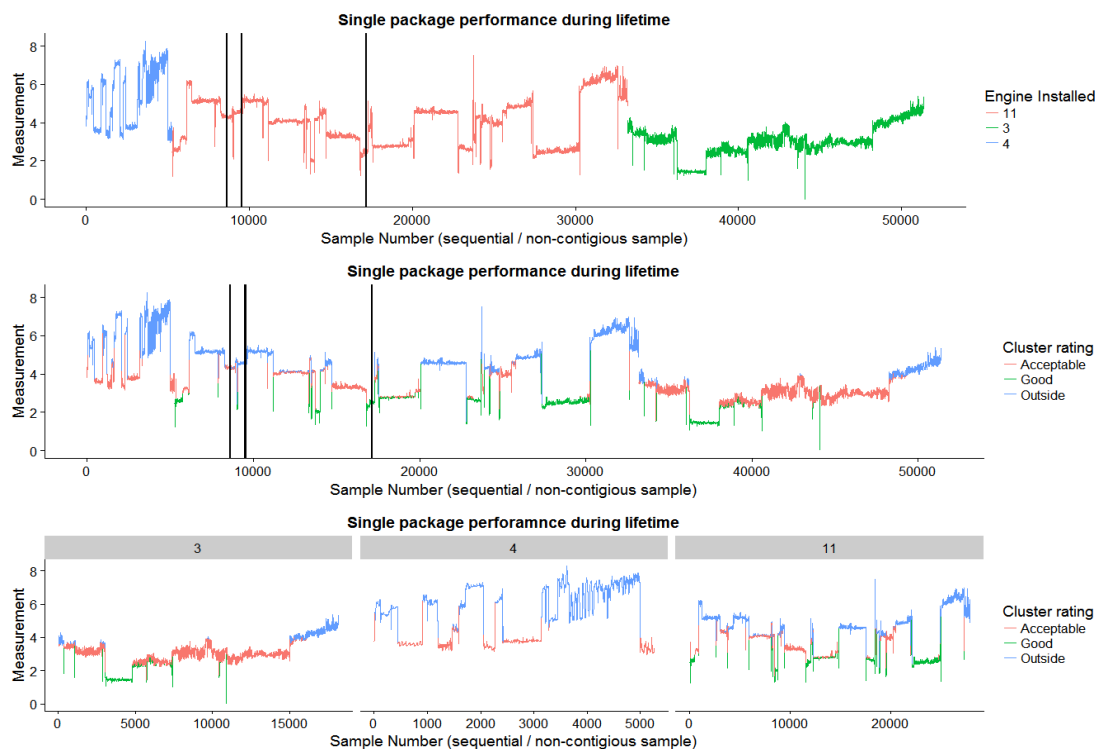


Figure 4: Time series plot of Package 3 (as shown in Figure 4), showing the different Engines installed as different colours (the change in colour indicating and Option 2 repair has been undertaken) are also lined. The point at which Option 1 repairs took place are also depicted by the black vertical lines.

6. Future work

Now that the majority of the preparatory work has been undertaken, it is now possible to consider the costs implications of the actions undertaken. Rather than simply consider the physical costs of installation and fitting, it is anticipated that more novel features will be includable within the future model. These issues include but are not limited to:

- Contractual penalties and bonuses,
- Benefit from parts of recovered (i.e. credit for parts)
- Cost of loss of useful life of parts (i.e. those that will not be useable within an overhaul)
- Cost to user of non-availability of Gas Turbine
- Costs or benefits occurred in performance of replacement unit (potentially between +3% and - 5% from designated rating), which could lead to increased profit for user.
- Costs incurred due to undertaking of no fault found investigations

Once inclusion of the above features has been completed it is anticipated an accurate working model can be produced and further problem domains can be integrated into the model.

The hope is that this will allow us to provide a labelled dataset from which we can introduce some of the more interesting deep learning approaches (such as that explored by Deutsch (Deutsch, 2017) which is directly applicable to this domain, and Sabour and Hinton's Capsnet architecture (Sabour, 2017)) which show great promise in extracting meaningful information and relationships and providing accurate predictions as to the remaining useful life in-turn provides an opportunity to better

optimize the whole lifecycle of the equipment. It also appears to support the possibility of transfer learning between seemingly dis-similar products but with the information from the surviving processes being applicable across them (i.e. all Gas Turbines obey the same fundamentals, but may be configured slightly differently).

7. Conclusion

This paper has presented an overview of a proposed system to capture and optimize the constraints of the decision making process for remedial activities on a Gas Turbine using quantitative data for it. The work conducted to date and documented within this paper, shows that overall goal of automating and optimizing the justification of certain remedial activities within the domain of Gas Turbine maintenance is achievable with the extensive cross-function data sets that are now available .

Acknowledgement

The work was supported by S.I.T Limited.

References

- Durugobo, R., Erkoyuncu, J. (2016) Critical Journal of Operations and Production Management, Vol. 36 , No.5, pp. 532-571
- Zhang, Y., Bingham, C., Ling, B., Gallimore, M., (2014) Machine fault detection by signal denoising - with application to industrial gas turbines, Measurement, 58, pp. 230-240
- Yang, Z., Ling, B., Bingham C., (2013) Fault detection and signal reconstruction for increasing operational availability of industrial gas turbines, Measurement 46, pp. 1938 - 1946
- Animah, I, Shafiee, M. Simms, N. (2016) Optimizing life extension and maintenance policy decisions for safety critical elements, *Proc. Of the 9th IMA International conference on Modelling in Industrial Maintenance and Reliability*, pp. 14 -16.
- Khan, S., Farnsworth, M., Erkoyuncu, (2016), A novel approach for No Fault Found decision-making, CIRP Journal of Manufacturing Science and Technology 17 ,pp. 18 - 31. Elsevier.
- Hartigan, J. A. and Wong, M. A. (1979). A K-means clustering algorithm. Applied Statistics 28, pp.100—108
- R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- Deutsch, J. He, D, (2017) Using Deep Learning-Based Approach to Predict Remaining Useful Life of Rotating Components, IEEE Transactions on Systems, Man, and Cybernetics, Vol 48, Issue 1 pp. 11-21.
- Sabour, S., Hinton, G., Frosst, N., (2017) Dynamic Routing between capsules [arXiv:1710.09829](https://arxiv.org/abs/1710.09829)
- Schechtman, E. (2002), Odds, Ratio, Relative Risk, Absolute Risk Reduction, and the Number Needed to Treat – Which of These Should We Use?, Value in Health, Volume 5, Number 5, pp.431 - 436

Vulnerability assessment modelling for railway networks

R. Meesit* and J. Andrews**

Resilience Engineering Research Group, University of Nottingham, University Park, Nottingham, NG7 2RD

*ratthaphong.meesit@nottingham.ac.uk ** john.andrews@nottingham.ac.uk

Abstract. Railway networks are prone to many different potential disruptive events such as technical failures (e.g. the failure of aging components), natural disasters (e.g. flooding) and intentional man-made disasters (e.g. trespass and suicide). Assessing the vulnerability of railway networks can help infrastructure managers to create the right preventive strategies to improve the robustness and the resilience of railway networks before the occurrence of disruptions. This study proposes a stochastic-vulnerability analysis model that enables the critical components of railway networks to be identified. The model is developed using a discrete event simulation technique. Its framework includes modules for assigning the disruption to the network components, predicting the network vulnerability, in terms of passenger delays and journey cancellations, and calculating the risk-based criticality of network components. Finally, an example application of the model is presented using a part of the East Midland railway network in UK.

1. Introduction

Over the past decade, railway networks have been significantly impacted by many different disruptive events such as technical failures and natural disasters. These events are difficult to predict, and some specific instances, such as flooding and earthquakes, are impossible to prevent. Thus, to avoid service disruptions the railway needs to be able to counter such occurrences. The robustness and the resilience of a railway network must be improved in order to maintain services and serve public demand during disruptions (Cabinet Office, 2010). To this end, networks would need to be enhanced to accommodate disruptions. Changes to the network are however limited through physical barriers and budget constraints. Infrastructure managers are required to make a decision on the effectiveness and priority of any improvements or added protection. This generates a need to develop a model that can support infrastructure managers in creating suitable prevention strategies to reduce the vulnerability of railway networks.

In recent years, various vulnerability analysis models have been proposed for railway networks. The majority of the initial models were based on the concept of complex network theory and used to assess the vulnerability of railway systems. For instance, Zhang et al. (2011) used complex network theory to investigate the reliability and the robustness of the Shanghai subway network. The global functionality loss (i.e. the combination of the degree of a node and the shortest path length between a node and the others) and the connectivity of railway lines (i.e. the connection of nodes on the impacted railway line) were calculated as the indicators of network vulnerability. The application of other topological indicators, such as global efficiency (i.e. the ability of a network to transfer the flow) and maximal connected subgraph (i.e. the maximum number of connected nodes in the network after disruptions), can be seen in the study of Deng et al. (2013).

Using only complex network analysis models to quantify vulnerability seems to be a useful method from the network structural point of view. However, it does not reflect the functional vulnerability of railway networks. For this reason, more recent attention has focused on estimating network flows of trains and passengers as the indicators of network vulnerability. In terms of the train flow, Ouyang et al. (2014) introduced a train flow model for vulnerability quantification. The model was constructed by considering the number of trains operating daily on each link. When a disruption happened (i.e. node removal), the accessibility of stations and the number of trains that can still operate, were calculated for the analysis. Hong et al. (2015) presented a model to evaluate the vulnerability of railway networks to flooding events. The effect of flooding events on network components was simulated using a Monte-Carlo simulation. Then, the method proposed by Ouyang et al. (2014) was adapted to estimate the performance of the railway network during flooding events.

For passenger flow, Zhang et al. (2015) adapted a gravity model to predict the movement of passengers between stations in order to evaluate the criticality of railway links. Sun et al. (2015) constructed a passenger flow model based on the passenger Origin and Destination ratio to investigating station vulnerability. Some passenger movement characteristics such as interchange times and route selections (i.e. in terms of travel time)

were included. The vulnerability indicators were passenger flow influence (i.e. flow volume affected by disruptions) and platform passenger flow (i.e. passenger volume at station during disruptions). In addition, a comprehensive rail passenger simulation model was developed by Pant et al. (2016) for assessing the vulnerability of a large-scale railway network. The model can estimate passenger trips lost during disruptions. The interdependency between a railway system and other infrastructures such as electricity substations and telecommunication equipment was analysed, and disruptions resulting from the failures of these different infrastructures and flood likelihoods in UK were modelled.

As can be seen from the literature, these vulnerability analysis models were mainly developed for identifying the critical components of railway networks and used complex network theory to construct macroscopic models. This type of model has advantages that they are reasonably simple to apply on a large-scale railway network and can also provide a quick solution for network robustness improvement. However, this method only pays attention to the analysis of a network topology. Even though recent studies have attempted to take into account some railway properties such as train and passenger flow, the results obtained are approximations since they do not consider other significant characteristics of railway networks such as the traffic control systems, the conditions of train movements and passenger behaviour when a disruption occurs.

The aim of this study is to introduce the use of a stochastic microscopic simulation model to assess the vulnerability of railway networks. This detailed modelling approach is able to predict the system performance of railway networks during disruptions and identify the critical components of railway networks in terms of risk-based criticality. These capabilities will enable infrastructure managers to establish the priority for which enhancement features are incorporated into the railway.

2. Railway network performance estimation

The railway network simulation model (Meesit and Andrews, 2017) is applied in this study to estimate the key performance indicators (KPIs) of railway networks (e.g. train delays, service cancellations, passenger delays and passenger journey cancellations) during disruptions. The model was developed using a stochastic-discrete event simulation technique. Its framework consists of two main modules: railway network modelling and passenger modelling. The description of each module is presented as follows.

2.1 Railway network modelling

The first module comprises three submodules: infrastructure module, control system module and operational module. The infrastructure module represents the railway network infrastructure assets such as track sections, stations and junctions. These infrastructure items are modelled as static resources with their significant attributes such as block lengths and speed restrictions. A state is then assigned to describe the possible status of each resource. Each of their states are considered as a set of mutually exclusive and exhaustive variables such as “*occupied or available*” and “*working or failed*”. The control system module creates train movement conditions accounting for the state of the block sections. For example, if the next two blocks are available (Green), a train can run to the next block at the permitted speed or at full operational speed. If only one block ahead is available (Yellow), a train needs to reduce its speed and prepare to stop at the next signal. However, if the block ahead is unavailable (Red), a train must stop and wait until it is free. Finally, the operational module produces the input data required for the simulation. Three main aspects (trains, service routes, and timetables) are considered. Trains are modelled as dynamic objects running on the network. The service routes of trains are modelled using the order of track section IDs, and for timetables, only departure times at terminal stations are required for the model. Then, the detailed timetables of train arrivals and departures at each station (in both normal and disruptive situation) are generated during the simulation and used as passenger information.

2.2 Passenger movement modelling

Passenger movement modelling is separated into three main parts: passenger arrivals at a station, passengers alighting/boarding a train and passenger behaviour during a disruption. Passenger arrivals at a station are simulated by a Poisson process. Then, the number of passenger arrivals is distributed to each destination station based on an origin-destination (O-D) matrix and saved into a passenger object which is created to contain the significant information about passengers such as the number of passengers, arrival time at origin, expected

arrival time at destination and the best route choice (i.e. route vector). This passenger object is properly stored in the passenger vector at an origin station waiting for a train to go to a destination. Passengers alighting and boarding happens when a train is at a station. Passengers alight from a train if the first destination in a route vector matches with the station. If a station is an interchange station, passengers alight from a train and wait for their connection. However, if a station is the final destination, passengers alight from a train and leave the network. Then, by comparison with the scheduled timetable, passenger delays are collected. Boarding occurs after passengers leaving the train have alighted. Passengers at a station compare their destination with the calling stations of the train. If their destination matches those of the trains calling stations, passengers then get on the train. Otherwise, they must wait for another train. It is noted that all trains and stations are modelled using passenger vectors. Thus, these processes are simulated by moving the passenger objects between the passenger vectors of stations and trains. The movements of passenger objects are limited by the capacity of stations and trains. Finally, passenger behaviour during a disruption is simulated in the model as presented in Figure 1. Before the occurrence of a disruption, all passengers use a normal timetable as travel information. Once a disruption happens, passengers are assumed to travel based on the disrupted timetable. A disrupted timetable refers to information provided to passengers during a disruption (e.g. real-time information announced at stations or displayed on smartphones). During this period, existing passengers on the network reconsider their routes according to the new information. If no route is found or the expected travel time is longer than the acceptable travel time (set in the model), passengers cancel their journeys. After recovering from a disruption, passengers then return to using the normal timetable.

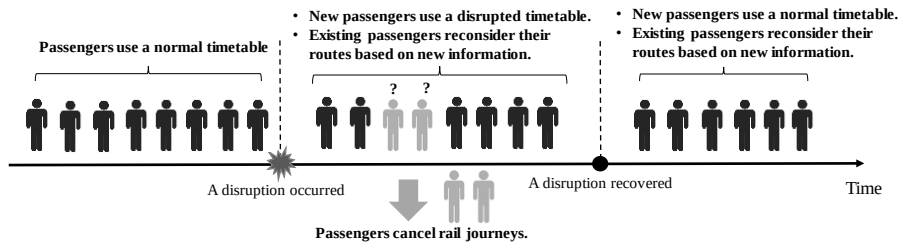


Figure 1. Passenger behaviour during a disruption simulated in the model.

3. Railway network vulnerability analysis model

The main purpose of the vulnerability analysis model is to identify the critical components of the railway network (e.g. stations or railway links). A critical component is defined as a component that, if it fails, causes large negative impacts on the network (Jönsson et al., 2008). Thus, to achieve the aim, three main steps are proposed: assigning a disruption, assessing vulnerability, and evaluating the risk-based criticality. These steps are described as follows.

3.1 Assigning a disruption to tested components

A disruptive event is simulated by changing the state of the failed network component to ‘unavailable’. This affects the movement conditions set, as described in Section 2.1, resulting in a line blockage. Thus, to conduct a vulnerability analysis, a disruption is assigned to each testing component one at a time. Either a short duration incident (less than 1 hr.) or a large impact disruption (more than 1 hr.) can be used in a test. However, only a large impact disruption is considered in this study to demonstrate the model. Then, the railway network simulation is applied to estimate the vulnerability of the railway network due to the section failure disruption.

3.2 Assessing the vulnerability of a railway network

As described in Section 2, four KPIs of the railway network, (train delays, train service cancellations, passenger delays and passenger journey cancellations) are predicted during a simulation. However, in order to evaluate the vulnerability of railway networks from the perspective of rail users, there is a focus on the passenger delays and journey cancellations. Passenger delay is defined as the difference between the actual arrival time (AT_p) and the expected arrival time (ET_p) of a passenger (p) at a destination station. The summation of all passenger delays over the network (PD) can be calculated using Equation 1.

$$PD = \sum_p^P (AT_p - ET_p) \quad (1)$$

Passenger journey cancellations (*PC*) are defined as the number of passengers who cancel their journeys when the delay time exceeds the defined threshold. This number is collected and converted to equivalent passenger delays using a weighting factor (*w*). This is then incorporated into Equation 1 to obtain the total passenger delays which is used as the vulnerability indicator, *V* (see Equation 2).

$$V = PD + PC(w) \quad (2)$$

where: *w* is the weighting factor to convert passenger journey cancellations to equivalent delay minutes. In this study it is assumed that 1 passenger journey cancellation is equal to 60-minute passenger delays (*w* = 60).

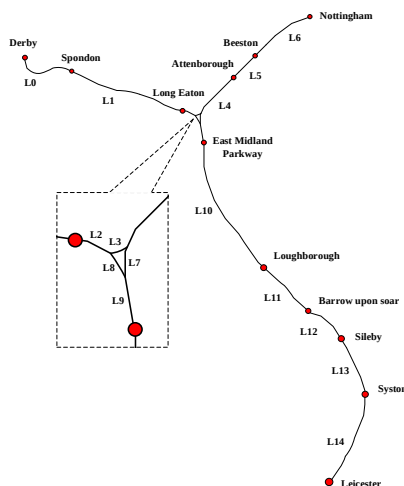
3.3 Evaluating risk-based criticality

The risk-based criticality (*RC*) of a network component is defined as the product of the vulnerability and the frequency of the network section disruption (historical data). The frequency data can be specific to events of interest such as flooding or landslides, if the focus is on a particular disruption type, or it can be the overall frequency from all disruptive events on the railway link (useful for setting annual improvement budgets). In this study, *RC* is applied to prioritise the criticality of railway network section. Sections with the higher predicted *RC* are used to prioritise infrastructure manager actions.

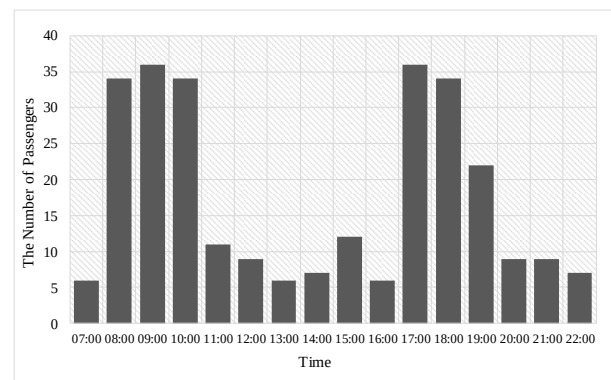
4. Model Application

4.1 Case study

A part of the East Midland railway network is used to demonstrate the application of the model. The network consists of 12 stations and 15 links (double track). The total length of the network is approximately 58 km as displayed in Figure 2(a). In addition, it is assumed there are 3 services operate daily from 6:00 to 22:00. The first service (*S1*) operates between Derby and Nottingham with the frequency of 1 train per hour. The other two services operate between Derby and Leicester (*S2*) and Nottingham and Leicester (*S3*) with a frequency of 2 trains per hour. All trains on the network are the British rail class 158 (3 coaches), and they stop at every intermediate station along their routes. For passengers, the peak travel hours are 7:00 to 10:00 and 16:00 to 19:00. The acceptable delay of passengers was set to 45 minutes. Passengers arrive at each station at different mean arrival rates. Then, they are distributed to their destinations based on the O-D matrix of passengers input to the model based on historical journey patterns indicated through ticket sales. An example of the number of passenger arrivals at a station (Attenborough) from a simulation is depicted in Figure 2(b).



(a) A part of the East Midland railway network



(b) The Attenborough station

Figure 2. A tested network (a) and the number of passenger arrivals at the station from a simulation (b).

4.2 Results

The evaluation of the link criticality is presented as an example in this study. The model software was developed in C++. The effects of large impact disruptions with mean durations of 5, 10 and 24 hrs (*SD* = 0.5, 1 and 2 hrs respectively) were considered with occurrence times of 13:00, 13:15, 13:30 and 13:45. Then, the average vulnerability of each link was calculated from the results of 2,000 simulations (at which point convergence had occurred).

4.2.1 The vulnerability of the railway links

Figures 3(a), 3(b) and 3(c) present the results obtained from running the model. It is found that the occurrence times of the large impact disruptions has little effect on the vulnerability prediction. Even though the vulnerability of each link was slightly different when the occurrence time was changed, there was no significant difference in the rank of the railway link criticality. Therefore, any large impact disruption can be applied and considered representative when performing the railway link criticality analysis.

Based on the findings, the study selected a test using a 24 hr-disruption for the analysis. The averaged vulnerability of each railway link resulting from the different occurrence times was calculated. After that, the highest vulnerability of the railway link was normalised to 1. Then, the vulnerability of other links was specified as its proportion of the highest value. This method allows the vulnerability of the railway links to be presented in the range of 0 to 1 as shown in Figure 3(d).

In Figure 3(d), Links 9, 10, 11, 12, 13 and 14 were all high criticality links on this network. Their vulnerabilities were approximately 43% higher than the next group of critical links, which are links 0, 1, 2, 4, 5 and 6. The least critical links of this network are those located at the junction (e.g. links 3, 7 and 8). This is because if one of these links is down, it affects only one of the three service routes. Thus, it is still possible for passengers to travel to all stations on the network, but their journey times will increase significantly due to the limited availability of the train services on the network.

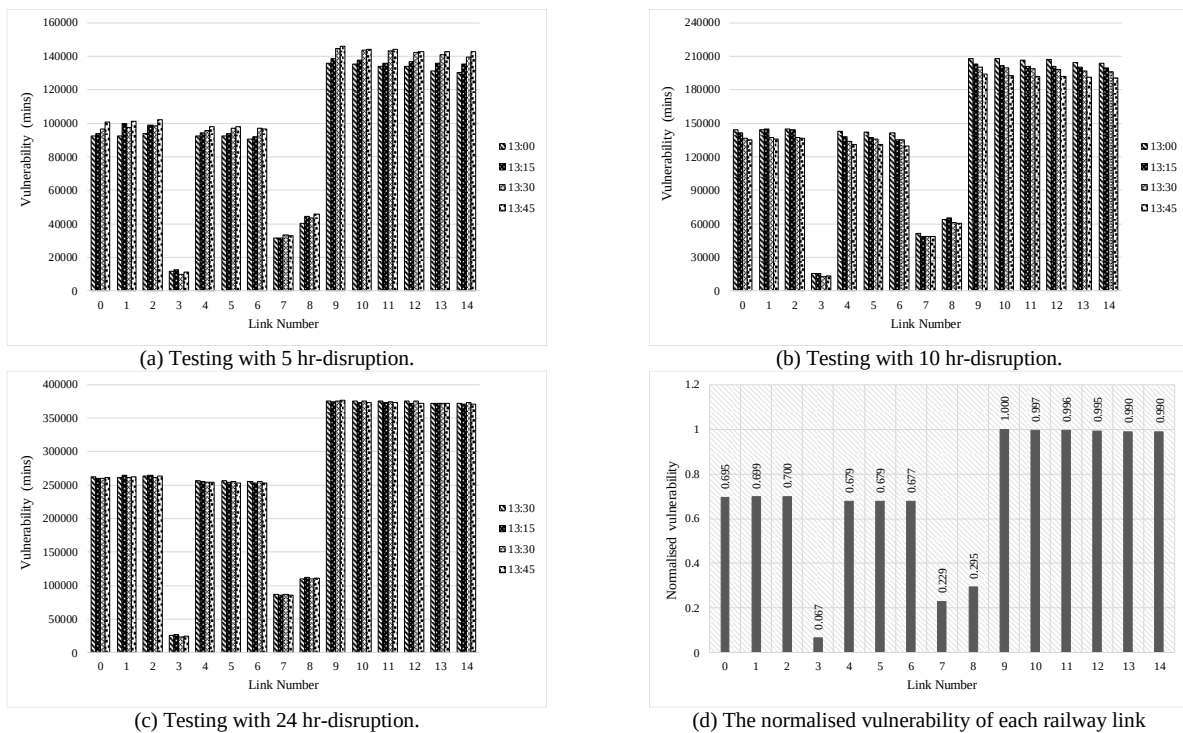


Figure 3. The vulnerability of each railway link.

4.2.2 The risk-based criticality of the railway links

The risk-based criticality of each railway link was estimated by using the normalised vulnerability and the overall frequency of the railway link disruptions from the past 10 years. To prove the concept of the model, this frequency data was reasonably assumed by considering the three main causes of disruptions on the railway network: technical failures (e.g. track failure, signaling failure and bridge collapse), natural disasters (e.g. flooding, landslide/landslip, and fallen tree) and man-made disasters (e.g. accident, trespass and suicide). Only the large impact disruptions were taken into account (See Table 1).

Table 1 illustrates the risk-based criticality of each railway link on the tested network. Link 11 is predicted as the most critical link on this network. Its risk exposure is 6.970 which represents an increase of 16.53% compared to the second critical link (i.e. link 10). The least critical link on this network was link 3 which has a risk exposure of 0.337. Therefore, based on these results, the limited budgets available for enhancing the robustness of the railway network would be most productively directed towards link 11. This would be the top priority, and link 3 would be the least priority that requires improvement for protection.

Table 1. The risk-based criticality of the railway links on the network.

Link No.	The frequency of railway link disruptions				Risk-based criticality
	Technical failures	Natural disasters	Man-made disasters	Total	
0	3	0	1	4	2.781
1	3	2	0	5	3.495
2	2	0	1	3	2.101
3	5	0	0	5	0.337
4	4	1	1	6	4.076
5	3	0	1	4	2.716
6	5	0	0	5	3.385
7	5	0	0	5	1.147
8	6	0	0	6	1.772
9	4	1	0	5	5.000
10	4	1	1	6	5.981
11	4	2	1	7	6.970
12	4	0	1	5	4.975
13	3	0	0	3	2.971
14	3	0	1	4	3.961

5. Conclusion

A stochastic railway network vulnerability assessment model is presented in this paper. The model applied two vulnerability indicators: passenger delays and passenger journey cancellations to identify the critical section links of a railway network. The application of the model was demonstrated using a section of the East Midland railway network, and the assessment of railway link criticality using different large impact disruptions was performed. In the future, the model will be further developed to investigate the criticality of other network components (e.g. stations and points). Other KPIs (e.g. train delays and train service cancellations) will also be used and compared. Then, the vulnerability analysis conducting with small impact disruptions will be studied, and the more detail about network component disruptions will be considered using a fault tree approach. This will enable the model to provide better results about network component criticality in both a qualitative and quantitative point of view.

Acknowledgements

Ratthaphong Meesit is in receipt of a PhD scholarship to study at University of Nottingham from the Royal Thai Government. Professor John Andrews is the Royal Academy of Engineering and Network Rail Chair in Infrastructure Asset Management. Both authors would like to gratefully acknowledge these organisations.

References

- Cabinet Office. (2010). Strategic framework and policy statement on improving the resilience of critical infrastructure to disruption from natural hazards. London, UK.
- Deng, Y., Li, Q., Lu, Y., Yuan, J. (2013). Topology vulnerability analysis and measure of urban metro network: The case of Nanjing. *Journal of Networks*, **8**(6), 1350–1356.
- Hong, L., Ouyang, M., Peeta, S., He, X., Yan, Y. (2015) Vulnerability assessment and mitigation for the Chinese railway system under floods. *Reliability Engineering and System Safety*, **137**, 58–68.
- Jönsson, H., Johansson, J., Johansson, H. (2008) Identifying critical components in technical infrastructure networks. *Proc. The Institution of Mechanical Engineers, Part O: Risk and Reliability*, **222**(2), 235–243.
- Meesit, R., Andrews, J. (2017) Railway network resilience modelling. *Transportation Research Part B: Methodological*, Submitted.
- Ouyang, M., Zhao, L., Hong, L., Pan, Z. (2014) Comparisons of complex network based models and real train flow model to analyze Chinese railway vulnerability. *Reliability Engineering and System Safety*, **123**, 38–46.
- Pant, R., Hall, J. W., Blainey, S. P. (2016) Vulnerability assessment framework for interdependent critical infrastructures: case-study for Great Britain's rail network. *European Journal of Transport and Infrastructure Research*, **16**(1), 174–194.
- Sun, D., Zhao, Y. & Lu, Q.-C. (2015) Vulnerability analysis of urban rail transit networks: A case study of Shanghai, China. *Sustainability*, **7**(6), 6919–6936.
- Zhang, J., Xu, X., Hong, L., Wang, S., Fei, Q. (2011) Networked analysis of the Shanghai subway network, in China. *Physica A: Statistical Mechanics and its Applications*, **390**, 4562–4570.
- Zhang, Z., Li, X., Li, H. (2015) A quantitative approach for assessing the critical nodal and linear elements of a railway infrastructure. *International Journal of Critical Infrastructure Protection*, **8**, 3–15.

Cost effective component swapping to increase system reliability

A. Najem^{1,2}, F.P.A. Coolen¹

¹ Department of Mathematical Sciences, Durham University, UK

² College of Science and Theoretical Studies, Saudi Electronic University, Saudi Arabia

Abstract

One of the strategies that might be considered to enhance reliability and resilience of a system is swapping components when a component fails, so replacing it by another component from the system which is still functioning. This paper considers cost effective component swapping to increase system reliability. The cost is discussed in two scenarios, namely fixed cost and time dependent cost for system failure.

1. Introduction

As an alternative to the use of additional components to provide increased redundancy, the use of standby components, maintenance activities or increased component reliability, Najem and Coolen (2018) introduce component swapping, this is the possibility to replace a failed component by another component in the system which has not yet failed. This is logically restricted to components which are of the same type. It seems that increasing in system reliability through such component swapping has not received much attention in the literature, yet in some scenarios it can be an attractive opportunity to prevent a system from failing. In practice, if this activity can be done at low cost, this could enable preparation of substantial repair activities, or it may be deemed to leave the system reliable enough to complete its mission.

Samaniego (1985) introduced the system signature as a tool for a assessment for systems consisting of components of a single type, by modelling the structure of a system and separating it from the random failure time of components. Coolen and Coolen-Maturi (2012) introduced the survival signature as an alternative to the signature, which can fulfil a similar role to the signature, but can also be used for reliability assessment for systems with multiple types of components. Consider a system that consists of m components of $K \geq 2$ types, with m_k components of type $k \in \{1, 2, \dots, K\}$ and $\sum_{k=1}^K m_k = m$. Assume that the random failure times of components of the same type are exchangeable, while full independence is assumed for the random failure times of components of different types. Let the state vector $\underline{x}^k = (x_1^k, x_2^k, \dots, x_{m_k}^k) \in \{0, 1\}^{m_k}$ be the state vector representing the state of the system components of type k , with $x_i^k = 1$ if the i th component of type k functions and $x_i^k = 0$ if not. The labelling of the components is arbitrary but must be fixed to define $\underline{x}^k$. Let $\underline{x} = (\underline{x}^1, \underline{x}^2, \dots, \underline{x}^K) \in \{0, 1\}^m$ be the state vector for the overall system. The structure function $\phi : \{0, 1\}^m \rightarrow \{0, 1\}$, defined for all possible $\underline{x}$, takes the value 1 for a particular state vector $\underline{x}$ if the system functions and 0 if the system does not function for the state vector $\underline{x}$. The survival signature is defined as the probability that the system functions, given that exactly l_k of type k components function, for $l_k = 0, 1, \dots, m_k$, for each $k = 1, 2, \dots, K$, and is denoted by $\Phi(l_1, l_2, \dots, l_K)$ (Coolen and Coolen-Maturi, 2012).

There are $\binom{m_k}{l_k}$ state vectors $\underline{x}^k$ with exactly l_k of its m_k components $x_i^k = 1$, so with $\sum_{i=1}^{m_k} x_i^k = l_k$. We denote the set of these state vectors for components of type k by S_l^k . Let $S_{l_1, \dots, l_K}$ denote the set of all state vectors for the whole system for which $\sum_{i=1}^{m_k} x_i^k = l_k$, for $k = 1, 2, \dots, K$. Because of the assumption that the failure times of m_k components of type k are exchangeable, all the state vectors $\underline{x}^k \in S_l^k$ are equally likely to occur. Thus, $\Phi(l_1, l_2, \dots, l_K)$ is given by following equation:

$$\Phi(l_1, l_2, \dots, l_K) = \left(\prod_{k=1}^K \binom{m_k}{l_k} \right)^{-1} \times \sum_{\underline{x} \in S_{l_1, \dots, l_K}} \phi(\underline{x}) \quad (1)$$

Let $C_t^k \in \{0, 1, \dots, m_k\}$ denote the number of type k components in the system that function at time $t > 0$. The reliability of the system $R_s(t) = P(T_S > t)$ is

$$R_s(t) = \sum_{l_1=0}^{m_1} \dots \sum_{l_K=0}^{m_K} \left[\Phi(l_1, \dots, l_K) \prod_{k=1}^K P(C_t^k = l_k) \right] \quad (2)$$

If we further assume, in addition to exchangeability of failure times of components of the same type, that these failure times are conditionally independent and identically distributed, with the probability distribution for the failure time of components of type k specified by the cumulative distribution function (CDF) $F_k(t)$, then

$$R_s(t) = \sum_{l_1=0}^{m_1} \dots \sum_{l_K=0}^{m_K} \left[\Phi(l_1, \dots, l_K) \prod_{k=1}^K \left(\binom{m_k}{l_k} [F_k(t)]^{m_k-l_k} [1-F_k(t)]^{l_k} \right) \right] \quad (3)$$

Coolen and Najem (2018) implement this concept of survival signature to quantify the reliability of systems if some components can be swapped. In this paper we aim to include considerations of costs. In section 2 of this paper, we first provide a brief introductory overview of the concept of components swapping. In section 3, we consider the cost of component swapping with fixed penalty for system failure and in section 4, we consider the cost of component swapping with time dependent penalty for system failure. In each section, we discuss some related examples. We end the paper with some concluding remarks.

2. Components swapping

Assume that there are fixed swapping rules, which prescribe upon failure of a component precisely which other component takes over its role in the system, if possible and if the other component is still functioning. Assume further that such a swap of components takes neglectable time and does not affect the functioning of the component that changes its role in the system nor its remaining time till failure. Under these assumptions, the effect of such a component swap can be reflected through the system structure function, and hence it can be taken into account for computation of the system reliability through the survival signature.

If we define a swapping case i as a regime of specified swaps that will occur if specific components fail, then let $\phi_{C_i}(\underline{x})$ denote the system structure function. Compared to the system's structure function without swapping opportunities, $\phi(\underline{x})$, ϕ_{C_i} will typically be equal to 1 for some $\underline{x}$ for which ϕ was equal to 0, reflecting the benefit from component swapping upon failure of a component. Let $\Phi_{C_i}(l_1, l_2, \dots, l_K)$ denote the survival signature given the defined swapping case i in place, so

$$\Phi_{C_i}(l_1, l_2, \dots, l_K) = \left(\prod_{a=1}^K \binom{m_a}{l_a} \right)^{-1} \times \sum_{\underline{x} \in S_{l_1, \dots, l_K}} \phi_{C_i}(\underline{x}). \quad (4)$$

Let T_S^i denote the random system failure time with swapping case i in place. Then, the reliability of the system with swapping case i in place $R_{s,i}(t) = P(T_S^i > t)$ is

$$R_{s,i}(t) = \sum_{l_1=0}^{m_1} \dots \sum_{l_K=0}^{m_K} \left[\Phi_{C_i}(l_1, \dots, l_K) \prod_{k=1}^K \left(\binom{m_k}{l_k} [F_k(t)]^{m_k-l_k} [1-F_k(t)]^{l_k} \right) \right] \quad (5)$$

It is important to notice here that the swapping case i is entirely reflected in the system survival signature. Crucially, the components have kept the same failure time distributions and the same assumptions apply, that is failure times of components of the same type remain conditionally independent and identically distributed, and failure times of components of different types remain independent. Hence, we can obtain the difference in reliability by the following equation:

$$R_{s,i}(t) - R_s(t) = \sum_{l_1=0}^{m_1} \dots \sum_{l_K=0}^{m_K} \left[\left[\Phi_{C_i}(l_1, \dots, l_K) - \Phi(l_1, \dots, l_K) \right] \prod_{k=1}^K \left(\binom{m_k}{l_k} [F_k(t)]^{m_k-l_k} [1-F_k(t)]^{l_k} \right) \right] \quad (6)$$

2.1. Example

Consider the system in Figure 1(a), which consists of four components of two types, $m_1 = m_2 = 2$. We want to examine the reliability of this system in the case that if the component A fails but the components B still functions, component A will be swapped by B, we refer to this case as case 1. We refer to the original case of the system when there is no swapping as case 0. $\Phi(l_1, l_2)$ and $\Phi_{C_1}(l_1, l_2)$, are given in Table 1, presenting only non zero values.

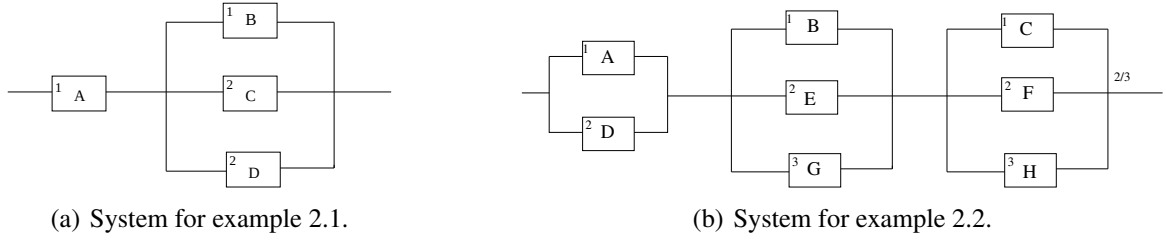
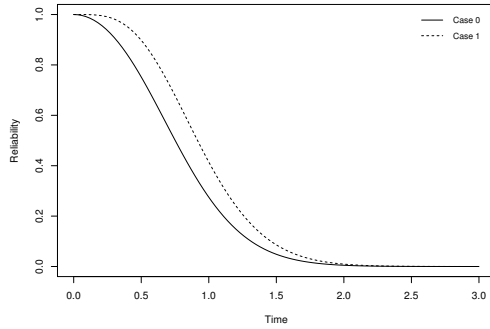


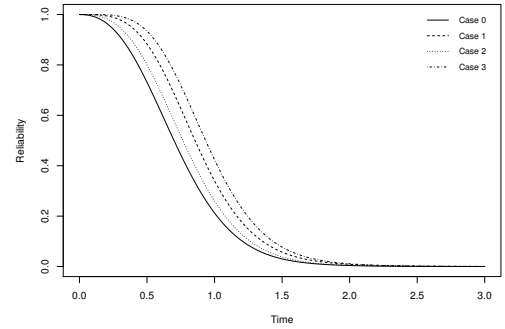
Figure1. Systems for examples.

l_1	l_2	Φ	Φ_{C_1}	l_1	l_2	Φ	Φ_{C_1}	l_1	l_2	Φ	Φ_{C_1}
1	1	1/2	1	2	0	1	1	2	2	1	1
1	2	1/2	1	2	1	1	1				

Table 1. $\Phi(l_1, l_2)$ and $\Phi_{C_1}(l_1, l_2)$ of system in Figure 1(a)



(a) For the system in Figure 1(a).



(b) For the system in Figure 1(b).

Figure2. The reliability before and after swapping cases.

Let the probability distributions of the failure times of the type 1 and 2 components have CDF $F_1(t)$ and $F_2(t)$, respectively. The reliability of the system in case 0 is $R_s(t) = [F_1(t)][1 - F_1(t)][1 - [F_2(t)]^2] + [1 - F_1(t)]^2$ and in case 1 is $R_{s,1}(t) = 2[F_1(t)][1 - F_1(t)][1 - [F_2(t)]^2] + [1 - F_1(t)]^2$. Figure 2(a) presents $R_s(t)$ and $R_{s,1}(t)$ if the failure times of type 1 components have a CDF $F_1(t) = 1 - e^{-t^2}$, and the failure times of type 2 components have a CDF $F_2(t) = 1 - e^{-t}$.

2.2. Example

Consider the system in Figure 1(b), consisting of 8 components of 3 types, $m = 8$ and $k = 3$. The notations A, B, C, D, E, F, G and H represent the specific components, and the notations 1, 2 and 3 represent the type of components. This system consist of three sub systems. The first subsystem is a parallel system consisting of components A and D, the second subsystem is a parallel system consisting of components B, E and G, and the third subsystem is a 2-out-of-3:G system consisting of components C, F and H.

The reliability of this system might be enhanced by a variety of swapping opportunities. In this example, we compare three swapping cases. In case 1, we assume that we are able to swap only type 2 components, in case 2, we assume that we are able to swap only type 3 components and in case 3, we assume that we are able to swap both type 2 and type 3 components, but a swap can only occur between components of the same type. We refer to original case without any swapping opportunity as case 0. With $m_1 = 3$, $m_2 = 3$ and $m_3 = 2$, $\Phi(l_1, l_2, l_3)$, $\Phi_{C_1}(l_1, l_2, l_3)$, $\Phi_{C_2}(l_1, l_2, l_3)$ and $\Phi_{C_3}(l_1, l_2, l_3)$ are given in Table 2. In this table we

l_1	l_2	l_3	Φ	Φ_{C_1}	Φ_{C_2}	Φ_{C_3}	l_1	l_2	l_3	Φ	Φ_{C_1}	Φ_{C_2}	Φ_{C_3}	l_1	l_2	l_3	Φ	Φ_{C_1}	Φ_{C_2}	Φ_{C_3}
0	2	2	1/3	1	1/3	1	2	1	1	2/9	2/3	4/9	1	3	1	0	1/3	1	1/3	1
0	3	1	1/2	1/2	1	1	2	1	2	5/9	1	5/9	1	3	1	1	2/3	1	1	1
0	3	2	1	1	1	1	2	2	0	1/3	2/3	1/3	2/3	3	1	2	1	1	1	1
1	1	2	2/9	2/3	2/9	2/3	2	2	1	1/2	5/6	7/9	1	3	2	0	2/3	1	2/3	1
1	2	1	2/9	2/3	4/9	1	2	2	2	7/9	1	7/9	1	3	2	1	5/6	1	1	1
1	2	2	5/9	1	5/9	1	2	3	0	2/3	2/3	2/3	2/3	3	2	2	1	1	1	1
1	3	0	1/3	1/3	1/3	1/3	2	3	1	5/6	5/6	1	1	3	3	0	1	1	1	1
1	3	1	2/3	2/3	1	1	2	3	2	1	1	1	1	3	3	1	1	1	1	1
1	3	2	1	1	1	1	3	0	1	1/2	1/2	1	1	3	3	2	1	1	1	1
2	0	2	1/3	1/3	1/3	1/3	3	0	2	1	1	1	1							

Table 2. $\Phi(l_1, l_2, l_3)$, $\Phi_{C_1}(l_1, l_2, l_3)$, $\Phi_{C_2}(l_1, l_2, l_3)$ and $\Phi_{C_3}(l_1, l_2, l_3)$ of system in Figure 1(b)

present only nonzero values. The reliability of this system can be calculated straightforwardly from this table by implementing equation (3) in case 0 and equation (5) in cases 1, 2 and 3. Figure 2(b) presents the reliability of the system in each case, if we assume that the failure times of type 1 components have CDF $F_1(t) = 1 - e^{-t^2}$, the failure times of type 2 components have CDF $F_2(t) = 1 - e^{-t}$, and the failure times of type 3 components have CDF $F_3(t) = 1 - e^{-t/2}$.

3. Component swapping with fixed penalty for system failure

Suppose that we have a system needs to continue functioning for a certain period of time $[0, \tau]$. If the system fails at any time t before the fixed time τ , a penalty cost needs to be paid. This penalty cost is fixed, independent of the failure time, denoted by c_p . If the system fails before τ , it can benefit from different swapping opportunities. An upfront cost need to be paid to enable each swapping opportunity. Let c_i denote the cost to enable swap i . We need to consider up front which opportunity of swapping cases will minimize the cost. We refer to case in which there is no swapping case as case 0. Let $C_0(\tau)$ denote the expected cost in case 0,

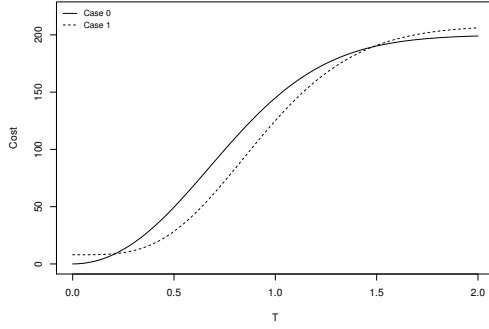
$$C_0(\tau) = c_p P(T_S < \tau) = c_p(1 - R_s(\tau)) \quad (7)$$

Let $C_i(\tau)$ denote the expected cost in case i ,

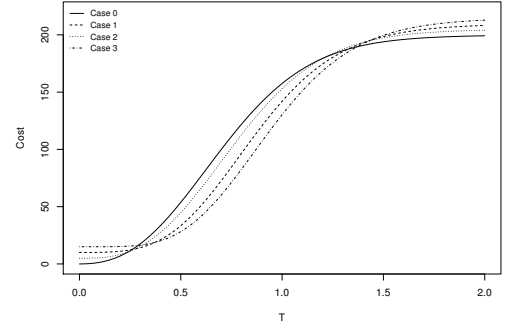
$$C_i(\tau) = c_i + c_p P(T_S^i < \tau) = c_i + c_p(1 - R_{s,i}(\tau)) \quad (8)$$

3.1. Example

Consider the system in Figure 1(a) and we keep the same scenario for the swapping case and the failure time distribution as in example 2.1. We assume that the system needs to continue functioning for a period of time $[0, \tau]$. Assume that the penalty cost if the system fails down in the period $[0, \tau]$ is fixed and is $c_p = 200$, and the case 1 swapping cost is $c_1 = 8$. So, $C_0(\tau) = 200 \left[1 - \left[F_1(\tau)[1 - F_1(\tau)][1 - F_2(\tau)]^2 + [1 - F_1(\tau)]^2 \right] \right]$, and $C_1(\tau) = 8 + 200 \left[1 - \left[2[F_1(\tau)][1 - F_1(\tau)][1 - F_2(\tau)]^2 + [1 - F_1(\tau)]^2 \right] \right]$. If $\tau = 1$, the cost in case 0 would be 145.01 and in case 1 would be 125.08, which means that if $\tau = 1$, it is good to take the opportunity of the swapping case. Figure 3(a) illustrates how the expected cost would change in case 0 and 1, depending on the value of τ . The opportunity of components swap will minimize the cost of the failure only if $0.21 < \tau < 1.48$ because if τ is small failure is unlikely and if τ is large failure is very likely even with a component swap.



(a) For the system in Figure 1(a)



(b) For the system in Figure 1(b)

Figure3. The expected cost with fixed penalty.

3.2. Example

Consider the system in Figure 1(b) and we keep the same scenario for the swapping cases and the failure time distributions as in example 2.2. Assume that if the system fails before the fixed time τ , the penalty cost is fixed and is $c_p = 200$. The cost to enable swapping cases 1, 2 and 3 are $c_1 = 10$, $c_2 = 5$ and $c_3 = 15$, respectively. If we assume that $\tau = 1$, $C_0(1) = 157.53$, $C_1(1) = 142.00$, $C_2(1) = 152.98$ and $C_3(1) = 130.30$, which means that the opportunity of swapping case 3 should be taken to minimize the expected cost. We plot $C_0(\tau)$, $C_1(\tau)$, $C_2(\tau)$ and $C_3(\tau)$ as functions of τ in Figure 3(b). It is clear from this figure that for most values of τ it is either optional not to prepare for any swaps ($\tau < 0.24$ or $\tau > 1.37$), or to prepare for both swaps ($0.29 < \tau < 1.37$), this is explained by the same reason as discussed in example 3.1.

4. Component swapping with time dependent penalty for system failure

In this section we assume that the penalty cost that needs to be paid if the system fails before a fixed time τ is dependent on the failure time. Let c represent the cost per unit of time if the system does not function in the period $[0, \tau]$. This system can benefit from different swapping opportunities if it fails at any time before τ . An upfront cost needs to be paid to enable each swapping opportunity, let c_i denote the upfront cost to be paid to enable swap i . In case 0, if the system fails at time $T_S \in [0, \tau]$, then the down time is $\tau - T_S$ and the down time cost is equal to $c(\tau - T_S)$. Let $C_0(\tau)$ denote the expected cost in case 0,

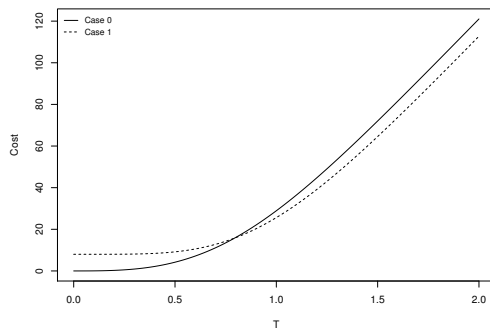
$$C_0(\tau) = cE(\tau - T_S) = c \left[\tau - \int_0^\tau R_s(t) dt \right] \quad (9)$$

Let $C_i(\tau)$ denote the expected cost in case i ,

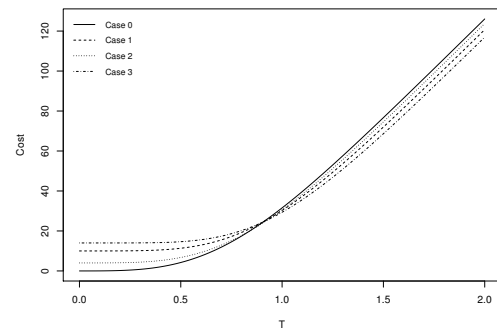
$$C_i(\tau) = c_i + cE(\tau - T_S^i) = c_i + c \left[\tau - \int_0^\tau R_{s,i}(t) dt \right] \quad (10)$$

4.1. Example

Consider the system in Figure 1(a) and, we keep the same scenario for the swapping case and the failure time distribution as in example 2.1. For the period of time $[0, \tau]$, we assume that the penalty failure cost per unit of time is $c = 100$. The swapping cost in case 1 is $c_1 = 8$. The expected cost in case 0 is $C_0(\tau) = 100 \left[\tau - \int_0^\tau [F_1(t)][1 - F_1(t)][1 - [F_2(t)]^2] + [1 - F_1(t)]^2 dt \right]$, and in case 1 is $C_1(\tau) = 8 + 100 \left[\tau - \int_0^\tau [2F_1(t)][1 - F_1(t)][1 - [F_2(t)]^2] + [1 - F_1(t)]^2 dt \right]$. If $\tau = 1$, then the expected cost in case 0 is 28.91 and in case 1 is 25.63, which means that it is better to take the opportunity of case 1 swap. Figure 4(a) illustrates how the expected cost would change in case 0 and 1, depending on the value of τ . If τ is small, the system is unlikely to fail,



(a) For the system in Figure 1(a)



(b) For the system in Figure 1(b)

Figure4. The expected cost with with time dependent penalty.

while for large τ the system is likely to fail but the swapping case delayed the failure time, leading to lower penalty costs.

4.2. Example

Consider the system in Figure 1(b) and we keep the same scenario for the swapping cases and the failure time distributions as in example 2.2. Let $c = 100$ and the cost to enable swapping cases 1,2 and 3 are $c_1 = 10$, $c_2 = 4$ and $c_3 = 14$, respectively. If $\tau = 2$, the expected cost in cases 0, 1, 2 and 3, would be 126.06, 120.72, 123.77 and 116.96, respectively. Thus, to minimize the cost we should prepare for swapping case 3. Figure 4(b) illustrates how the expected cost would change in cases 0, 1, 2 and 3, depending on the value of τ .

5. Concluding remarks

In this paper we considered cost effective component swapping to increase system reliability. We discussed the cost of component swapping with fixed penalty for system failure and with time dependent penalty. Considering a single period $[0, \tau]$ is often practically relevant and convenient for the analysis, as shown in this paper. A topic for further research is consideration over an infinite, or undefined, length of time, for example by the use of renewal theory (Tijms, 1990). Further interesting topics for future research are different cost structures and consideration of swapping, component standby, spares and maintenance activities.

References

- Najem, A. and Coolen, F.P.A (2018). System reliability and component importance when components can be swapped. *In submission*.
- Samaniego, F.J. (1985). On closure of the IFR class under formation of coherent systems. *IEEE Transactions on Reliability*, 34, 60-72.
- Coolen, F.P.A and Coolen-Maturi, T. (2012). Generalizing the signature to system with multiple types of components. *In: Complex Systems and Dependability*, W. Zamojski et al (Eds.). Springer, 115-130.
- Tijms, H. (1990). *Stochastic Modelling and Analysis*. Chichester: Wiley.

Reliability Modelling of Dispensing Processes in Community Pharmacy

M. Naybour* and Rasa. Remenye-Prescott* and M.J. Boyd**

* Resilience Engineering Research Group, Faculty of Engineering, University of Nottingham, University Park, Nottingham, NG7 2RD: evxmtn@nottingham.ac.uk; r.remenye-prescott@nottingham.ac.uk

** Division of Pharmacy Policy and Practice, School of Pharmacy, University of Nottingham, University Park, Nottingham, NG7 2RD matthew.boyd@nottingham.ac.uk

Abstract. Studies of error rates in community pharmacies have reported error rates of between 0.014% and 3.3% per item dispensed. This suggests up to 36 million items per year may contain errors in England. In addition, literature shows that patient satisfaction with services is directly related to waiting times. There is a need for a method to model pharmacy efficiency balancing safety and waiting times, ensuring that the reliability of the dispensing process is not compromised. In this paper a Coloured Petri Net (CPN) approach is proposed for analysing reliability and efficiency of community pharmacy. A pharmacy team work to complete dispensing and non-dispensing tasks, where non-dispensing tasks require staff to be temporarily removed from the dispensing process. The proposed approach is useful to investigate what affects the error rates and long waiting times, and provides modelling-based evidence to decision makers, looking to optimise staffing levels and improve the reliability of dispensing.

1. Introduction

1.1 Background

It has become increasingly apparent that healthcare systems are responsible for some level of iatrogenic harm to patients (Health Foundation, 2011), and community pharmacies represent no exception to this. So much so, that the World Health Organization (WHO) has identified them as a key area for safety improvements in primary care (WHO, 2012). For example, dispensing error rates of between 0.014% (Knudsen et al, 2007) and 3.3% (Franklin & O'Grady, 2007) have been reported in literature. In addition, patient satisfaction with their pharmacy service provision is influenced by waiting times (Afolabi & Erhun, 2003), and satisfied patients are more prone to return visits to specific providers. Given that pharmacies operate in a competitive market environment, there is a clear incentive to try and reduce their patient waiting times and to increase levels of patient satisfaction and patient retention. Doing this while also maintaining very high standards of safety presents a challenge for community pharmacy policy makers, and it is one that this paper addresses.

1.2 Reliability engineering

Risk assessment has become a common feature for industries which involve some level of risk in their operations (Andrews, 2009). Commonly seen examples of reliability engineering techniques are fault and event trees. Cohen et al. used both techniques to conduct an assessment of community pharmacy dispensing (Cohen et al, 2012). These models used Boolean logic to relate events in a way such that independent events can combine together to create system level failures. By assigning each event a probability of occurrence, a study of the fault tree structure can provide both qualitative and quantitative analyses of the undesirable outcome of the process.

Another technique commonly used for reliability engineering applications are Markov models (Boyd, 1998). These are used when systems exhibit dynamic behaviour, such as repairable components or inspection strategies. The models have the Markov property, meaning that the future behaviour of the system is independent of the past, and instead only dependent on the present.

Petri Nets are capable of modelling process which exhibit concurrency (Schneeweiss, 1999). This makes them a good candidate modelling tool for community pharmacies, which have a large number of components, as well as events, such as the final accuracy check, which are dependent on earlier stages of the dispensing process. Furthermore, colours can be introduced to Petri Net models, creating so called Coloured Petri Nets (Jensen, 1996). This type of Petri Nets enable information specific to individual tokens to be transmitted around the net, allowing the complex properties of prescriptions to be modelled at a higher resolution.

The primary healthcare sector has not been a conventional area for the application of reliability engineering techniques. Where work has been done, it has focused on the secondary healthcare settings. Petri Net modelling approaches have been proposed in a number of secondary healthcare settings, such as hospital departments, hospital information systems, and mental health providers (Dammach & Horton, 2007). The novel element of this paper is the ability to investigate the effects of various prescription checking strategies and team make-up on the reliability and efficiency of the dispensing process using a single PN model.

2. Community pharmacy dispensing process

2.1 The main stages of dispensing

Figure 1, shows the six key stages of the dispensing process (Langley & Belcher, 2009; NPSA, 2007).

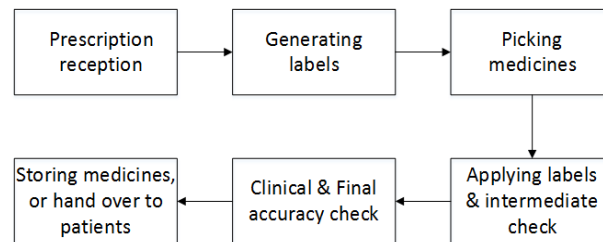


Figure 1. Dispensing process flow chart

The process begins when a patient hands over a prescription, then the labels are printed and the required medicines are picked. Then an intermediate check is performed by the member of staff who has completed the previous stages. The labels are applied to medicine boxes, before the completed prescription is passed onto a pharmacist who completes clinical and final accuracy checks. After the final accuracy check, the prescription is handed back to the patient or stored in the pharmacy to be collected later. The final accuracy check is the final failsafe of the process, beyond which errors are unlikely to be identified.

2.2 Resources

Community pharmacy teams are made up of a group of qualified and semi-qualified personnel, including pharmacists, Accuracy Checking Technicians (ACTs), dispensers and counter assistants. Teams range in size, from a single pharmacist up to large teams of people. The only legal requirement in terms of staff is that there must be a responsible pharmacist present on site. Non-staff resources used in the pharmacy include prescription forms, medicines, labels, labelling stations and a private room for consultations.

2.3 Non-dispensing tasks

In addition to the dispensing requirements of a community pharmacy, many non-dispensing tasks must also be completed (Davies et al, 2014). These tasks temporarily take staff away from the dispensing process, slowing down the movement of prescriptions. The non-dispensing tasks modelled in this paper are: stock management, patient counselling and advanced pharmacy services. Lunch hours for dispensers have also been considered in the model.

2.4 Failure modes

Reliably providing customers with accurate prescriptions in the shortest time possible is the main goal of community pharmacies. Thus the process can be considered to be a failure if either the process takes a period of time such that the customer is unhappy with, or the prescription provided is inaccurate.

Prescription inaccuracies can come in many forms; the three main components of a prescription which may contain inaccuracies are: labels, contents and label application. An example of a labelling inaccuracy would be a label indicating to take a higher dose than the amount prescribed. Contents errors can be as basic as items being dispensed which are not on the prescription form, and label application errors involve labels being applied to the wrong medicine boxes.

If an erroneous prescription is given to a patient, this is recorded as a dispensing error. If instead an error is found in a prescription at the final accuracy check, this is recorded as a near miss (Chua et al, 2003).

2.5 Defining process reliability and efficiency

The dispensing process reliability, R , is defined in equation 1:

$$R = \frac{p_{cc}}{p_{total}} \quad (1)$$

where p_{cc} is the number of completely correct prescriptions dispensed, and p_{total} is the total prescriptions dispensed.

Efficiency is commonly calculated as a ratio between the required resources to maintain a process and the useful output produced. This paper focusses on a set of efficiency indicators related to how fast prescriptions are processed by the pharmacy and how many prescriptions are completed in a day.

3. Proposed CPN-based methodology

A Coloured Petri Net (CPN) was developed to model a manual, rather than automated, dispensing process. The full model of the process is shown in Figure 2. Prescriptions enter the pharmacy from the left (transitions 2 and 4), and as staff complete the stages of dispensing prescriptions are moved from left to right.

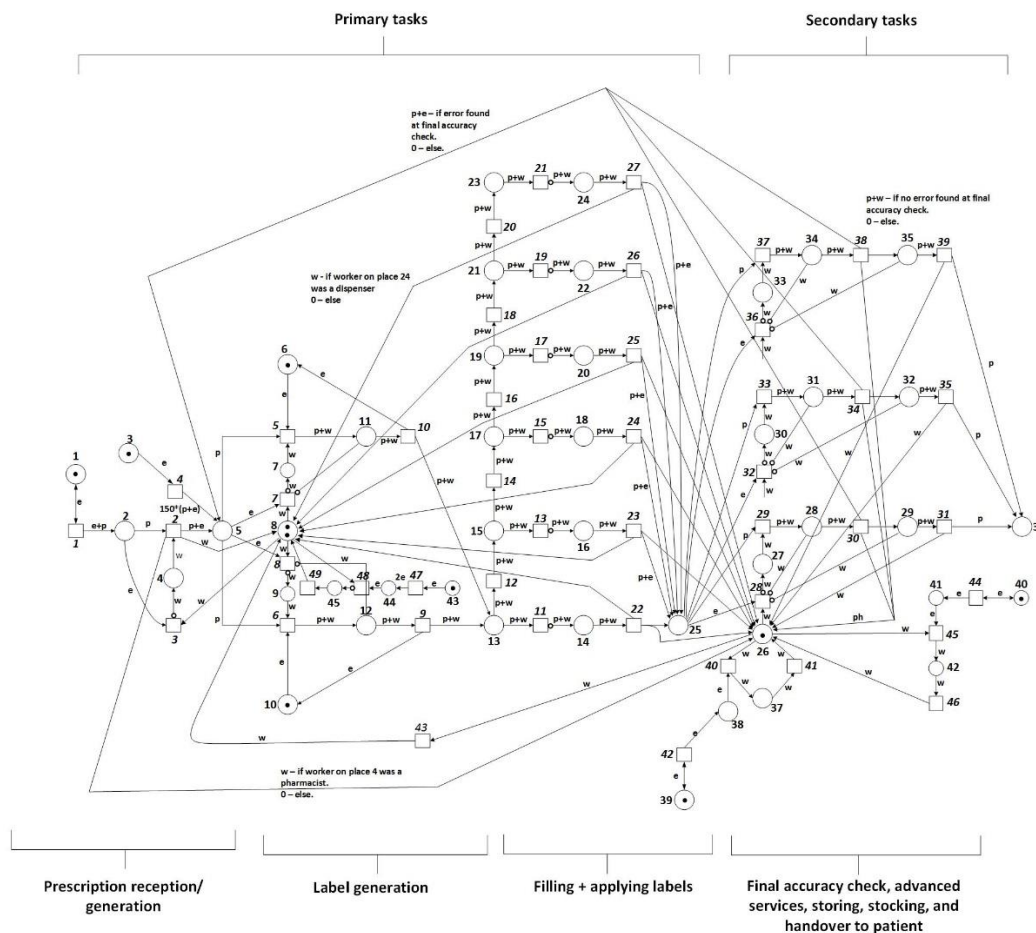


Figure 2. A CPN model for community pharmacy dispensing

3.1 Model places and transitions

Tables 1 and 2 show the descriptions of each place and transition in the CPN respectively. Each place is allocated a set of token types which are able to occupy it, where the token types used in this net are: e (basic), w (staff) or p (prescription). Transitions are given a distribution controlling their delay, and they are also denoted as processor transitions or not; note that processor transitions model stages which are dependent on the number of items in the prescription. Transition times are sampled from the distribution a number of times,

equal to the number of items in the prescription being processed, taking the sum as a delay value for the transition. Exponential, uniform and normal distributions were used, some transitions had a deterministic delay.

Table 1. Model places

Place	Description	Type
1	Walk-in task generator	e
2	Customer at counter	e
3	Delivery task generator	e
4	Staff receiving	w
5	Prescriptions to be dispensed	p, e
6, 10	Labelling stations available	e
7, 9	Staff member choosing prescription	w
8	Staff available for primary tasks	w
11, 12	Staff member generating labels	w, p
13, 15, 17	These places are used to	w, p
19, 21, 23	separate staff into parallel work streams	
14, 16, 18	Staff are assembling, and	w, p
20, 22, 24	applying labels to prescriptions	
25	Prescriptions waiting for secondary dispensing tasks	p, e
26	Pharmacists available to complete secondary tasks	w
27, 30, 33	Pharmacist allocated to complete secondary tasks for a prescription	w
28, 31, 34	Pharmacists is checking a prescription	w, p
29, 32, 35	Pharmacist is handing out/storing for delivery	w, p
36	All completed prescriptions	p
37	Advanced service being completed	w
38	Advanced service waiting	e
39	Advanced service task generator	e
40	Stocking task generator	e
41	Stocking waiting	e
42	Stocking task being completed	w
43	Dispenser lunch break generator	e
44	Lunch break ready to be taken	e
45	A dispenser is on their lunch break	w

Table 2. Model transitions

Transition	Description	(Y/N) *
1	Walk in generation: Exponential(0.0033)	N
2	Receive a prescription: Uniform(30, 60)	N
3	Move staff to counter: Deterministic(0)	N
4	Delivery generation: Det(6000)	N
5, 6	Staff choose prescription: Uni(5,10)	N
7, 8	Allocate a staff member: Det(ϵ)	N
9, 10	Label generation: Det(15)	Y
11- 21	Spreaders: Det(ϵ)	N
22- 27	Filling & label application: Normal(50,10)	Y
28, 32, 36	Pharmacist allocation: Det(ϵ)	N
29, 33, 37	Choose prescription: Uni(10, 15)	N
30, 34, 38	Final accuracy check: Uni(5,10)	Y
31, 35, 39	Hand out and counsel: Exp(0.025)	N
31, 35, 39	Store for delivery: Exp(0.05)	
40	Allocate to advanced service: Det(ϵ)	N
41	Complete advanced service: Uni(300, 600)	N
42	Advanced service generator: Exp(0.00006)	N
43	Move pharmacist primary: Det(10)	N
44	Stocking task generator: Det(6600)	N
45	Allocate to stocking: Det(ϵ)	N
46	Finish stocking: Uni(300, 900)	N
47	Begin triggering of lunch break: Det(7200)	N
48	Dispenser goes for lunch: Det(ϵ)	N
49	Dispenser finished lunch: Det(3600)	N

3.2 Model assumptions

There are two main categories of task in the model - primary and secondary tasks. Primary tasks may be completed by any member of staff, and secondary tasks are strictly limited to pharmacists. Many assumptions about staff behaviour and pharmacy operation were made in the modelling process. The assumptions related to how staff behave are as follows:

- Probability distributions for task timings and error probabilities do not depend on the level of staff.
- Dispensers are limited to primary tasks only; pharmacists can complete both types.
- Once the primary stages of dispensing are started, the same member of staff completes the remaining primary stages; stages 2, 3 and 4 in Figure 1.
- A customer arriving in the pharmacy is seen to as soon as a member of staff becomes available.
- Lunch hours are taken by dispensers (one at a time); pharmacists do not stop for lunch.

The remaining assumptions relate to the model:

- Opening hours are between 9am - 5 pm.

- Walk-in prescriptions are given priority over deliveries, and the number of items in prescriptions is sampled from a Geometric(0.35) distribution.
- Delivery prescriptions enter the pharmacy at 10am, as a group of 150 prescriptions.
- There are only 2 labelling stations used to generate labels.
- 15 minutes is the threshold beyond which a walk-in is considered to have been delayed.

3.3 Prescription modelling

Prescriptions are modelled using colours in the CPN model, where the colours represent prescription properties. The properties that alter token behaviour are: the type of prescription, the number of items, and the presence of errors. Other colours are used to keep track of how many times a prescription has been corrected during its dispensing journey, or how long it has taken to dispense.

3.4 Failures and checks

The occurrence of failures in the process is modelled by Bernoulli random variables. The following error occurrence probabilities have been taken from (Cohen et al., 2012) (labelling (0.06), filling (0.05), label application (0.03), final accuracy check (0.05)), or assumed (intermediate accuracy check (0.1)).

Both accuracy checks are dependent of the state of the prescription being checked. Correct prescriptions pass through checks unhindered, however, prescriptions containing errors are spotted and attempts are made to rectify them. Checks can fail to spot errors. Upon finding an error, there are a number of ways staff can deal with them. This paper investigates two intermediate check protocols. The first protocol (a), is that if an error is found, the prescription form is sent back to be dispensed again from the start. Using the second protocol (b), staff finding erroneous prescriptions attempt to rectify through repeating the related task. Staff “repair” each component of an erroneous prescription with the failure probabilities, described above. Note that the final accuracy check always uses protocol (a). It is assumed that intermediate checks do not increase the time taken to complete primary dispensing tasks since they are done simultaneously.

4. Pharmacy simulation scenarios and their analysis

4.1 Scenario specification

Four checking strategies were used to demonstrate the ability to evaluate pharmacy performance using the CPN model and simulation results are shown in Table 3. The first four scenarios used the same staff set up of 1 pharmacist and 2 dispensers, and the set of non-dispensing requirements was the same. Scenario 1 had no accuracy checks, and scenario 2 used a final accuracy check. Scenarios 3 and 4 were extensions of scenario 2, i.e. scenarios 3 (a) and 4 (b) introduced intermediate checks using protocols (a) and (b) respectively. The last two scenarios considered 2 pharmacist and 2 dispensers. Each scenario was simulated 6000 times.

4.2 Results and analysis

The results indicate that the different checking strategies had variable effects on the dispensing process. Incorporating a final accuracy check (2) reduced the number of dispensing errors produced significantly although it slowed down the process in comparison with using no checks (1). The use of both intermediate checking protocols further reduced errors. Protocol (a) was better than (b) in terms of reliability. Protocol (a) also performed slightly better in terms of efficiency, as can be seen by comparing the number of deliveries completed, and the mean time taken to dispense walk-ins. Thus these results indicate that it may be beneficial for pharmacies wanting to incorporate an intermediate accuracy check into their processes, to use a check that sends erroneous prescriptions to be dispensed again. This is influenced, however, by assumptions about how such checks operate. It may be the case that when repeating a “faulty” prescription, repeated error probabilities are lower (due to increased concentration).

The last two scenarios (2+2) had an additional pharmacist on the team; as a result, more prescriptions were dispensed and the reliability was retained.

Table 3. Simulation results

Scenario	Efficiency					Reliability		
	Deliveries completed	Total completed	Advanced services completed	Delayed	WI Dispense time mean (sec)	R	Near misses	Dispensing errors
1	136	236	1.9	18.5	624	0.863	0	32.5
2	111	212	1.8	25.0	713	0.993	29.7	1.57
3 (a)	113	213	1.8	24.0	696	0.996	16.6	0.86
4 (b)	112	213	1.8	24.2	699	0.995	18.3	0.97
3(a) (2+2)	150	251	1.9	7.7	482	0.996	20.7	1.12
4(b) (2+2)	150	251	1.8	6.9	469	0.996	20.8	1.13

5. Conclusion

It has been shown that the CPNs can be used to model the dispensing process in a community pharmacy, and the model can evaluate process efficiency and reliability. All main steps of dispensing are modelled, taking account of staff roles, task error probabilities and remedial actions. The model could be used to support decisions in better staffing levels and improved reliability of dispensing. Future work is focussing on in-field data collection in UK pharmacies and using the CPN model within an optimisation framework.

6. Acknowledgements

The authors would like to acknowledge and thank the ESPRC for their support and funding, without which, this research would not be possible (Grant ref EP/M50810X/1).

6. References

- Afolabi M & Erhun W (2003) Patients' response to waiting time in an out-patient pharmacy in Nigeria. *Tropical Journal of Pharmaceutical Research*
- Andrews J (2009) System reliability modelling: the current capability and potential future developments. *Journal of Mechanical Engineering Science*
- Boyd M (1998) An Introduction to Markov Modelling: Concepts and Uses. *Reliability and Maintainability Symposium*
- Cohen M et al. (2012) Risk models to improve safety of dispensing high-alert medications in community pharmacies. *Journal of American Pharmacists Association*
- Dammasch K, Horton G (2007) Active Tokens for Modelling Mental Health Care with Coloured Stochastic Petri Net. *2007 Innovations in Information Technologies (IIT)*
- Davies J, Barber N, Taylor D (2014) What do community pharmacists do?: results from a work sampling study in London. *International Journal of Pharmacy Practice*
- Franklin B & O'Grady K (2007) Dispensing errors in community pharmacy: frequency, clinical significance and potential impact of authentication at the point of dispensing. *International Journal of Pharmacy Practice*
- Health Foundation (2011) Health Foundation Research scan: Levels of Harm. *The Evidence Centre*
- Jensen K (1996) Coloured Petri nets: basic concepts, analysis methods, and practical use. *Berlin: Springer*
- Knudsen P, Herborg H, Mortensen A, Hellebek A (2007) Preventing medication errors in community pharmacy; frequency and seriousness of medication errors. *Qual Saf Health Care*
- NPSA (2007) Design for patient safety: A guide to the design of the dispensing environment. *National Patient Safety Agency*
- Langley C, Belcher D (2009) Applied pharmaceutical practice. *London: Pharmaceutical Press*
- Schneeweiss W (1999) Petri nets for reliability modeling: (in the fields of engineering safety and dependability). *Hagen: Liolo-Verlag*
- WHO (2012) Safer primary care a global challenge summary of inaugural meeting. World Health Organization

Joint maintenance scheduling and routing optimization for geographically dispersed production systems

H.S.H. Nguyen, P. Do, B. Iung and H.C. Vu

University of Lorraine, CRAN, UMR 7039, Campus Sciences, BP 70239, Vandoeuvre-les-Nancy, 54506, France

Abstract We develop a joint optimization between maintenance planning based on dynamic grouping approach and maintenance team routing for a geographically dispersed production system (GDPS). First, the interactions between the maintenance planning and routing scheduling are analyzed and formulated. A joint implementation of Local Search Genetic Algorithm (LSGA) and a new maintenance routing algorithm based on Branch and Bound (BAB) is then proposed to solve the joint optimization problem. The uses and advantages of the proposed joint optimization approach are illustrated through a numerical example of a GDPS consisting of fifteen components located in five different sites.

1. Introduction

In the literature, a large number of maintenance models has been proposed for multi-component systems, see for instance (Nicolai and Dekker.,2008; Vu et al.,2014; Keizer et al. 2017). However, majority of them investigate for centralized production systems (CPS) containing components which are located in close proximity to each other. In the framework of GDPS system, whose production is done over a number of production sites which are geographically located far apart from each other, the existing maintenance models cannot be directly adapted. In fact, planning a maintenance strategy for GDPS is a more complex work than CPS due to the dispersed location of production sites which need to schedule optimal maintenance routes. In addition, a GDPS system some specific constraints which is not exist in CPS such as disruption, possibility of accident, weather or road condition, etc (Hameed and Vatn.,2012) should be considered in maintenance models. Recently, some maintenance strategies have been developed for GDPS systems, see Diaz-Ramirez et al (2014); Camci (2015). However, these works mainly focus on the problem of maintenance routing rather than maintenance planning and optimization ones. Besides, the number of PM activities in a short-term planning period is always assumed to be known in advance, and then introduced as a constraint for the maintenance routing model.

The aim of paper is to develop a dynamic grouping maintenance strategy for a GDPS system which integrate both maintenance scheduling and routing into joint maintenance model simultaneously. The economic dependence is here considered in the proposed model. Moreover, to solve optimization problem for maintenance model, the joint implementation of a Local Search Genetic Algorithm (LSGA) and a new route optimization algorithm based on Branch and Bound (BAB) is consider. This work ensures that obtained result is global optimal solution.

The rest of this paper is organized as follows. Section 2 is devoted to describe mathematical model for a GDPS system. Section 3 focuses on performing joint optimization for GDPS integrating both maintenance scheduling and routing. A numerical example investigated to validate the performance of the proposed approach is shown in section 4. Finally, the conclusions deduced from this work are presented in the last section.

2. Problem descriptions

We consider a GDPS containing n components located at different production sites dispersed over a large area. The components are subjected to random failures which is modeled by Weibull distribution. To prevent the system from an intensive regime of failures, preventive maintenance (PM) is carried out at predetermined times. A component is considered to be in “as good as new” state after the PM. Whereas, corrective maintenance (CM) is immediately carried out after the component failures in order to restore the failed components into their operational state as soon as possible. After a minimal repair action, component is in “as bad as old” state. It should be noted that in this paper, we consider that to perform a CM action, local resources (repair team, repair tool and spare part) at maintenance sites are enough. Otherwise, to do a PM action, the external maintenance resources from the maintenance center are needed. For this reason, when PM is performed jointly

on a group of components situated in different sites, the maintenance resources have to be transported from the maintenance center, then passed all the maintained sites and finally returned to the departure center.

Assume that a group of several components, denote G^k , are preventively maintained together. These components are located in m_s different sites. The itinerary of maintenance team/spare part to the PM of all components of the group should contains all m_s sites. The itinerary has to be optimized and could be represented by decision variable X_{jv} . $X_{jv} = 1$ if site j is the v^{th} site visited by the maintenance team, otherwise, X_{jv} is equal to 0. The whole maintenance itinerary can be represented by the following matrix $X_{(n_s \times m_s)}$ which n_s is the number of all sites and m_s is the number of sites in group G^k :

$$X = \begin{pmatrix} X_{11} & \dots & X_{1m_s} \\ \vdots & \ddots & \vdots \\ X_{n_s 1} & \dots & X_{n_s m_s} \end{pmatrix} \text{ with } \begin{cases} \sum_{j=1}^{n_s} X_{jv} = 1, \forall v = [1, 2 \dots m_s] \\ \sum_{v=1}^{m_s} X_{jv} \leq 1, \forall j = [1, 2 \dots n_s] \end{cases}$$

There are a lot of possible itineraries to visit all n_s sites of group G^k . The total maintenance cost of the group depends on the itinerary that we selected. In the following paragraph, we will present how to calculate the maintenance cost of G^k as a function of a itinerary X .

Maintenance costs. The maintenance cost of group G^k contains the following parts: cost of spare part ($C_{G^k}^{sp}$), downtime cost (G^{kdt}), labor cost ($C_{G^k}^{lb}$), site preparation cost ($S_{G^k}^0$), and travel cost ($S_{G^k}^{tr}$).

$$\begin{aligned} C_{G^k}^p &= C_{G^k}^{sp} + C_{G^k}^{dt} + C_{G^k}^{lb} + S_{G^k}^0 + S_{G^k}^{tr} \\ &= \sum_{i \in G^k} C_i^{sp} + R_{j_{G^k}}^{dt} \cdot \sum_{j \in G^k, i \in j} w_i + R_{G^k}^{lb} \cdot w_{G^k} + \sum_{j=1}^{n_s} \sum_{v=1}^{m_s} X_{jv} \cdot S_j \\ &\quad + R^{tr} \cdot \left(L_{0v_1} + L_{v_{end}0} + \sum_{j=1}^{n_s} \sum_{v=2}^{m_s} \sum_{m=1, m \neq j}^{m_s} X_{jv} \cdot X_{m(v-1)} \cdot L_{m,j} \right) \end{aligned} \quad (1)$$

where, $R_{j_{G^k}}^{dt}$ and w_i are the downtime cost rate and the maintenance duration of component i located at site j ; $R_{G^k}^{lb}$ and w_{G^k} are the labor cost rate and the total maintenance duration of group G^k ; R^{tr} is travel cost rate; L_{0v_1} is the distance between the maintenance center and the first site of the itinerary; L_{0v_1} is the distance between the maintenance center and the last site of the itinerary; L_{mj} is the distance between site m and site j ; S_j is the site preparation cost of site j . When several components of the same site are maintained together, only one S_j has to be paid.

Grouping economic profit. To evaluate the effectiveness of grouping maintenance strategy, grouping economic profit is normally used. The grouping economic profit is defined as the difference between the total maintenance costs that have to be paid when the components are grouped and when they are separately maintained. The grouping economic profit of group G^k , denoted EPG , can be expressed as follows

$$\begin{aligned} EPG(G^k, t_{G^k}, L_{G^k}) &= \sum_{i \in G^k} C_i^p - C_{G^k}^p - H_{G^k} \\ &= \left(\sum_{i \in G^k} S_{ij} - S_{G^k}^0 \right) + \left(\sum_{i \in G^k} S_{ij} - S_{G^k}^{tr} \right) - (C_{G^k}^{lb} - \sum_{i \in G^k} R_i^{lb} \cdot w_i) - H(G^k, t_{G^k}) \\ &= \Delta S_{site}(G^k) + \Delta S_{tr}(G^k, L_{G^k}) - \Delta C_{lb}(G^k) - H(G^k, t_{G^k}) \end{aligned} \quad (2)$$

where,

- $\Delta S_{site}(G^k)$ is the site-preparation cost saving when several components of the same site are preventively replaced. $\Delta S_{site}(G^k) = \sum_{j \in G^k, i \in j} S_{ij} - \sum_{j=1}^{n_s} \sum_{v=1}^{m_s} X_{jv} \cdot S_j$

- $\Delta S_{tr}(G^k, L_{G^k}) = \Delta S_{tr}(\overline{G^k}) - \Delta S_{tr}(G^k)$ is the travel cost saving. $\Delta S_{tr}(\overline{G^k})$ is the travel cost when the components are maintained separately. It means that the maintenance team travels from the maintenance center, then performs the PM on only one component at a time, and returns to the maintenance center. $\Delta S_{tr}(G^k)$ is the total travel distance when the maintenance team performs the maintenance actions on all the components of

the group before returns to the maintenance center.

$$\Delta S_{tr}(G^k, L_{G^k}) = R^{tr} \cdot \left[\sum_{j \in G^k} \sum_{i \in j} (L_{0j} + L_{j0}) - \left(L_{0v_1} + L_{v_{end}0} + \sum_{j=1}^{n_s} \sum_{v=2}^{m_s} \sum_{m=1, m \neq j} X_{jv} \cdot X_{m(v-1)} \cdot L_{mj} \right) \right]$$

- Labor penalty cost is the difference between the labor cost paid for the maintenance of group G^k and that of its components. We consider that the labor cost rate of components and a group of components may be different. $\Delta C_{lb}(G^k) = R_{G^k}^{lb} \omega_{G^k} - \sum_{i \in G^k} R_i^{lb} \cdot \omega_i$

- $H(G^k, t_{G^k})$ is the penalty cost occurred due to the change of the maintenance dates of the components in the group. Note that to be grouped in the same group, the maintenance dates of individual components in the group have to be modified. The detail calculations of this penalty cost can be found in Do et al. (2014).

In a short-term planning horizon, there are many PM activities and a grouping solution of this horizon therefore may contains several groups. A grouping solution or grouping structure can be defined as a collection of mutually exclusive groups $SG = \{G^1, G^2, \dots, G^e\}$ with $G^h \cap G^k = \emptyset$ and $G^1 \cup G^2 \cup \dots \cup G^e$ covers all maintenance activities in the planning horizon. According to this definition, a grouping structure is represented as

$$Y = \begin{pmatrix} y_{11} & \dots & y_{1b} \\ \vdots & \ddots & \vdots \\ y_{a1} & \dots & y_{ab} \end{pmatrix} \text{ with } y_{ab} = \begin{cases} k, & \text{group number} \\ 0, & \text{otherwise} \end{cases}$$

where, a and b are index of component and site respectively. $y_{ab} = k$ means that the component a located in site b is in k^{th} group.

The total economic profit of a grouping structure is equal to the sum of the economic profit of its groups.

$$EPS(Y, X) = \sum_{k=1}^{ng_{max}} EPG(G^k, t_{G^k}, L_{G^k}) \quad (3)$$

where, ng_{max} is the total number of group in grouping structure Y .

3. Joint optimization of maintenance scheduling and routing

The aim of this section is to present the proposed joint optimization between maintenance grouping and routing (see Fig. 1 for more details). The optimization process can be divided into two following phases: (a) optimization at level of group of components to maximize the group economic profit (EPG); (b) optimization at grouping structure level where all groups of components of a grouping solution are considered to maximize the economic profit the considered short-term horizon (EPS).

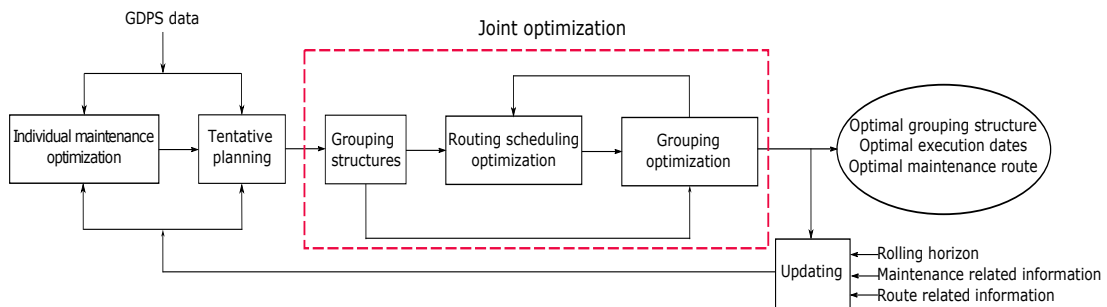


Figure 1. Joint optimization for dynamic grouping maintenance

3.1. Optimization at the group level

In the first phase, we assume that the grouping structure is known. With each group of a grouping structure, we find the optimal itinerary for the maintenance team to do the maintenance of all components of the group at the lowest penalty and travel costs. To this end, a new joint optimization approach based on the Branch and Bound is proposed. Branch and bound method is a critical enumeration of the search space. It enumerates, but constantly tries to rule out parts of the search space that cannot contain the best solution, by using lower estimated bounds. In more details, consider the iteration q ($q = 1, \dots, m_s$) being the q^{th} site of the itinerary. The

lower bound at this iteration, denoted by LB_q , is defined as the lowest maintenance cost of the considered group (corresponding to the maximum of group economic profit) that we can obtain if a site j is decided to be the q^{th} visited one. Given that the optimal itinerary has to be optimal for the maintenance grouping and routing at the same time, the lower bound is designed as $LB_q = LBF_q + LBE_q$, where LBF_q denotes the total of travel cost and penalty cost of the first part of the itinerary starting from the maintenance center and finishing at the q^{th} maintained site. This part has been already planned, LBF_q is then can be calculated exactly. Otherwise, LBE_q is the total of travel and penalty cost of the remaining part of the itinerary starting from the q^{th} maintained site and finishing at the maintenance center. Given that in the second part, the itinerary has not been planned yet, exact calculations of LBF_q are impossible. In our work, this second part is estimated by considering the shortest time needed to visit all the remaining sites of the second part.

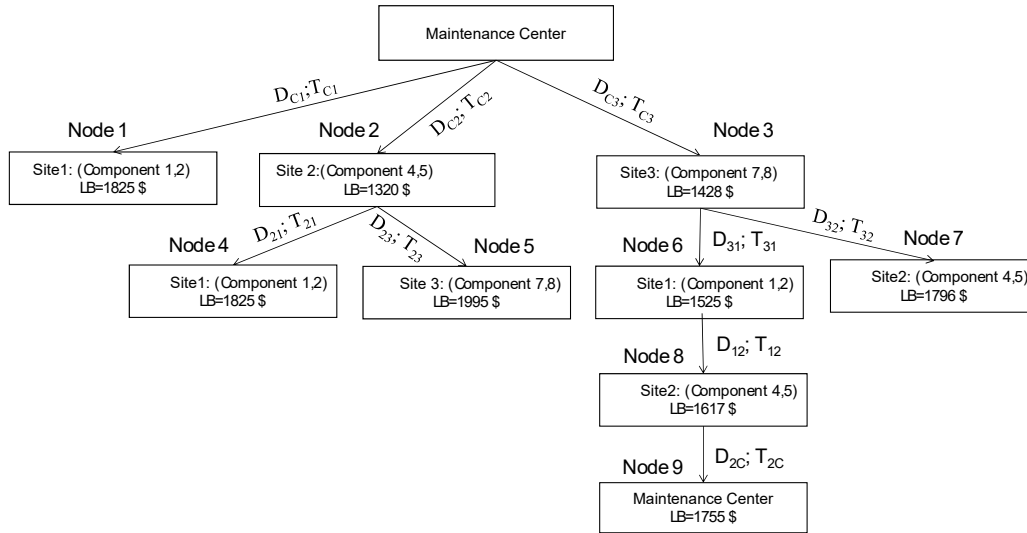


Figure 2. Search tree representation of the developed joint optimization approach

At each iteration q , a node (a maintenance site) with the lowest LB_q value is chosen for expansion. The expanded and generated nodes are called the parent and the child nodes respectively. After expansion, lower bounds of all child nodes are recalculated. The space of solutions is updated to cover all unexpanded nodes of the previous iterations and the new child nodes. The process is repeated until all the maintained sites are planned. To facilitate the understanding, let us consider an example in Figure 2. The BAB process is applied to a group containing 8 components. The components $\{1,2\}$, $\{4,5\}$, $\{7,8\}$ are located in the three different sites 1, 2, 3 respectively. The BAB process starts at the maintenance center (node 0, level 0). Since at this level there is only one maintenance center note, the expansion is directly applied to this node. Nodes 1, 2 and 3 are three child ones. The process jumps to level 1 ($q = 1$). The lower bound is evaluated for all nodes of this level. Node 2 is then selected as the expansion node of level 1, since its lower bound value is the smallest one among all the nodes of the level. The expansion is then done on node 2. Nodes 4 and 5 are generated as children nodes. The process goes to level 2. The lower bound is evaluated for nodes 4 and 5. Node 3 is then selected for expansion. The process then returns to the level 1, and so on. The process is stopped at node 9, and the optimal maintenance itinerary is to start at the maintenance center, then travel across sites 3, 1, 2 respectively, and come back to the maintenance center.

3.2. Optimization at the grouping solution level

In this second phase, the joint optimization between maintenance grouping and routing is considered at grouping solution level. To solve the joint optimization problem, a joint implementation of a Local Search Genetic Algorithm (LSGA) and the modified BAB approach (developed in the previous subsection) is proposed. For more details, LSGA generates first a random population of grouping structures. To evaluate the performance (total economic profit) of a grouping structure, the modified BAB is applied to find the optimal itineraries of all groups in a grouping structure. Promising grouping solutions will be selected based on their

total economic profit. At each iteration, local search, elitism, crossover and mutation operators are applied to the promising solutions. The best solution is found after certain iterations (Ombuki and Ventresca, 2004).

4. Numerical examples

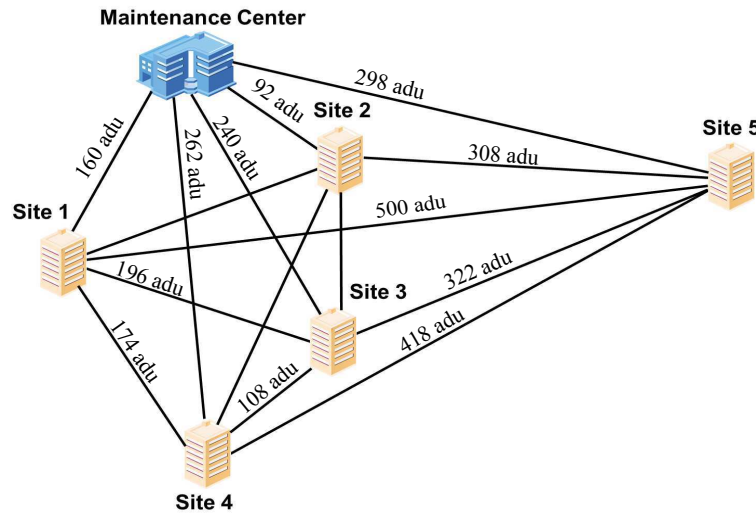


Figure 3. GDPS containing 3 production sites

The aim of this section is to demonstrate the performance of the proposed joint optimization between maintenance grouping and maintenance routing, named JOMSR. To do this, the performance of JOMSR will be compared to the separate optimization approach named SOMSR. Most steps of SOMSR are the same as that of JOMSR, except that the maintenance grouping and maintenance routing are separately considered. The penalty costs of the grouping maintenance are not considered during the finding of the optimal itineraries. The optimal itineraries are determined based on the total travel distance only. Both JOMSR and SOMSR are applied to find the best grouping solution within the planning horizon $PH = [0, t_{51} + \omega_5] = [0, 8460.7]$, and optimal repair team itineraries of a GDPS containing 15 components located in 5 different sites (Figure 3). The given data and the tentative maintenance dates of 15 components are reported in Table 1, and the labor cost rates R^{lb} are fixed at 100, 200, 300 for the required skill levels of repair team 1, 2, 3 respectively. It should be noticed that, in this study, all parameters are given in arbitrary units, e.g., arbitrary time unit (atu), arbitrary distance unit (adu) and arbitrary cost unit (acu).

The results provided by JOMSR are shown in Tables 2 and 3. The joint optimization helps to reduce up to 59.85% travel distance of the maintenance team as well as the travel cost. The reduction of travel distance is very important since it is not only meaningful from the economic point of view, but also the sustainable one (reduce energy consumption, travel-related risks, environmental negative impacts). Given the above performance, the grouping maintenance also leads to slight increases of the labor and the CM costs which is around 7.44%.

Similarly, the results provided by the use of SOMSR are reported in Tables 4 and 5. By comparing the obtained results of SOMSR and JOMSR, we can conclude that both JOMSR and SOMSR provide the same grouping structure, however, the optimal itinerary and the best departure time of the maintenance team to maintain the group G^1 are not the same. Given a higher performance in reducing the travel distance as well as the travel cost of SOMSR when compared to the JOMSR, the routing solution provided by SOMSR is not optimal. The reason is that SOMSR consider only the travel distance during the finding of the optimal route. The obtained itinerary is then only optimal for maintenance routing but not for the grouping maintenance. The travel cost of the itinerary is minimal, but the penalty costs related to the maintenance date are too large. Overall, it leads to a lower performance of SOMSR when compared to that of JOMSR. In our case study, JOMSR helps to save up to 4.79% of the total maintenance cost when compare to SOMSR. The joint optimization is then important to ensure the global optimality of the maintenance grouping and routing.

Table 1. Data and individual optimization at component level

Unit	λ_i	β_i	C_i^{sp}	C_i^c	w_i	t_i^e	sl_i	S_i^{tr}	x_i^*	ϕ_i^*	t_i^1
1	2499	2.85	1757	561	30	1398	1	4800	6861.5	4.1452	5526.5
2	3257	2.77	2546	669	48	2729	2	4800	10592	4.5875	7955.5
3	2423	3.72	3651	628	63	781	3	4800	5852.3	10.613	5071.3
4	2815	2.87	1874	665	27	2418	1	2760	6846.6	3.5731	4428.6
5	3505	2.74	2676	722	45	2649	2	2760	10972	4.1107	8415.7
6	2327	3.69	3397	619	66	1142	3	2760	5728.5	11.076	4613.5
7	2645	2.82	1704	517	33	3281	1	7200	8303.6	4.4215	5022.6
8	3057	2.76	2341	547	51	3799	2	7200	11512	5.0947	7818.1
9	2624	3.54	3313	611	72	1035	3	7200	7246.7	10.881	6244.7
10	2446	2.88	1656	445	30	2498	1	7860	7778.5	4.6119	5346.5
11	2963	2.74	2449	628	45	2999	2	7860	10564	5.3044	7661.1
12	2327	3.73	3332	690	66	1155	3	7860	5730.6	12.951	4575.6
13	2698	2.84	1659	689	36	1546	1	8940	7331.3	4.5637	5908.3
14	3359	3.16	2432	609	51	3598	2	8940	9227.8	5.0827	5629.8
15	2483	3.65	3532	314	72	1819	3	8940	7564.3	8.8367	5796.3

Table 2. Grouping solution provided by JOMSR

Optimal grouping structure	ΔS_{tr}	ΔS_{Site}	ΔC_{lb}	H	EPG	EPS
$G_1 = \{3^1, 6^1, 9^1, 12^1, 13^1, 14^1, 15^1\}$	31530	380	12300	3938.5	15671.5	40462.8
$G_2 = \{1^2, 4^2, 7^2, 10^2\}$	12570	0	0	257.9	12312.1	
$G_3 = \{2^3, 5^3, 8^3, 11^3\}$	12570	0	0	90.8	12479.2	

Table 3. Maintenance itineraries and departure times provided by JOMSR

Groups	Departure times	Maintenance routes	Travel distances
$G_1 = \{3^1, 6^1, 9^1, 12^1, 13^1, 14^1, 15^1\}$	4992.1	$0 \rightarrow 2 \rightarrow 1 \rightarrow 4 \rightarrow 3 \rightarrow 5 \rightarrow 0$	1194
$G_2 = \{1^2, 4^2, 7^2, 10^2\}$	5073.2	$0 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 1 \rightarrow 0$	670
$G_3 = \{2^3, 5^3, 8^3, 11^3\}$	7852.4	$0 \rightarrow 1 \rightarrow 4 \rightarrow 3 \rightarrow 2 \rightarrow 0$	670

5. Conclusions

In this work, a joint optimization between the maintenance grouping and routing is modeled and formulated for geographically dispersed production systems with taking into account the interactions between components at both component and site levels. To solve the joint optimization problem, the SOMSR optimization process is proposed based on the LSGA and the BAB algorithms. The numerical examples shown that the grouping maintenance is an interesting approach for the maintenance management of GDPS systems, and the joint optimization between the maintenance scheduling and routing is better than the separate optimization. In the

Table 4. Grouping solution provided by SOMSR

Optimal grouping structure	ΔS_{tr}	ΔS_{Site}	ΔC_{lb}	H	EPG	EPS
$G_1 = \{3^1, 6^1, 9^1, 12^1, 13^1, 14^1, 15^1\}$	31980	380	12300	5468.8	14591.2	
$G_2 = \{1^2, 4^2, 7^2, 10^2\}$	12570	0	0	257.9	12312.1	39382.5
$G_3 = \{2^3, 5^3, 8^3, 11^3\}$	12570	0	0	90.8	12479.2	

Table 5. Maintenance itineraries and departure times provided by SOMSR

Groups	Departure times	Maintenance routes	Travel distances
$G_1 = \{3^1, 6^1, 9^1, 12^1, 13^1, 14^1, 15^1\}$	4914.8	$0 \rightarrow 2 \rightarrow 5 \rightarrow 3 \rightarrow 4 \rightarrow 1 \rightarrow 0$	1164
$G_2 = \{1^2, 4^2, 7^2, 10^2\}$	5073.2	$0 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 1 \rightarrow 0$	670
$G_3 = \{2^3, 5^3, 8^3, 11^3\}$	7852.4	$0 \rightarrow 1 \rightarrow 4 \rightarrow 3 \rightarrow 2 \rightarrow 0$	670

future work, other kinds of component interactions (such as stochastic and/or functional dependencies) and production constraints will be investigated and integrated in the proposed joint optimization model.

References

- Grall A, Dieulle L, Berenguer C, Roussignol M (2002). Continuous-time predictive-maintenance scheduling for a deteriorating system. *IEEE transactions on reliability*, 51(2), 141-150.
- Dieulle L, Berenguer C, Grall A, Roussignol M (2003). Sequential condition-based maintenance scheduling for a deteriorating system. *European Journal of operational research*, 150(2), 451-461.
- Nicolai RP, Dekker R (2008). Optimal maintenance of multi-component systems: a review. *In complex system maintenance handbook*, 263-286. Springer, London.
- Vu HC, DO P, Barros A, Berenguer C (2014). Maintenance grouping strategy for multi-component systems with dynamic contexts. *Reliability Engineering & System Safety*, 132, 233-249.
- Dekker R, Wildeman RE, Van der Duyn Schouten FA (1997) A review of multi-component maintenance models with economic dependence. *Mathematical Methods of Operations Research*, 45(3), 411-435.
- Zhou Y, Zhang Z, Lin TR, Ma L (2013) Maintenance optimisation of a multi-state series-parallel system considering economic dependence and state-dependent inspection intervals. *Reliability Engineering & System Safety*, 111, 248-259.
- Keizer MCO, Flapper SDP, Teunter RH (2017) Condition-based maintenance policies for systems with multiple dependent components: A review. *European Journal of Operational Research*, 261(2), 405-420.
- Hameed Z, Vatn J (2012) Role of grouping in the development of an overall maintenance optimization framework for offshore wind turbines. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, 226(6), 584-601.
- Diaz-Ramirez J, Huertas JI, Trigos F. (2014) Aircraft maintenance, routing, and crew scheduling planning for airlines with a single fleet and a single maintenance and crew base. *Computers & Industrial Engineering*, 75, 68-78.
- Camci F. (2015) Maintenance scheduling of geographically distributed assets with prognostics information. *European Journal of Operational Research*, 245(2), 506-516.
- Do P, Vu HC, Barros A, Berenguer C (2015) Maintenance grouping for multi-component systems with availability constraints and limited maintenance teams. *Reliability Engineering & System Safety*, 142, 56-67.
- Ombuki BM, Ventresca M (2004) Local search genetic algorithms for the job shop scheduling problem. *Applied Intelligence*, 21(1), 99-109.

Maintenance engineering management of a LNG plant critical system: a case study of gas turbine dry gas seal

E. Pereira, B.M Alkali & O. Niculita

Department of Engineering, School of Engineering and Built Environment, Glasgow Caledonia University, Glasgow, G4 0BA, United Kingdom. Eloi.Pereira@gcu.ac.uk, Babakalli.alkali@gcu.ac.uk, Octavian.Niculita@gcu.ac.uk

Abstract: This paper focuses on maintenance management of critical equipment of a New Liquefied Natural Gas (LNG) plant. The challenges associated with the risk of failure affect the operational conditions of the plant equipment, and also the recommendations by the original equipment manufacturer (OEM) to maintain critical system is a difficult task. In this paper a case study of a new gas turbine design within an LNG plant is considered. The maintenance cost, operation and availability are some of the most important concerns to the gas turbine operators. Majority of the gas turbine maintenance are related to the replacement and repair of hot section static parts. A review of the operating and maintenance practices for heavy-duty gas turbines and others critical equipment, with emphasis placed on the types of inspections and factors that influence maintenance schedules is presented. The Reliability Centred Maintenance (RCM) modeling and criticality analysis approach is used to analyse historical failure and maintenance data. To compliment the accuracy of data information careful consultation with the operators is conducted and we test the goodness of fit of using a Weibull model for the analysis. A maintenance optimisation model is discussed and an attempt is made to model the failure pattern to determine and recommend cost effective maintenance management schedule for the gas turbine dry gas seal.

1. Introduction

There are a number of approaches an organisation can take to maintain rotating machinery, and often an organization will practice a number of different maintenance philosophies at once. To detect a problem well before failure occurs, and to make changes to improve equipment reliability, it is important to understand why machines fail in the first place. The source of machine failure can start on the design and end with poor maintenance practices and operating conditions. The way the machine is manufactured, the way it is installed, and the way it is overhauled all contribute to the ultimate life of the machine. All rotating machinery, like gas turbines and gas compressors, will vibrate. The level of the vibration and the pattern of the vibration tell us something about the internal condition of the rotating component. (Vibration training inter oil PNG 2010). Maintenance and reliability engineers deploy two basic strategies to insure asset reliability: predictive maintenance and proactive maintenance. No matter how often we continuously monitoring an equipment, we cannot influence the probability of failure we can only hope to find the problem early enough that corrective action can be planned for a time that is least disruptive and costly to the organisation. In this paper, the dry gas seal is investigated due to several sources of failure which can affect seal integrity (such as wrong installation, prolong reverse rotation, material compatibility, excessive vibration or alignment, loss of film stiffness across the sealing faces, ingress of solids or liquids into the seal cavities will cause a reduction in performance). The Reliability Centered Maintenance (RCM) approach discussed by Moubray (1997) is an adequate approach to determine the maintenance requirement of complex physical asset in its operating context. In this paper RCM is conducted on the dry gas seal to give a better insight of the failure modes and link of the failure modes to the historical failure of dry gas

seal types. The dry gas seal is considered as a single point failure resulting to production losses. A brief technical description and the operational process of the dry gas seal component as discussed in Section 2.

2. Dry gas seal

Dry gas seals are designed to operate with clean dry gas. The sealing gas is normally supplied during operation from the discharge flow of a selected compressor. The dry gas seal has the following advantages;

- System simplicity: Simplification of the system and therefore greater reliability and reduced maintenance (elimination of the overhead tank for the seals, of the oil pumps, of the traps, of the gas separators)
- No process gas contamination: Elimination of oil transfers, through seals, in the compressed gas.
- Increased efficiency: Economy of the system due to the reduction of oil consumption through seal, to the energy consumption (no oil pumps), to the improved reliability and availability.
- Minimization of process gas discharge: To the atmosphere during normal running (Dry gas seals secondary vent versus oil seals traps vent).
- Reduction of fire risk due to burst of pipes or pipe fitting at high pressure.

The dry gas seal consist of stationary and rotating parts as illustrated in the cross sectional view diagram in Figure 1. During start-up conditions, the seal gas is supplied from the fuel gas system or another compressor source.

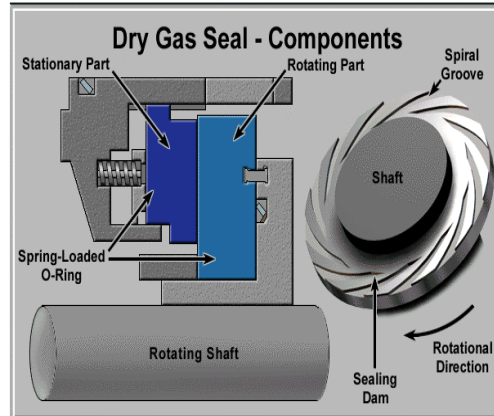


Figure 1: Cross sectional view of dry gas seal components

The seal gas goes through a duplex set of filters, a differential pressure control valve with an orifice by-pass (for minimum flow). There are different type of dry gas seal which include; single seal, double seal and the tandem seal.

Single seal

Single seal is used where gas is sealed at relatively moderate pressure, and where a small amount of leakage into the atmosphere does not present a problem.

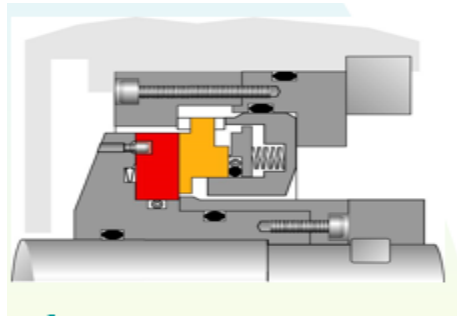


Figure 2: Single seal configuration

Double seal

The double seal has sealing modules facing each other and sometimes fitted with common rotating element. Double seal arrangement provides a positive barrier to process gas through the use of buffering with inert gas. Buffer gas is normally set at higher pressure than sealed process gas, leakage between inert gas and process gas can occur. The double seal prevents any trace of process gas from escaping to atmosphere. The diagram in Figure 3 give and example of double seal configuration. (ADGAS 2006)

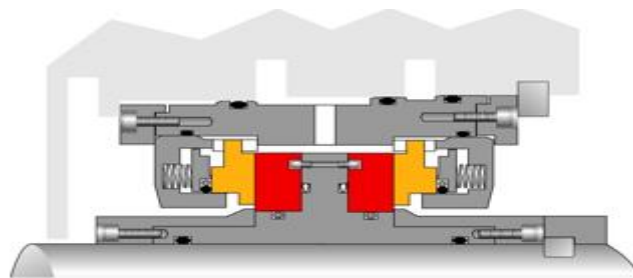


Figure 3: Double seal

Usage of a double seal arrangement is mainly to seal toxic or hazardous gases or where process is being protected from outside contaminants.

Tandem seal

The Tandem seal is type of dry gas seal and the tandem seal arrangement consist of a primary seal and a secondary seal, contained within a single cartridge. An example of the tandem seal set is presented in Figure 4. The sealing modules can share sealing load, or more commonly, one seal would handle full pressure drop while other seal would run as a standby or back up seal with zero pressure differential, this back up seal then functions as an additional barrier between the process gas and atmosphere. Higher pressure may require triple tandems where two seals share the sealing load and the third is a backup and barrier seal, (Stahley, 2001).

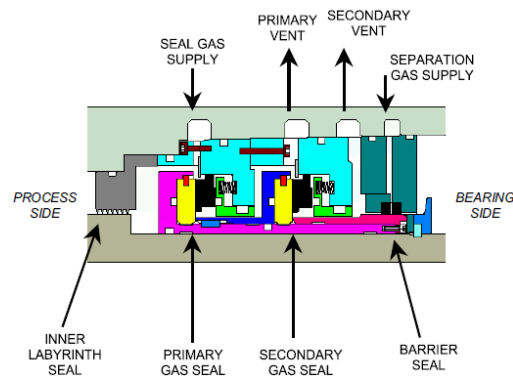


Figure 4: Tandem seal

Dry gas seal maintenance and failure

The dry gas seal has a variety of challenges related to maintenance, storage and failure. The seal require adequate maintenance, and proper care should be utilized to keep the gas seal in their optimum performance condition during operation. The dry gas seal is always exposed to oil or other contaminations. Although the seals are normally reliable in operation, however they can be damaged as a result of incorrect assembly or handling, and the spear seals must be assembled with the appropriate installation plates in order to hold the rotating and stationary elements in alignment.

If the seals are to be fitted to the compressor, it is necessary to ensure that the seal faces are not contaminated with oil. This is necessary to prevent the seal faces from wringing together. During any storage period it is necessary to keep the seals free from all solid and liquid contaminants either which can damage the seal faces during compressor start up and operation.

3 Reliability Centered Maintenance

The Reliability Centred Maintenance (RCM) concept is one of most useful tools that is established by the aeronautics business group in the early of 1970s (Brauer D. and Brauer G. 1987). This maintenance tool had been developed for failure analysis of equipment to reduce maintenance tasks and unnecessary works to propose adequate maintenance program (Demers 2005). By doing this, the availability and reliability are inclined to increase and hence decrease maintenance costs (Ramli and Arffin 2012). These are the seven basic steps of the RCM process which can lead to the achievement of best maintenance practices, based on increasing the reliability, safety and cost saving (Moubray 1997). The RCM modeling approach is considered in this paper.

3.1 RCM investigation and analysis

RCM is conducted on the on the dry gas seal and the result indicates that debris from the seal gas filter was freed into the clean side of the filter, this debris preferentially flowed through the drive end seal of the compressor. The increased seal gas flow across the seal is attributed to one:

- Wear of seal face due to debris
- Seal hang-up due to debris

The possible causes of the seal failure are believed to be:

- Debris passing the seal gas filters due to saturation and an impulse of gas being introduced.
- Saturated seal gas filters due to insufficient cleaning of the seal gas lines after construction work.

Contributing factors of the seal failure are believed to be:

- Debris introduced from the seal drain.
- Seal gas Differential Pressure Control Valve being 100% open when high-pressure seal gas is introduced.

The seal gas filters the high differential pressures lead to debris had possibly passed through the seal gas filter and flow to the seals. This validates that debris was in the seal gas flow to the dry gas seals. From our investigation we found that 99% of the DGS failures during commissioning is due to none seal related problems, such as foreign contaminates in the seals cavity, seals installation errors, alignment problems, piping problems, liquid inside piping and compressor casing, instrument problems. Contact between the faces is normally a primary cause of gas seal failure, face contact due to mechanical operation are, wrong installation., prolong reverse rotation., material compatibility, excessive vibration-alignment, reverse pressurization, loss of film stiffness across the sealing faces, damaged elastomers during installation.

It was also deduced that 80% of the seal failures are due to seal gas contaminations, continuous supply of clean and dry seal gas is one of the most important requirements for trouble free operation of dry gas seal. A study from DGS (2012) concluded that dry gas seal reliability is equal to compressor availability. Dry gas seal contamination is predominantly caused by inadequate quality and or pressure of the gas in the seal cavity. Ingress of solids and or liquids into the seal cavities will cause a reduction in performance. Reduction in performance will eventually lead to instability of the fluid film between the rotating and stationary sealing faces resulting in loss of aerodynamic lift. Loss of aerodynamic lift at the seal faces can lead to a catastrophic failure scenario. (DGS 2012)

4. Maintenance optimisation

Many different types of maintenance optimization models have been developed and examples of such models can be found in (Dekker 1996; Dekker and Scarf 1998). The literature gives a nice overview of maintenance models distinguishing such policies as age replacement, block replacement, failure limit, repair limit, sequential, and repair counting; optimization criteria such as cost, availability, failure rate limits, cost and reliability; different time horizons over which the criteria are measured, and different types of repair. It is acknowledged that one of the reasons that many models have found limited application is that the assumptions made are typically rather strong. (Christer 1984) noted the preventive maintenance activity is considered to be the most difficult maintenance to model in the field of maintenance.

4.1 Weibull model

In estimating the failure rate of the dry gas seal the hazard function is used and a Weibull model is considered as adequate as it is a monotonic function capable of capturing the improvement and deterioration

of the dry gas seal. For the Weibull model for $t \geq 0$, this implies that the failure rate $\lambda(t) = h(t)$, therefore the expression of the hazard function of the Weibull Power law process is given as

$$h(t) = abt^{b-1} \quad (1)$$

The Reliability function is

$$R(t) = 1 - F(t) = \exp(-at^b) \quad (2)$$

and then the density function of the Weibull distribution is then

$$f(t) = abt^{b-1} \exp(-at^b) \quad (3)$$

The likelihood function for the power-law process will then be proportional to the product of the total probability of the intensities and it is given as

$$L(a, \beta; D) \propto \left\{ \prod_{i=1}^n f(t_i) \right\} \exp(-at_0^b). \quad (4)$$

Where D is the observed data set and α, β is the scale and shape parameters respectively. The probability of the underground cable failure intensities can be differentiated with respect to the parameter (α, β) and the resulting derivatives are set to zero to give the maximum. Thus, the log-likelihood is given as

$$\ell(a, b; D) = n \log a + n \log b + (b-1) \sum_{i=1}^n \log t_i - at_0^b. \quad (5)$$

Differentiating the function (5) above with respect to the shape parameter β and setting the derivative to zero. Multiplying both side of equation (6) by the shape parameter β

$$\frac{\partial \ell}{\partial b} : 0 + \frac{n}{b} + \sum_{i=1}^n \log t_i - at_0^b \log t_0 = 0 \quad (6)$$

we then substitute the scale parameter in equation (7) above, this can be obtained by differentiating the

$$\frac{\partial \ell}{\partial a} : n + b \sum_{i=1}^n \log t_i - a (\log t_0) b t_0^b = 0, \quad (7)$$

equation (5) with respect to α and setting the derivative to zero

$$\Rightarrow n + b \left(\sum_{i=1}^n \log t_i - n \log t_0 \right) = 0. \quad (8)$$

Therefore, the model closed form MLEs for the shape parameter β and scale parameter α is given as

$$b = \frac{n}{n \log t_0 - \sum_{i=1}^n \log t_i}, \quad a = \frac{n}{t_0^b}.$$

The maximum likelihood estimate of the shape and scale parameters is given above.

4.2 Weibull analysis

The Weibull distribution has been adopted to analyze the time-to-failure of dry gas seal. When it is used in predicting number of failures for those aging related populations of seals and where failure data is more likely to be homogeneous. In cases of early failures, the present authors applied the linear model to determine the goodness of fit of the Weibull distribution to two sets of data; one from the double seal and tandem seal respectively. As shown in Figure 5, each data point represents a failure that has been processed and the lines represent the linear fit of these points. R-Square is the degree of fitting. The closer R-Square value is to 1, the better the fitting is. Also in Figure 5, the horizontal axis is the logarithmic of “t” which stands for “time-to-failure” of the failed cables, while the vertical axis “F(t)” represents the cumulative distribution function of seals. The Weibull model give a good fitting result, as can be seen in Figure 5(a). However with the tandem seal failures, The R-square value of approximately 0.826 is not acceptable as shown in Figure 5(b) showing that the set of data does not obey Weibull distribution. The reason may be due to a lack of data homogeneity.

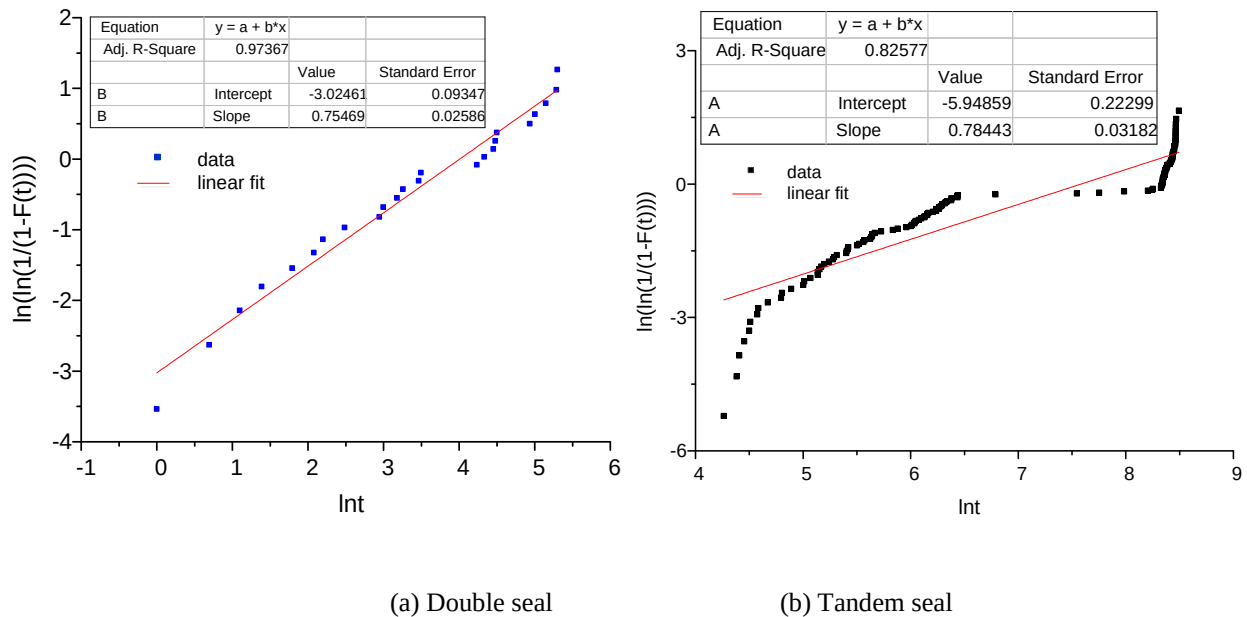


Figure 5: Analysis of Double and Tandem seal using Weibull distribution

5. Discussions

The main assumption of the proposed model is that T_i are drawn randomly from a Weibull distribution. From the data we can observe what causes the dry seal to fail. The actual data consists of a sequence of failure modes, which are followed by either maintenance or a failure. However, this reflects a practical issue that is nowadays quite often the case: Underlying systems reliability may be measured during maintenance. However, during operation there are lots of things, often external to the gas turbine seal failure modes, which can cause the seal to fail. In this way it is possible to have a high measured reliability during maintenance times and a much lower availability in practice. Condition monitoring information of the seals is also required to model how changes to current set up would affect availability. Investigation into the dataset collected is still underway and we intend to fit the proposed model and present results in future publications.

6. Conclusions

This paper presents failure analysis results conducted on dry gas seals of a GT in an LNG plant. The results from the application show that the approach is used to establish a link with reliability analysis. The RCM modeling approach is used to conduct Weibull analysis model. The application of the Weibull model to the dry gas seal failure data is still on-going and the results will be presented at the conference.

Acknowledgement

I would like to thank Dr Babakalli and Dr Octavian, for all the technical support and guidance to get this paper done on a professional manner.

References

- John S. Stahley (2001), Dry Gas Seal System Design Standards for Centrifugal Compressor Applications.
John S. Stahley (2001), Design, Operation, and Maintenance Considerations for Improved Dry Gas Seal Reliability in Centrifugal Compressors
API Standard 614, 1999 Lubrication Shaft-Sealing and Control-Oil Systems for Special-Purpose Application, Fourth Edition, API, Washington, D.C
Dry Gas Seal Training Manual 2012 Maintenance and Operation
Thorp, Joseph. (2008). Improving Dry Gas Seal Reliability
Brauer, D. and Brauer, G. (1987) Reliability-centered maintenance. Reliability, IEEE Transactions on, 36(1), pp.17-24. ISSN 0018-9529
Demers, J. (2005) Reliability centred maintenance. CIM Bulletin. Vol. 98(1086), pp. 50. ISSN 0317-0926.
Moubray, J. (1991) Reliability-centered maintenance. New York: Industrial Press.
Moubray, J. (1997) Reliability-centered maintenance. New York: Industrial Press.
Ramli, R. & Arffin, M.N. (2012) Reliability Centered Maintenance in Schedule Improvement of Automotive Assembly Industry. American Journal of Applied Sciences. Vol. 9, no. 8, pp. 1232-1236. ISSN 1546-9239
Compressor training manual 2011 Operation and Maintenance
Adgas das Inland compressor training 2006 Turbomachinery

- Christer, A. H. and W. M. Waller. 1984. Operational Research Approach to Planned Maintenance: Modelling PM for a Vehicle Fleet. *Journal of the Operational Research Society* **35**: 967-984.
- Crocker, J. and U. D. Kumar. 2000. Age-related maintenance versus reliability centred maintenance: a case study on aero-engines. *Reliability Engineering & System Safety*, Volume 67, Issue 2, February 2000, Pages 113-118
- Dekker, R., 1996. Applications of maintenance optimization models: A review and analysis. *Reliability Engineering & System Safety* **51**(3): 229-240.
- Dekker, R. and P. A. Scarf. 1998. On the impact of optimisation models in maintenance decision making: the state of the art. *Reliability Engineering & System Safety* **60**(2): 111-119.

A selective maintenance model based on FITradeoff method

L. Ribeiro*, P. Dantas** and C. Cavalcante***

* Departamento de Eng. de Produção, Universidade Federal de Pernambuco, Brazil- lucasfredericoeing@hotmail.com

** Departamento de Eng. de Produção, Universidade Federal de Pernambuco, Brazil- priscillapld@hotmail.com

***Departamento de Eng. de Produção, Universidade Federal de Pernambuco, Brazil- cristiano@ufpe.br

Abstract. Some aspects of a system operation dynamics may prevent excellence in organizations performance. This is imperative for mission-oriented systems since all necessary interventions have to be undertaken during the stoppages between successive missions. Maintenance models that deal with it, however, do not take into account that aspects, such as the length of stoppages or constraints on the maintenance resources. These constraints limit the set of actions that can be carried out during each stoppage. Thus, it should be selected the maintenance actions that would best improve system performance, observing constraints. Traditionally, models that address this issue seek to optimize one of the performance indicators of a system. Thus, an original contribution of this paper is to optimize conflicting indicators through a multicriteria perspective by considering the preferences of the decision maker. Therefore, it has been used FITradeoff (Flexible and Interactive Tradeoff) method in order to asses a global performance value to each alternative. At last, the obtained plan is compared to the one obtained in a traditional way.

1. Introduction

The performance of organizations depends on the choice of an appropriate maintenance strategy (BOGONOV, MARSEGUERRA and ZIO, 2000). This is evident for mission-oriented systems, since the set of maintenance actions to be executed in this system depends on the duration of stoppages between missions (PANDEY et al., 2013).

In addition to the fact that the majority of the models dedicated to determining maintenance strategy, in that case, neglect the length of the stoppages, they ignore constraints on the maintenance resources, which also limits the set of actions to be performed (MAILLART et al., 2009).

Thus, one should identify, among the possible actions, those that best improve system performance, while not violating constraints. This problem is called Selective Maintenance Problem (LIU and HUANG, 2010). The traditional formulations of this problem aim to determine a subset of actions that produces a great system performance in terms of maintenance cost, average system reliability or maintenance duration.

Depending on the context, system performance is better described in terms of one of those indicators. Therefore, it the indicators sometimes as the objective function, sometimes as a constraint. In a complementary way, Ribeiro, Cavalcante and Rodriguez (2017) embedded the decision maker (DM) aspirations into the model as a reliability constraint. Besides, authors pointed the importance of contemplating DM preferences in a deeper perspective - through treating selective maintenance as a multiple objective problem. This insight figures out as the original contribution of this paper.

2. Methodology

Multicriteria Decision Aid (MCDA) approach considers DM preferences in order to unify objectives in a multiple objective environment. Preferences, in turn, are based on a subjective evaluation of each decision criteria (DE ALMEIDA et al., 2015).

This evaluation, however, is not an easy task since it depends on the knowledge and understanding of the objectives and preferences by the DM (GUSMÃO and MEDEIROS, 2016). This fact, combined to the compensatory characteristic of the selective maintenance criteria, delimitate the choice of a Multicriteria Decision Method to tackle with the problem.

Among the possible choices, we elect FITradeoff (Flexible and Interactive Tradeoff) method to elicitate DM's preferences. This method, developed by (De Almeida et al., 2015), is based on the traditional trade-off procedure, conserving its axiomatic structure, but demanding less cognitive effort from the DM, due to the use of partial information (DELL'OVO et al., 2017). Thus, instead of requiring an indifference relation between two values, this method requires a strict preference, turning the process softer.

FITradeoff uses a decision support system (DSS) to achieve results. While the DM shows their preferences, the DSS evaluates the alternatives through a linear programming problem (LPP). The process finishes when a unique solution is found. This process is illustrated in Figure 1.

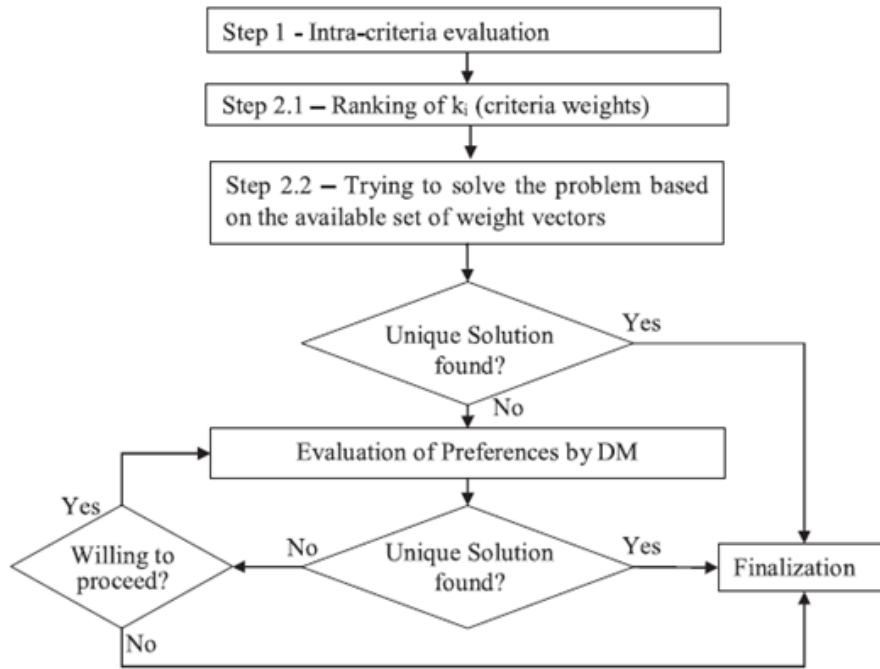


Figure 1 - Source: De Almeida et al (2015)

2. System Description and Criteria Definition

Consider a system S composed by n subsystems S_i ($i=1, \dots, n$) connected in series, each consisting of N_i components C_{ij} ($j=1, \dots, N_i$) arranged in parallel. This system operates M missions, with known duration $U(m)$ ($m=1, \dots, M$), such that, after a mission, a stoppage, also with known duration $D(m, l)$ ($m=1, \dots, M-1$; $l=m+1$), figures as an opportunity for executing maintenance.

Assume that each of the components may be in operational or failed state. Failure occurs within a mission and is related to the component's virtual age (NAKAGAWA, 1988). Setting $A_{ij}(m)$ and $B_{ij}(m)$, the virtual ages of component C_{ij} , respectively, at the beginning and at the end of the m -th mission, the reliability $R_{ij}(m)$ of component C_{ij} is obtained from:

$$R_{ij}(m) = \exp \left(- \int_{A_{ij}(m)}^{B_{ij}(m)} h_{ij}(t) dt \right). \quad (1)$$

where $h_{ij}(t)$ corresponds to the risk function of component C_{ij} . From equation 1, it follows that the subsystem S_i and system S reliabilities, denoted by $R_i(m)$ and $R(m)$, are given by:

$$R_i(m) = 1 - \prod_{j=1}^{N_i} (1 - R_{ij}(m)). \quad (2)$$

$$R(m) = \prod_{i=1}^n R_i(m). \quad (3)$$

Such system is minimally repaired, if a component fails during a mission. This intervention restores the operational state of the component, has negligible duration and a cost cmr_{ij} inherent to component C_{ij} . Within a stoppage, components can be preventively maintained, either perfectly or imperfectly. These effects are modeled by an age reduction coefficient $\alpha(p) \in [0, 1]$, inherent to action p , that updates components' virtual ages.

In this sense, the P preventive actions that can be performed on system's components are displayed in vector $VPM = [a_1, \dots, a_p, \dots, a_P]$. For every action a_p ($p = 1, \dots, P$) it is defined an age reduction coefficient $\alpha(p)$, a cost $cpm(a_p)$ and a duration $dpm(a_p)$, as well as a set of components on which the action can be performed $Comp(a_p)$.

The total maintenance cost C_{total} consists in the sum of three parcels: the minimal repair cost CMR; the preventive maintenance cost CPM; and a penalty cost CPAR, related to the system failure probability during a mission. These values are explicit by the equations:

$$CMR = \sum_{m=1}^M \sum_{i=1}^n \sum_{j=1}^{N_i} cmr_{ij} \int_{A_{ij}(m)}^{B_{ij}(m)} h_{ij}(x) dx. \quad (4)$$

$$CPM = \sum_{i=1}^n \sum_{j=1}^{N_i} \sum_{m=1}^{M-1} \sum_{p=1}^P (cpm(a_p) \cdot \alpha_p(C_{ij}, m)). \quad (5)$$

where $a_p(C_{ij}, m)$ is a binary decision variable which assumes non-zero value if the action a_p is performed on component C_{ij} at the m -th stoppage.

$$CPAR = \sum_{m=1}^M \left(1 - \prod_{j=1}^n (1 - \prod_{i=1}^{N_i} (1 - R_{ij}(m))) \right) cpar. \quad (6)$$

where $cpar$ is a fee, related to system failure probability.

Total maintenance duration DPM is given by the expression:

$$DPM = \sum_{m=1}^{M-1} \sum_{p=1}^P \sum_{i=1}^n \sum_{j=1}^{N_i} dpm(a_p) \cdot a_p(C_{ij}, m). \quad (7)$$

3. Numerical Example

In order to illustrate the application of the multicriteria approach, it have been made a numerical application. Therefore, we used an instance based on that found in Ribeiro, Cavalcante and Rodriguez (2017). The instance consists in a series-parallel system consisting on $n=2$ subsystems, compound, respectively, by $N_1 = 2$ and $N_2 = 4$ components. The reliability of each component is established from the risk function of a Weibull distribution, which shapes γ_{ij} and scale β_{ij} parameters, as well as the minimum repair costs cmr_{ij} are detailed in table 1.

Table 1. Parameters of system components.

C_{ij}	C_{11}	C_{12}	C_{21}	C_{22}	C_{23}	C_{24}
β_{ij}	1.2	1.4	2.8	1.5	2.4	2.2
γ_{ij}	26	35	40	35	30	45
cmr_{ij}	3	6	7	4	6	3

This system is oriented to perform 2 missions of 25 time units each, such that between this missions it is planned a stoppage lasting for 5 time units; $cpar$, in turn, is settled in 35 monetary units. Furthermore, Vector VPM and the parameters of the P preventive actions that can be performed on the system are shown in table 2.

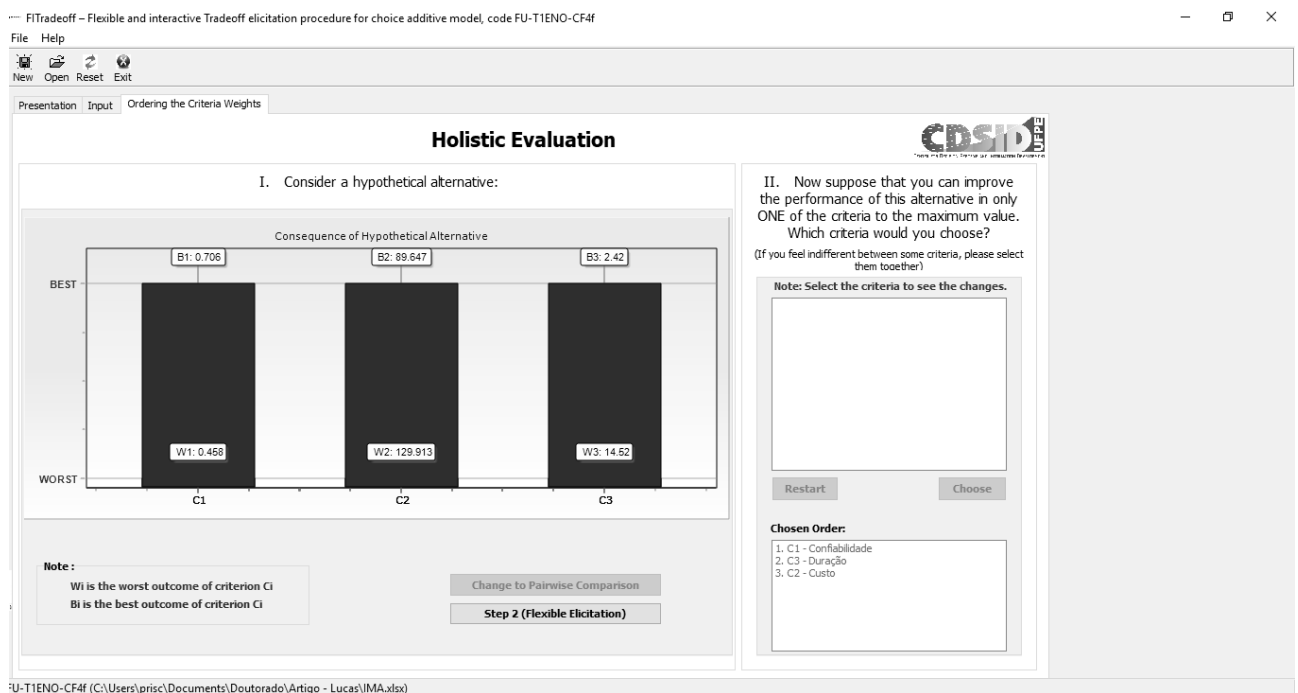
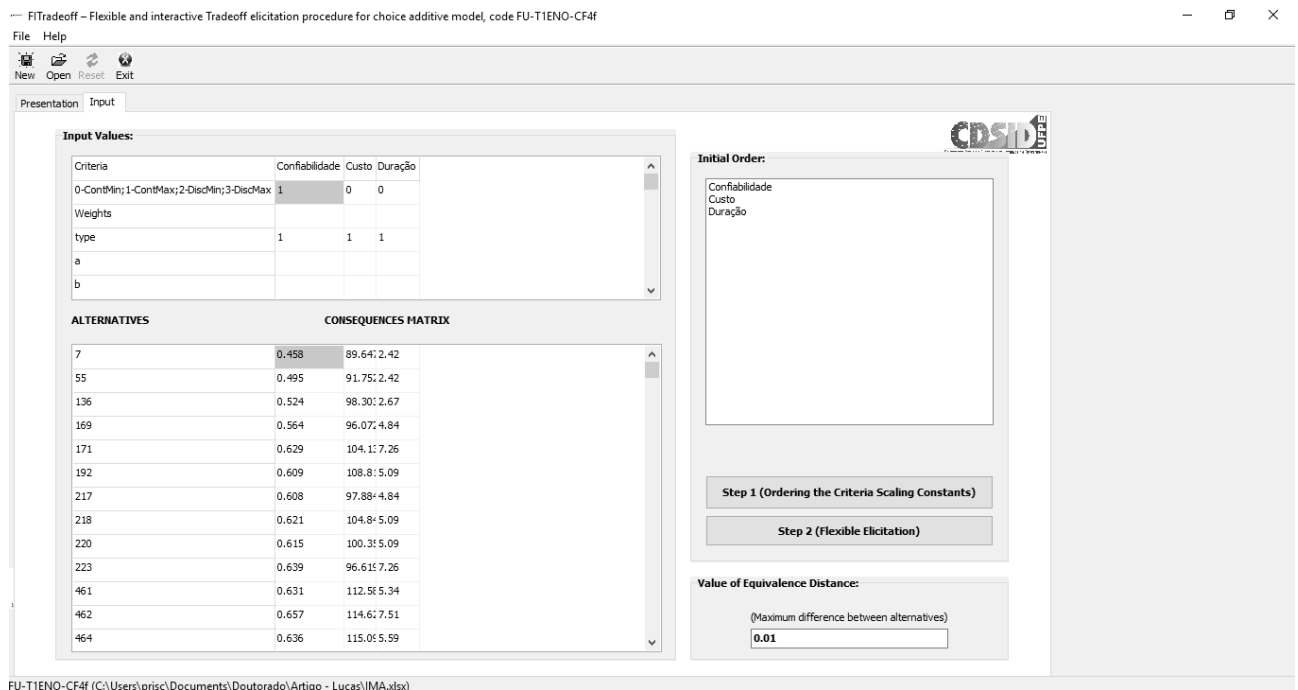
Table 2. Vector VPM configuration

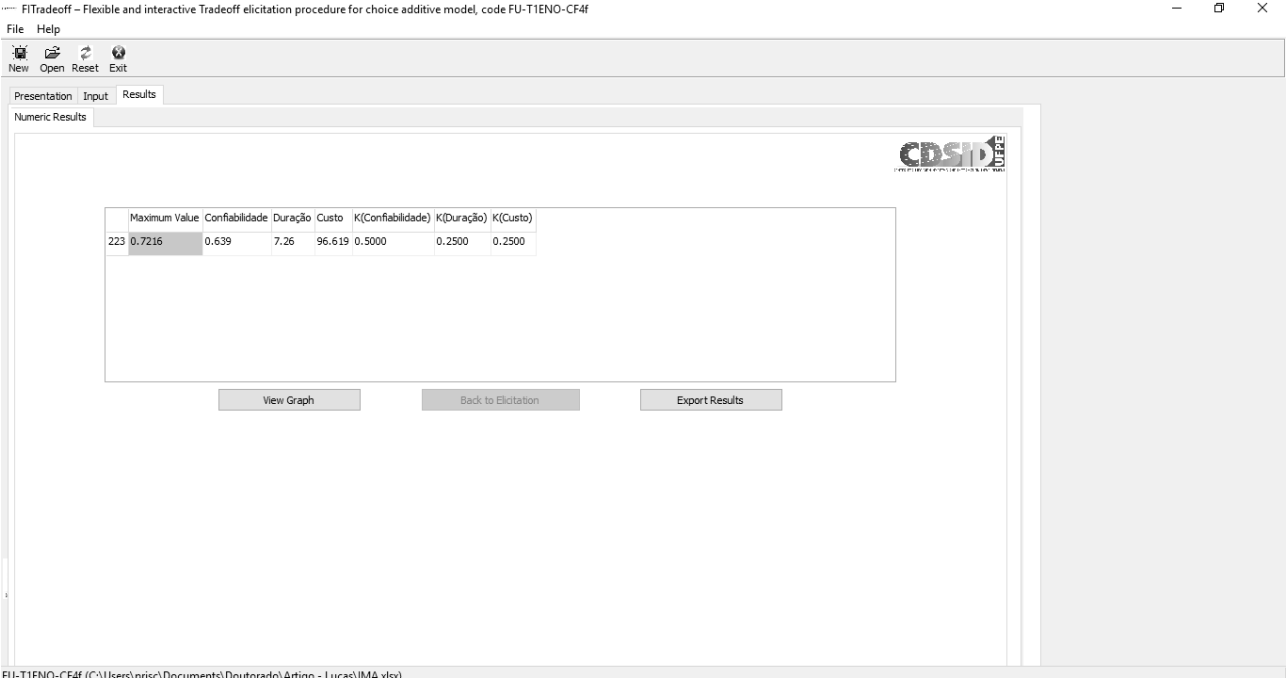
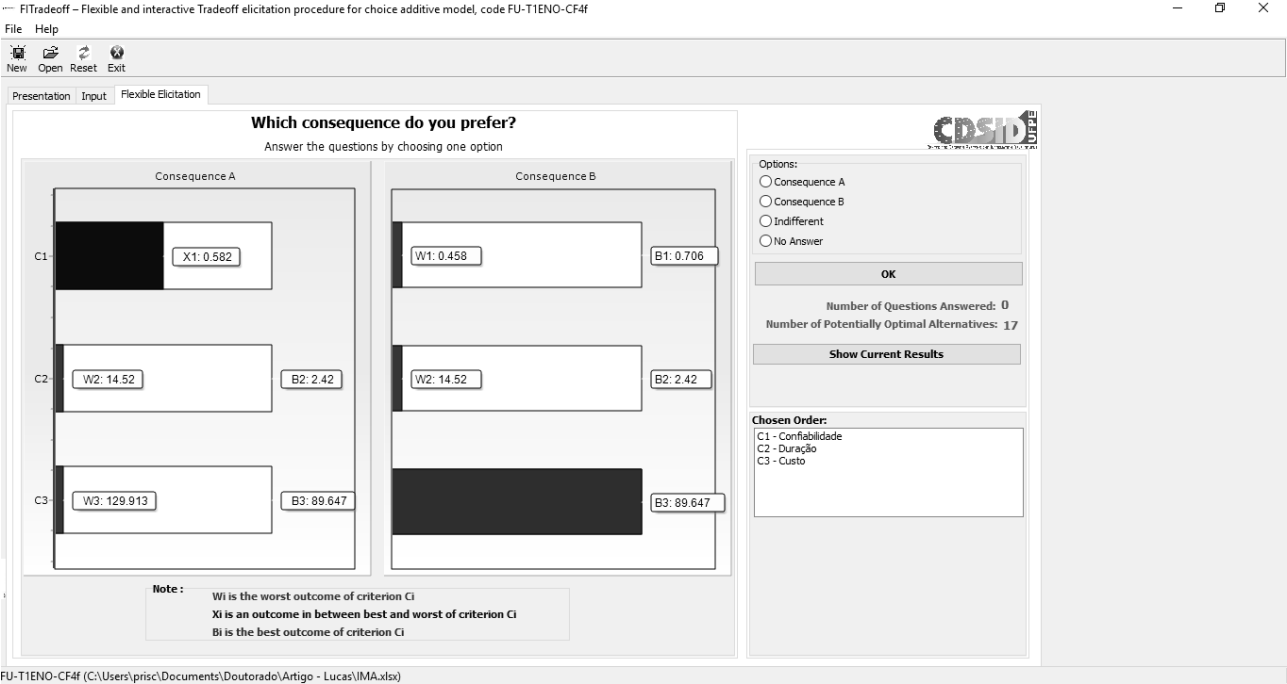
p	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Comp	C_{11}			C_{12}			C_{21}			C_{22}			C_{23}			C_{24}		
α	1.0	0.6	0.0	1.0	0.5	0.0	1.0	0.6	0.0	1.0	0.4	0.0	1.0	0.3	0.0	1.0	0.4	0.0
com	0.0	8.3	12.5	0.0	9.2	15.4	0.0	6.0	13.9	0.0	7.9	12.0	0.0	6.7	14.0	0.0	7.8	16.5
dpm	0.0	0.3	2.4	0.0	0.4	1.8	0.0	0.3	1.8	0.0	0.5	2.3	0.0	0.5	3.0	0.0	0.5	2.6

Due to the small magnitude of the instance, it was efficient using exhaustive enumeration to list all the maintenance plans that could be executed during the stoppage. Each plan figures as an alternative for the FITradeoff method. Plans are represented by matrices such that a cell value indicates the kind of action to be performed in the component given by the matrix column. For that, components are sorted from the first component in the first subsystem to the last component in the last subsystem.

Overall, $3^6 = 729$ alternatives have been generated. of wich 152 were dominated ones. FITradeoff method was applied for the lasting alternatives. For this we used a FITradeoff application software developed by CDSID, a laboratory at Federal University of Pernambuco, Brazil.

As previously announced, the preference relations (instead of indifference relations) of FITradeoff method speed up the decision procedure. We made a simulated application of the method (Figure 2), firstly a rank of the weight criteria was indicated, as presented in Figure 3, and, after three simple questions (Figure 4), it was found the alternative that better fits DM preferences (Figure 5). This alternative is plan number 223 [1 6 9 10 15 16], which indicates that components 2, 3 and 5 should be replaced in the stoppage.





Applying the traditional approach in a service context (minimize maintenance duration, observing cost and reliability constraints), alternative 217 [1 6 9 10 13 16] – that indicates a replacement on components 2 and 3 during the stoppage - figures as the best one. Table 3 states a comparison between the two alternatives.

Table 3. Comparison between the performance indicators for both alternatives.

Alternative	217	223
Reliability	0.61	0.64
Cost	97.9	96.6
Duration	4.8	7.3

While the traditional approach gives the alternative with the smallest maintenance duration, observing the cost (<100) and reliability (>60) constraints; FITradeoff gives an alternative that translates DM preferences. Thus, multicriteria approach clearly modifies problem perspective.

4. Conclusions

This paper contributes to the Selective Maintenance since it brings a new perspective of treatment to the problem. A numerical application illustrate the easy and intuitive use of the method. In comparison to the traditional approach, multicriteria method conducted to a solution that fits DM preferences. Future work should invest in larger horizon (multiple missions) problems, following the literature trend.

Acknowledgement

The research of L. Ribeiro have been supported by Fundação de Amparo à Ciência e Tecnologia do Estado de Pernambuco (FACEPE); P. Dantas and C. Cavalcante research is funded by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

References

- Bogonovo E, Marseguerra M, Zio E (2000) A Monte Carlo methodological approach to plant availability modelling with maintenance, aging and obsolescence. *Reliability Engineering & System Safety* 67(1), pp. 61–73.
- De Almeida AT, De Almeida JA, Costa AP, De Almeida-Filho AT (2016) A new method for elicitation of criteria weights in additive models: flexible and interactive tradeoff. *European Journal of Operational Research* 250 (1), pp. 179–191.
- Dell'Ovo M, Frej EA, Oppio A, Capolongo S, Morais DC, De Almeida AT (2017) Multicriteria Decision Making for Healthcare Facilities Location with Visualization Based on FITradeoff Method. In: *Decision Support Systems VII. Data, Information and Knowledge Visualization in Decision Support Systems: Third International Conference, ICDSST* pp.23-44.
- Gusmão AP; Medeiros CP (2016) A model for selecting Information system using the FITradeoff. *Hindawi Publishing Corporation – Mathematical Problems in Engineering*.
- Liu Y, Huang HZ (2010) Optimal selective maintenance strategy for multi-state systems under imperfect maintenance. *IEEE Transactions on Reliability* 59(02) pp. 356–367.
- Ribeiro L, Cavalcante C, Rodriguez, T (2017) Selective maintenance for multi mission oriented systems. In: *Risk, Reliability and Safety: Innovating Theory and Practice* (Walls, Revie & Bedford, Eds) pp. 1179-1184.
- Maillart LM, Cassady CR, Rainwater C, Schneider K (2009) Selective maintenance decision-making over extended planning horizons. *IEEE Transactions on Reliability* 58(03), pp. 462 – 469, 2009.
- Nakagawa T (1998) Sequential imperfect preventive maintenance policies. *IEEE Transactions on Reliability* 37(3), pp.295-308.
- Pandey M, Zuo MJ, Moghaddass R, Tiwari MK (2013) Selective maintenance for binary systems under imperfect repair. *Reliability Engineering and System Safety* 113, pp.42 – 51.

Selective maintenance model for mission-oriented systems considering heterogeneous missions and budget constraints

L. Ribeiro* and C. Cavalcante**

* Departamento de Eng. de Produção, Universidade Federal de Pernambuco, Brazil- lucasfredericoeing@hotmail.com

** Departamento de Eng. de Produção, Universidade Federal de Pernambuco, Brazil- cristiano@ufpe.br

Abstract. Maintenance planning is a challenge for organizations. This is particularly evident for mission-oriented systems, since maintenance must to be performed within stoppages. In this sense, models indicate a set of actions to be performed in each stoppage by considering the operational dynamics of a system. We propose a model that address this issue in an innovative way, by considering a system working under different operational conditions – i.e. heterogeneous missions – and subject to a budget constraint. We used genetic algorithm (GA) in order to solve this problem have. Finally a case study illustrates the model applicability.

1. Introduction

System's performance is dependent on effective assistance. This is a challenge for mission-oriented systems since interventions have to be done during stoppages between missions (LIU, XIE and KUO, 2016). Maintenance models, however, usually neglects the system's operational dynamics as well as constraints on the maintenance resources (DO and BARROS, 2017; RICE, CASSADY and NACHLAS, 1998). These aspects, however, limit the set of actions that may be executed (RICE, CASSADY and NACHLAS, 1998; MAILLART et al., 2009). Thus, one should elect, the actions that best improve system's performance, not violating such aspects (KHATAB and AGHEZZAF, 2016).

In this sense, one contribution of this paper is adopting a limited maintenance budget. Moreover, we deal with maintenance costs in an innovative way considering a penalty fee related to system failures during missions. We also consider a system working under different operational conditions - heterogeneous missions.

Further, we explore service-oriented systems. In this context, stoppages implies in loss of service to customers; so its duration must be minimized. Despite this objective is uncommon in literature (CASSADY, POHL and MURDOCK, 2001), it may increase system's efficiency.

2. Problem formulation

First, it is necessary to point that a great part of this section is derived from Ribeiro, Cavalcante and Rodriguez (2017).

Consider a system S composed by n subsystems S_i ($i=1, \dots, n$) connected in series, each consisting of N_i components C_{ij} ($j=1, \dots, N_i$) arranged in parallel. This system must perform a sequence of M missions, with known duration $U(m)$ such that between two consecutive missions must to be a stoppage, also with known duration $D(m, l)$ ($l=m+1$).

Components may be in operational or failed state. Failure occurs within a mission and is related to component's virtual life (NAKAGAWA, 1998). Setting $A_{ij}(m)$ and $B_{ij}(m)$ the virtual ages of component C_{ij} , respectively, at the beginning and at the end of the m -th mission, the reliability $R_{ij}(m)$ of component C_{ij} is determined by the equation:

$$R_{ij}(m) = o(m) \exp \left(- \int_{A_{ij}(m)}^{B_{ij}(m)} h_{ij}(t) dt \right). \quad (1)$$

where $h_{ij}(t)$ corresponds to the risk function of component C_{ij} . The parameter $o(m)$, in turn, models heterogeneity through the relation between the designed and the expected system's operation condition, i.e. it assumes values smaller than one if the expected operation condition is harder than the designed one. This parameter generalizes the problem stated in Ribeiro, Cavalcante and Rodriguez (2017) and modifies the subsystem S_i and system S reliabilities, respectively denoted by $R_i(m)$ and $R(m)$, impacting in following equations.

$$R_i(m) = 1 - \prod_{j=1}^{N_i} (1 - R_{ij}(m)). \quad (2)$$

$$R(m) = \prod_{i=1}^n R_i(m). \quad (3)$$

Such system is minimally repaired, if a component fails during a mission. This intervention restores the operational state of the component, has negligible duration and a cost cmr_{ij} inherent to component C_{ij} . Within a stoppage, components can be preventively maintained, either perfectly or imperfectly. These effects are modeled by an age reduction coefficient $\alpha(p) \in [0,1]$, inherent to action p , that updates components' virtual ages.

The set of the P preventive actions that can be performed on system's components is displayed in vector $VPM = [a_1, \dots, a_p, \dots, a_P]$. Every action a_p ($p = 1, \dots, P$) is attached to age reduction coefficient $\alpha(p)$, a cost $cpm(a_p)$ and a duration $dpm(a_p)$, as well as a set of components on which the action can be performed $Comp(a_p)$.

Additionally to system's reliability, it is important to highlight others indicators. Thus, total maintenance cost C_{total} consists in the sum of three parcels: the minimal repair cost CMR ; the preventive maintenance cost CPM ; and a penalty cost $CPAR$, related to the system failure probability during a mission.

$$CMR = \sum_{m=1}^M \sum_{i=1}^n \sum_{j=1}^{N_i} cmr_{ij} \int_{A_{ij}(m)}^{B_{ij}(m)} h_{ij}(x) dx. \quad (4)$$

$$CPM = \sum_{i=1}^n \sum_{j=1}^{N_i} \sum_{m=1}^{M-1} \sum_{p=1}^P \left(cpm(a_p) \cdot a_p(C_{ij}, m) \right). \quad (5)$$

where $a_p(C_{ij}, m)$ is a binary decision variable which assumes non-zero value if the action a_p is performed on component C_{ij} at the m -th stoppage.

$$CPAR = \sum_{m=1}^M \left(1 - \prod_{j=1}^n (1 - \prod_{i=1}^{N_i} (1 - R_{ij}(m))) \right) cpar. \quad (6)$$

where $cpar$ is a fee, related to system failure probability.

Total maintenance duration DPM is given by the expression:

$$DPM = \sum_{m=1}^{M-1} \sum_{p=1}^P \sum_{i=1}^n \sum_{j=1}^{N_i} dpm(a_p) \cdot a_p(C_{ij}, m). \quad (7)$$

In the case of service-oriented systems, performance is indirectly described via stoppage duration. Still, system's reliability $R(m)$ should not be smaller than technical recommendations $R(l)$ nor the maintenance budget BGT should be trespassed. Thus, determine a subset of VPM that minimizes the total maintenance duration (8), enables the achievement of the desired reliability levels (9) and honors the budget constraint (10).

$$P1: \text{Minimize } DPM \quad (8)$$

Subject to:

$$R(m+1) \geq R(l) \quad (9)$$

$$C_{total} \leq BGT \quad (10)$$

$$\sum_{v=1}^P a_v(C_{ij}, m) \leq 1 \quad (11)$$

$$a_v(C_{ij}, m) \in \{0,1\} \quad (12)$$

$$i = 1, \dots, n; j = 1, \dots, N_i; p = 1, \dots, P; m = 1, \dots, M-1; l = m+1 \quad (13)$$

Constraint (11) stands that no more than one action can be performed on a component within a stoppage. Further, we have chosen genetic algorithm as solution approach. For details, also refer to Ribeiro, Cavalcante and Rodriguez (2017).

3. Illustrative case

In order to demonstrate models' applicability, we took advantage of the coal transportation system described in (LIU, XIE and KUO, 2016). The system consists of five subsystems compound, respectively, by 3, 2, 3, 2 and 4 components. Components' lives follow Weibull(β, η) distributions which parameters and minimum repair costs cmr_{ij} are displayed in Table 1.

Table 1. System components' parameters.

C_{ij}	C_{11}	C_{12}	C_{13}	C_{21}	C_{22}	C_{31}	C_{32}	C_{33}	C_{41}	C_{42}	C_{51}	C_{52}	C_{53}	C_{54}
β_{ij}	1.5	2.4	1.6	2.6	1.8	2.4	2.5	2	1.2	1.4	2.8	1.5	2.4	2.2
η_{ij}	37.5	57	42	60	42	51	39	42	39	52.5	60	52.5	45	67.5
cmr_{ij}	3	4	3	5	2	3	6	5	3	6	7	4	6	3

This system is oriented to perform a sequence of 15 missions whose durations, operation conditions and reliability aspiration levels are shown in Table 2.

Table 2. Durations, operation conditions and reliability levels for system.

U(m)	11	12	11	13	11	13	10	10	13	11	11	13	13	10	13
o(m)	1	1.05	1	1.1	1	1.1	0.95	0.95	1.1	1	0.9	1	1	0.85	1
R(m+1)	0.8	0.8	0.8	0.8	0.8	0.75	0.75	0.75	0.75	0.7	0.7	0.7	0.7	0.7	-

The set of the preventive actions that can be performed on system's components and their features are presented in Table 3.

Table 3. Parameters of the P preventive actions that can be performed in the system.

P	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Comp(p)	C_{11}	C_{11}	C_{12}	C_{12}	C_{13}	C_{13}	C_{21}	C_{21}	C_{22}	C_{22}	C_{31}	C_{31}	C_{32}	C_{32}
$\alpha(p)$	0.6	0	0.46	0	0.52	0	0.68	0	0.41	0	0.36	0	0.55	0
cpm(p)	6.9	14.1	8.2	18	6.9	14.1	9.7	15.5	9	14.2	8.1	16	8.3	16.25
dpm(p)	0.25	0.55	0.35	0.65	0.25	0.37	0.3	0.58	0.37	0.75	0.25	0.75	0.87	1.5

P	15	16	17	18	19	20	21	22	23	24	25	26	27	28
Comp(p)	C_{33}	C_{33}	C_{41}	C_{41}	C_{42}	C_{42}	C_{51}	C_{51}	C_{52}	C_{52}	C_{53}	C_{53}	C_{54}	C_{54}
$\alpha(p)$	0.62	0	0.61	0	0.49	0	0.6	0	0.43	0	0.25	0	0.43	0
cpm(p)	7.3	16.5	8.3	12.5	9.2	15.4	6	13.9	7.9	12	6.7	14	7.8	16.5
dpm(p)	0.75	1.65	0.25	0.42	0.38	0.8	0.25	1	0.45	0.95	0.45	1.15	0.52	1.55

The fee $cpar$ resulting from the system's failure probability during a mission was settled in 100 currencies and the maintenance budget in 1900 currencies. As previously announced, we choose Genetic Algorithm as solution technic. Table 4 displays the maintenance plan obtained for the formulation in comparison to the homogeneous conditions case (P2).

Table 4. Maintenance plans obtained for cases P1 and P2.

Case/ Stoppage	P1	P2
1	3,10,11,18,20,22,27	5,9,15,17,21,25
2	1,8,10,14,16,20,22	7,10,12,13,18,20,22,23,26,27
3	1,4,6,8,10,14,16,18,22,23,27	2,5,9,11,18
4	3,5,7,10,11,21,26,27	2,3,6,7,12,17,19,21,24,26,28
5	1,4,13,19,23,26	1,10,11,13,15,19,22,24
6	2,5,7,10,14,15,18,19,21,23,28	6,7,11,19,25,28
7	2,3,5,7,9,12,20,24	1,4,6,8,15,20,21,24,25,28
8	1,6,8,10,11,13,21,25	2,10,11,16,18,19,21,23,26,28
9	3,6,8,10,13,20,22,25	3,7,11,18,22,23
10	1,4,5,8,9,11,22	1,5,7,9,23
11	4,5,7,10,15,17,19,21,25	1,5,8,10,17,21,23,26,28
12	1,5,7,9,11,13,16,17,20,25	4,6,7,9,14,19,22,23,25,27
13	2,5,9,14,16,18,21,23,26,27	1,3,5,15,18,26,28
14	2,6,7,14,17,26,27	2,7,11,13,16,18,19,22,23,25

At a first glance, there is no pattern between plans, however the total interventions number is similar in cases P1 (117) and P2 (113). The amount of perfect actions are also very similar between plans (55 for P1 and 48 for P2). This little difference may be explained by the necessity of more effective interventions in the missions in which the system is subject to severe operational conditions.

The more effective the intervention (replacement) the more expensive it is. Despite replacements are more frequent in case P1, its total maintenance cost is similar to the P2 cost (table 5). This is a consequence of a balance between severe and soft operational conditions for P1 (table 2). Thus it is necessary a skew to consider the effects of heterogeneity in models.

Table 5. Expected performance values for cases P1 and P2.

Case / Indicator	P1	P2
Maintenance Cost	1590.38	1586.36
Maintenance Duration	71.31	67.42
Average Reliability	0.91	0.90

4. Conclusions

Although a proper maintenance policy is essential to system's performance, organizations adopt models that despise its dynamics of operation. Selective maintenance models deals with such simplification. This paper contributes to that category of models, in part, by considering the system working under different operational conditions - heterogeneous missions. This assumption, however, should reflect the system's operating conditions, since adopting it brings unnecessary complexity to model in the case of homogeneous missions or balanced condition horizons.

Acknowledgement

The research of L. Ribeiro have been supported by Fundação de Amparo à Ciência e Tecnologia do Estado de Pernambuco (FACEPE); and C. Cavalcante research is funded by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

References

- Cassady CR, Pohl AE, Murdock WP (2001) Selective maintenance modeling for industrial systems. *Journal of Quality in Maintenance Engineering* 7(02), pp.104–117.
- Do P, Barros A, (2017) Maintenance grouping models for multicomponent systems. In *Mathematics Applied to Engineering* (Ran and Davim, Eds), Academic Press, pp. 147 -170.
- Khatab A, Aghezzaf EH (2016) Selective maintenance optimization when quality of imperfect maintenance actions are stochastic. *Reliability Engineering and System Safety* 150, pp. 182 – 189.
- Liu B, Xie M, Xu Z, Kuo W (2016) An imperfect maintenance policy for mission-oriented systems subject to degradation and external shocks. *Computers and Industrial Engineering* 102, pp. 21–32.
- Nakagawa T (1998) Sequential imperfect preventive maintenance policies. *IEEE Transactions on Reliability* 37(3), pp.295-308.
- Ribeiro L, Cavalcante C, Rodriguez, T (2017) Selective maintenance for multi mission oriented systems. In: *Risk, Reliability and Safety: Innovating Theory and Practice* (Walls, Revie & Bedford, Eds) pp. 1179-1184.
- Rice WF, Cassady CR, Nachlas JA (1998) Optimal maintenance plans under limited maintenance time. In: *Proceedings of the Seventh Industrial Engineering Research Conference*, s.d.

Reliability Demonstration Tests Considering Performance Degradation with Measurement Error

C. Ruiz*, H. Liao**, E.A. Pohl***

*University of Arkansas, Department of Industrial Engineering, caruizto@uark.edu

**University of Arkansas, Department of Industrial Engineering, liao@uark.edu

***University of Arkansas, Department of Industrial Engineering, epohl@uark.edu

Abstract Reliability demonstration tests (RDTs) have been widely used in engineering design to verify if a product has met a certain reliability requirement. Such tests are usually conducted and analyzed based on binomial theory for the number of failures or analysis of failure times. Unlike these traditional methods, a degradation-based RDT method is proposed in this paper. Appropriate implementation of this method will speed up reliability demonstration, especially for highly reliable products. However, a big challenge is the measurement error that cannot be avoided in degradation data collection. To incorporate the impact of measurement errors in degradation-based RDT, a random effects stochastic process model explaining both the evolution of product degradation and measurement error is proposed. Under this model, a statistical inference method based on an expectation-maximization algorithm is developed to estimate the model parameters. Moreover, the optimal design of degradation-based RDT is developed to minimize the total testing cost considering both the producer's and consumer's risks. A numerical example is presented to illustrate the use of the proposed RDT method in practice.

1. Introduction

Reliability Demonstration Testing (RDT) has been widely used to verify if a product has met a certain reliability requirement. In the literature, two groups of RDT methods have been developed for non-repairable systems based on either the number of failures (Leemis (2006)) or failure times (Lawless (1978); McKane et al. (2005)). Guo and Liao (2012) provide detailed discussions on the relationship between the two groups of methods and illustrate when they will lead to similar RDT designs. In practice, the method based on the number of failures is the most commonly used. This method is called "binomial test" as it uses the binomial equation for calculating the design variable based on the given quantities such as the confidence level, the maximum number of failures allowed, and lower bound on the demonstrated reliability. The drawback of this method is that the test must be conducted up to the target lifetime. Unlike the binomial test method, a parametric alternative is to perform reliability demonstration by fitting a failure time distribution based on observed failure times. The limitation of this method is that it would be difficult, if not impossible, to estimate the model parameters if no failures are observed in the test. For repairable systems, Guo et al. (2011) develop an RDT design method applicable for both Homogeneous Poisson Process and Non-homogeneous Poisson Process cases. The method can effectively determine the required sample size and testing time to demonstrate the required Mean-Time-Between-Failures. Apparently, these traditional methods are not efficient for highly reliable products with high reliability requirements.

Degradation data analysis has been extensively studied in reliability engineering (Meeker and Escobar (1998); Liao and Elsayed (2006); Ye and Chen (2014)). Compared to failure time data analysis, this type of approach can speed up reliability modeling without waiting for actual failures to occur. Moreover, it can provide valuable insights into the underlying physics of failure. Although degradation data analysis has shown many advantages in reliability estimation, degradation-based RDT has not been reported in the literature. This paper focuses on the design and analysis of degradation-based RDT, which will provide a new direction for improving the efficiency of RDT. To the best of our knowledge, this is the first study on the design and analysis of degradation-based RDT.

When performing degradation analysis, a big challenge is how to handle measurement errors. Without considering the impact of measurement errors, the developments of a degradation model for the actual degradation process and of the optimal RDT plan will be misled by the contaminated degradation data. To incorporate the

impact of measurement errors in degradation-based RDT, a stochastic process model explaining both the evolution of actual product degradation and measurement errors is proposed in this paper. In particular, a random effects Inverse Gaussian process is used to model the actual degradation processes for different units while the measurement error is described by a normal probability distribution.

In planning RDT, the producer's and consumer's risks are usually considered. In the context of RDT, the producer's risk is the probability of rejecting a product when its reliability is higher than the target value. Similarly, the consumer's risk is the probability of accepting a product when its reliability is lower than the target value. In this paper, the optimal design of RDT plan is developed to minimize the total testing cost considering both the producer's and consumer's risks. To assist in solving the optimization problem, a method for checking the problem feasibility is developed considering the producer's and consumer's risks.

The remainder of this paper is organized as follows. Section 2 provides the random effects degradation model considering measurement errors as well as the corresponding statistical inference method. The optimal design of degradation-based RDT is addressed in Section 3, where both the producer's and consumer's risks are considered. In Section 4, a numerical example is provided to illustrate the use of the proposed degradation-based RDT method in practice. Finally, Section 5 draws conclusions.

2. Degradation-based Reliability Modeling

2.1. Random effects degradation model

Consider a product whose performance degrades over time. We assume that the performance evolution of each unit of the product can be described by a monotone degradation process $Y(t)$, and the degradation increments $\Delta Y_j = Y(t_j) - Y(t_{j-1})$ in time interval $[t_{j-1}, t_j]$ follow an Inverse-Gaussian distribution (Ye and Chen (2014)) with probability density function (pdf):

$$f(\Delta y_j | \mu, \lambda, \Delta h_j) = \sqrt{\frac{\lambda \Delta h_j^2}{2\pi \Delta y_j^3}} \exp\left(-\frac{\lambda(\Delta y_j - \mu \Delta h_j)^2}{2\mu^2 \Delta y_j}\right), \quad (1)$$

where (μ, λ) are the mean and diffusion parameters, respectively, $\Delta h_j = h(t_j) - h(t_{j-1})$, and $h(t)$ is a monotonically increasing time transformation.

We assume that the RDT is conducted by testing n units of the product and each unit i has unit-specific parameters (μ_i, λ_i) drawn from a joint Truncated Normal-Gamma distribution (i.e., random effects) with pdf:

$$g_r(\lambda, \mu^{-1}) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} \exp(-\beta\lambda) \left(\frac{\lambda\tau}{2\pi}\right)^{1/2} \exp\left(-\frac{\lambda\tau(\mu^{-1} - \nu)^2}{2}\right) \frac{1}{1 - \Upsilon_{2\alpha}(-\nu\sqrt{\tau\alpha/\beta})}, \quad (2)$$

where $\Upsilon_{2\alpha}$ is the cumulative distribution function (cdf) of a Student-t distribution with 2α degrees of freedom. Under the random effects model, $f(\Delta y_j | \Delta h_j)$ can be derived as:

$$\begin{aligned} f(\Delta y_j | \Delta h_j) &= \int_0^\infty \int_0^\infty f(\Delta y_j | \mu, \lambda, \Delta h_j) g_r(\lambda, \mu^{-1}) d\mu d\lambda \\ &= \left[1 - \Upsilon_{2\alpha+1}\left(-(\Delta h_j + \tau\nu)\sqrt{\frac{2\alpha+1}{2\gamma(\Delta y_j + \tau)}}\right) \right] \frac{1}{1 - \Upsilon_{2\alpha}(-\nu\sqrt{\tau\alpha/\beta})} \times \\ &\quad \frac{\Delta h_j}{\sqrt{2\pi \Delta y_j^3}} \frac{\beta^\alpha}{\Gamma(\alpha)} \sqrt{\frac{\tau}{\Delta y_j + \tau}} \frac{\Gamma(\alpha + 1/2)}{\gamma^{\alpha+1/2}}, \end{aligned} \quad (3)$$

where

$$\gamma = \frac{\Delta h_j^2}{2\Delta y_j} + \beta + \frac{\tau\nu^2}{2} - \frac{(\Delta h_j + \tau\nu)^2}{2(\Delta y_j + \tau)}.$$

Let Z_{ij} be the j th degradation measurement of unit i , which is composed of the true degradation signal Y_{ij} and an error term ε_{ij} that follows the Normal distribution $N(0, \sigma^2)$ with mean 0 and known variance determined from the specifications of measurement equipment. As a result, the observed degradation increment of unit i in time interval $[t_{j-1}, t_j]$ can be expressed as: $\Delta z_{ij} = \Delta y_{ij} + \Delta \varepsilon_{ij}$, where $\Delta \varepsilon_{ij} = \varepsilon_{ij} - \varepsilon_{ij-1}$. Under this model, the

reliability of a new unit given a fixed failure threshold y^T and target lifetime T is:

$$R(T, y^T) = \int_0^{y^T} \int_{-\infty}^z f(z - \varepsilon | h(T)) \phi(\varepsilon / \sigma) / \Phi(z / \sigma) d\varepsilon dz, \quad (4)$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the pdf and cdf of the standard normal distribution, respectively. To evaluate the reliability value using this expression, Monte-Carlo (MC) integration will be performed for the inner integral, and a numerical routine will be implemented for the outer integral.

2.2. Model parameter estimation

Let $\Theta = \{\alpha, \beta, \nu, \tau\}$ be the parameter vector of the degradation process. To facilitate parameter estimation, an expectation-maximization (EM) algorithm is developed. First, consider the likelihood of the observed degradation increment Δz with complete data $f_z(\Delta z) = f(\Delta z - \Delta \varepsilon | \Delta \varepsilon, \lambda, \mu) g_r(\lambda, \mu^{-1}) \phi(\Delta \varepsilon / (\sqrt{2}\sigma))$. After testing n units with m measurements each and observing the matrix $Z = \{Z_{ij}, \forall i, j\}$ of degradation signals, the log-likelihood function for the complete information can be expressed as:

$$\begin{aligned} \ln \mathcal{L}(\Theta | Z) \propto & \sum_{i=1}^n \sum_{j=1}^m \left[\frac{\ln(\lambda_i \Delta h_{ij})}{2} - \frac{\lambda_i (\Delta y_{ij} - \mu_i \Delta h_{ij})^2}{2\mu_i^2 \Delta y_{ij}} \right] \\ & + m \sum_{i=1}^n \left[\alpha \ln(\beta) + (\alpha - 1/2) \ln(\lambda_i) - \lambda_i \beta + \frac{\ln(\tau)}{2} - \ln(\Gamma(\alpha)) - \frac{\tau \lambda_i \mu_i^{-2}}{2} + \nu \tau \lambda_i \mu_i^{-1} - \frac{\tau \nu^2 \lambda_i}{2} \right] \\ & - mn \ln \left[1 - \Upsilon_{2\alpha}(-\nu \sqrt{\tau \alpha / \beta}) \right] - \sum_{i=1}^n \frac{\Delta \varepsilon_{i1}^2}{2\sigma^2} - \sum_{i=1}^n \sum_{j=2}^m \frac{\Delta \varepsilon_{ij}^2}{4\sigma^2}, \end{aligned} \quad (5)$$

For the E-step, $\Delta \hat{y}_{ij} = E[\Delta y_{ij}]$ with $E[\Delta y_{ij}] = \Delta z_{ij} - E[\Delta \varepsilon_{ij}]$ and $E[\Delta \varepsilon_{ij}]$ is the expected value of a Truncated Normal distribution with an upper bound at $-z_{ij}$. Since the variance of measurement errors σ^2 is known, this estimate can be used for all iterations of the EM algorithm. In addition, using Bayesian results (Banerjee and Bhattacharyya 1979), we can calculate $E[\lambda_i | Z_i]$ and $E[\ln(\lambda_i) | Z_i]$ for each unit i with degradation signals $Z_i = \{Z_{ij}, \forall j\}$ as:

$$E[\lambda_i | Z_i] = \left(\alpha + \frac{m}{2} \right) \left[\beta + \frac{\tau \nu^2}{2} + \sum_{j=1}^m \frac{\Delta h_{ij}^2}{2\Delta \hat{y}_{ij}} - \frac{(\tau \nu + \sum_{j=1}^m \Delta h_{ij})^2}{2(\tau + \sum_{j=1}^m \Delta \hat{y}_{ij})} \right]^{-1} = \frac{\alpha'}{\beta'}, \quad (6)$$

$$E[\ln(\lambda_i) | Z_i] = \psi(\alpha') - \ln(\beta'), \quad (7)$$

where $\psi(\cdot) = \Gamma'(\cdot) / \Gamma(\cdot)$ is the digamma function, in which $\Gamma(\cdot)$ and $\Gamma'(\cdot)$ are the gamma function and its first derivative, respectively.

Note that $(\mu_i^{-1} | Z_i)$ follows a Truncated Student-t (Tt) distribution with lower bound 0, mean ν , standard deviation $\sigma = \sqrt{\beta / \tau \alpha}$, and 2α degrees of freedom. Moreover, the expected value and variance of a Tt distribution with lower bound of $-\nu / \sigma$, mean 0, standard deviation 1, and d degrees of freedom are (Kim 2008):

$$E[Tt | d, \Theta] = d^{d/2} \frac{\Gamma((d-1)/2)}{\Gamma(d/2)\Gamma(1/2)} \frac{(d + \nu^2 \tau \alpha / \beta)^{-(d-1)/2}}{2(1 - \Upsilon_d(-\nu \sqrt{\tau \alpha / \beta}))} \quad (8)$$

$$\text{Var}(Tt | d, \Theta) = \frac{d}{d-2} - E[Tt | d, \Theta] \left(\nu \sqrt{\tau \alpha / \beta} + E[Tt | d, \Theta] \right). \quad (9)$$

By defining $\tau' = \tau + \sum_{j=1}^m \Delta y_{ij}$, $\nu' = (\tau \nu + \sum_{j=1}^m \Delta h_{ij}) / \tau'$ and $\Theta' = \{\alpha', \beta', \nu', \tau'\}$, and using the transformation $E[\mu^{-1} | Z_i] = \nu + \sigma E[Tt]$, the expected values of μ_i^{-1} , $\lambda_i \mu_i^{-1}$, and $\lambda_i \mu_i^{-2}$ can be derived as:

$$E[\mu_i^{-1} | Z_i] = \nu' + \sqrt{\frac{\beta'}{\tau' \alpha'}} E[Tt | 2\alpha', \Theta']$$

$$E[\lambda_i \mu_i^{-1} | Z_i] = \frac{\alpha' \nu'}{\beta'} + \frac{\alpha'}{\beta'} \sqrt{\frac{\beta'}{\tau'(\alpha' + 1)}} E[Tt | 2\alpha' + 2, \Theta']$$

$$E[\lambda_i \mu_i^{-2} | Z_i] = \frac{\alpha'}{\beta'} \left[\nu'^2 + \frac{\beta'}{\tau' \alpha'} + \nu' \sqrt{\frac{\beta'}{\tau'(\alpha' + 1)}} E[Tt | 2\alpha' + 2, \Theta'] \right]$$

Note that not all the terms in Equation 5 are independent of Θ . For example, for the M-step we only need to numerically maximize:

$$\begin{aligned} \ln \mathcal{L}(\Theta|Z) \propto & (\alpha - 1/2) \sum_{i=1}^n E[\ln(\lambda_i)|Z_i] - \left(\beta + \frac{\tau v^2}{2} \right) \sum_{i=1}^n E[\lambda_i|Z_i] - \frac{\tau}{2} \sum_{i=1}^n E[\lambda_i \mu_i^{-2}|Z_i] + v\tau \sum_{i=1}^n E[\lambda_i \mu_i^{-1}|Z_i] \\ & + n \left[\alpha \ln(\beta) - \ln(\Gamma(\alpha)) + \frac{\ln(\tau)}{2} - \ln \left[1 - \Upsilon_{2\alpha}(-v\sqrt{\tau\alpha/\beta}) \right] \right] \end{aligned} \quad (10)$$

Alternatively if $P\{\mu^{-1} < 0\} \approx 0$, it is possible to approximate the maximum likelihood estimates by solving the following system of equations:

$$\begin{aligned} \hat{\beta} &= \frac{\alpha}{\sum_{i=1}^n E[\lambda_i|Z_i]/n} & \ln(\hat{\alpha}) - \ln\left(\sum_{i=1}^n E[\lambda_i|Z_i]/n\right) + \sum_{i=1}^n E[\ln(\lambda_i)|Z_i]/n - \psi(\hat{\alpha}) &= 0 \\ \hat{v} &= \frac{\sum_{i=1}^n E[\lambda_i \mu_i^{-1}|Z_i]}{\sum_{i=1}^n E[\lambda_i|Z_i]} & \hat{\tau} &= n \left[\sum_{i=1}^n E[\lambda_i \mu_i^{-2}|Z_i] - 2\hat{v} \sum_{i=1}^n E[\lambda_i \mu_i^{-1}|Z_i] + \hat{v}^2 \sum_{i=1}^n E[\lambda_i|Z_i] \right]^{-1} \end{aligned}$$

It is worth pointing out that the corresponding statistical inference methods developed in this paper can be used for other stochastic degradation process models, such as the Wiener process and gamma process.

3. Optimal Design of Degradation-based RDT

Consider a RDT experiment where n units will be tested for a fixed amount of time τ , and m degradation measurements will be taken on each unit periodically at equal time intervals such that the time between measurements is τ/m . Given the cost of a unit c_u that includes testing and unit manufacturing cost, and the measurement cost c_m , the decision maker wants to minimize the total cost of testing subject to performance guarantees in terms of the producer's risk r_p and the consumer's risk r_c by determining the optimal values of n and m . The optimal RDT design problem can be formulated as:

$$\begin{aligned} & \underset{n,m}{\text{minimize}} && c_u n + c_m n m \\ & \text{subject to} && P\{R(\hat{\Theta}) \leq R_0 | R(\Theta_p) > R_0, \Omega\} \leq r_p && \text{Producer's risk} \\ & && P\{R(\hat{\Theta}) \geq R_0 | R(\Theta_c) < R_0, \Omega\} \leq r_c && \text{Consumer's risk} \\ & && Lb \leq (n, m) \leq Ub && \text{Lower and upper bounds for } (n, m) \end{aligned}$$

where $\Omega = \{n, m\}$ represents the experimental design for a bounded design space, and the competing parameter sets Θ_p and Θ_c are specified by the producer and consumer, respectively.

We only discuss how to determine the feasibility of the producer's risk with techniques that are also applicable for the consumer risk's. The exact probability involves nm dimensional integral of the form:

$$P\{R(\hat{\Theta}) \leq R_0 | R(\Theta_1) > R_0, \Omega\} = \int_0^\infty \cdots \int_0^\infty I(R(\hat{\Theta}|Z) \leq R_0) f_Z(Z|\Theta_1, \Omega) dz_{11} \cdots dz_{nm}. \quad (11)$$

where $I(R(\hat{\Theta}|Z) \leq R_0)$ is an indicator function of the test being rejected for the null hypothesis: $R \geq R_0$ and the alternative hypothesis: $R < R_0$. Note that an EM algorithm for parameter estimation has to be used to compute $\hat{\Theta}$ for each design candidate. Moreover, since the sample size n tends to be small when minimizing the total cost of testing, it is not appropriate to use large-sample approximation for calculating the variance-covariance matrix of $\hat{\Theta}$, which only leaves a bootstrap procedure as a reasonable way for performing the hypothesis test. All these complications make the numerical integration of this probability computationally prohibitive.

Homem-de-Mello and Bayraksan (2015) solve a chance constrained stochastic optimization problem by constructing constraints of the form $P \geq \alpha$, where $\hat{P} = \sum_{k=1}^N I(\cdot)/N$ is estimated as the average of an indicator function and α is the minimum required probability. By constructing a test with $\rho < \alpha$ such that if it passes there is a $(1 - \beta)$ confidence that the constraint is satisfied. The quantity ρ is estimated as $\sup \rho \in (0, 1)$ such that $B(N\hat{P}, N, \rho) \leq \beta$ for $B(\cdot)$ being the cdf of the binomial distribution with the probability of success ρ and $N\hat{P}$ successes out of N trials. Using this method, we need to transform the producer's risk constraint to $1 - P\{R(\hat{\Theta}) \leq R_0 | R(\Theta_p) > R_0, E\} \geq 1 - r_p$, then specify β and N to do a Monte Carlo simulation of N tests for a given Θ_p .

4. Numerical Example

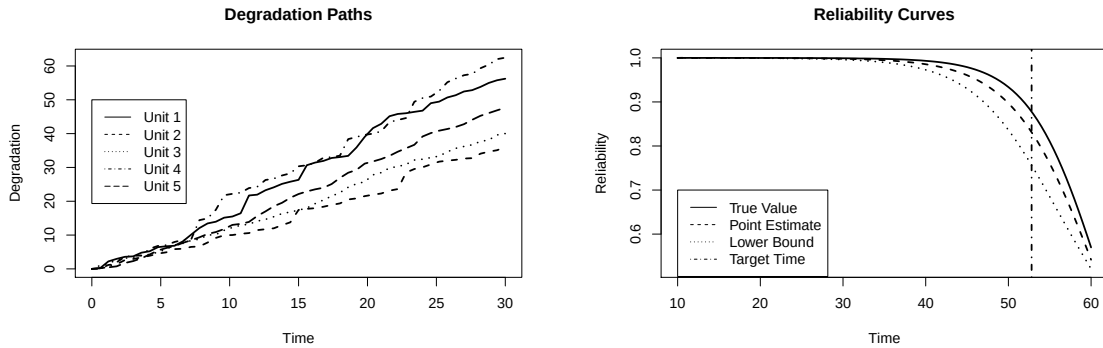
This section provides a numerical example to illustrate the parameter estimation for the proposed random effects degradation model and the design of optimal degradation-based RDT plan.

4.1. Problem setting

Assume a manufacturer wants to sell a new product to a consumer. The minimum reliability requirement for this product is $R_0 = 0.8779$ after approximately $T = 53$ days of operation. The degradation parameters that achieve this target are $\Theta = \{4, 2, 1, 50\}$. In the degradation test, a time transformation $h(t) = t^{1.12}$ is considered (so the target life time after the transformation is $h(T) = 85$ days), and measurement errors have mean 0 and variance $\sigma^2 = 0.001$. A test unit is considered failed if its degradation process crosses a failure threshold of 100 (i.e., soft failure). The RDT is conducted for a fixed 30 days.

4.2. Parameter estimation

To illustrate the proposed reliability estimation method for the random effects degradation model, we consider an RDT plan that takes 50 measurements on each of 50 units with model parameters drawn from Θ . Figure 1a shows the simulated degradation paths of 5 (out of 50) units. We implement the EM algorithm presented in Section 2.2 and use the function “auglag” from the package “nloptr” in R to solve the M-Step routine with initial solution of $\Theta_0 = \{1, 1, 1, 1\}$. The hypothesis test on the product’s reliability is conducted using a bootstrap procedure with 200 nonparametric bootstrap samples. A one-sided 95% confidence interval on the reliability is constructed using the pivot inversion procedure, and the acceptance is declared if the attained reliability lower bound is greater than R_0 . The parameter estimates from the simulated degradation data are $\hat{\Theta} = \{4.369, 2.416, 0.9932, 35.56\}$, the point reliability estimate is 0.8312, and the bootstrap lower bound of reliability is 0.7639. Figure 1b shows the true reliability of the product, the point estimate and the lower bound. The result is unfavorable to the producer because of the under-estimation of the product’s reliability.



(a) Simulated degradation paths (5 out of 50 units).

(b) True and estimated reliability functions.

Figure1. Simulated degradations paths and reliability functions.

4.3. Optimal RDT plan

Assume that the cost per unit c_u is 100, the cost of each measurement c_m is 2, and both the number of measurements and units to be tested have a lower bound of 20 and an upper bound of 100, respectively. According to the producer, the product degrades with parameters $\Theta_p = \{4, 2, 1.1, 50\}$, so that its reliability at the target 53 days is 0.9693. On the other hand, the consumer proposes the parameters $\Theta_c = \{4, 2, 0.9, 50\}$ with reliability of 0.6623. Without loss of generality, we consider that both the consumer’s and producer’s risks are 5%. The feasibility of each candidate test plan is evaluated using the method outlined in Section 3 with a 95% confidence on the feasibility ($\beta = 0.05$) and 100 independent simulations of the test plan. Table 1 presents the feasibility and costs of candidate test plans. ρ_c and ρ_p are the values obtained for the consumer and producer respectively, and a test plan is feasible only if both values are greater than 0.95. The optimal test plan requires the use of 50 units and taking 50 measurements on each test unit. The total cost is 10000.

Table 1: Costs and feasibility of candidate testing plans.

m	n	ρ_c	ρ_p	Cost	m	n	ρ_c	ρ_p	Cost	m	n	ρ_c	ρ_p	Cost
20	20	1	0.8803	2800	48	48	1	0.9452	9408	60	60	1	0.9801	13200
40	40	1	0.9133	7200	49	49	1	0.9452	9702	100	100	1	0.9735	30000
45	45	1	0.9463	8550	50	50	1	0.9769	10000	—	—	—	—	—

5. Conclusions

In this paper, a new methodology is proposed for the use of degradation data in modeling and planning RDT. The EM algorithm developed in this study can effectively estimate the parameters of the proposed random effects degradation model considering measurement errors. The result shows that the proposed degradation-based RDT can significantly speed up reliability demonstration. In addition, the proposed simulation methodology enables the feasibility assessment of an RDT plan involving both the consumer's and producer's risks.

It is worth pointing out that the numerical results suggest that the most important restriction in planning the RDT is the producer's risk. Sometimes, the consumer's risk may be dropped from the formulation without affecting the feasible region. In practice, the methodology proposed in this paper can be adapted to a Bayesian model in order to explore if the bias in the reliability estimation can be reduced.

References

- Banerjee, A. K and Bhattacharyya, G. K (1979) Bayesian results for the Inverse Gaussian distribution with an application. *Technometrics* **21**(2), 247–251.
- Guo, H and Liao, H (2012) Methods of reliability demonstration testing and their relationships. *IEEE Transactions on Reliability* **61**(1), 231–237.
- Guo, H, Liao, H, Gerokostopoulos, A and Mettas, A (2011) Design of reliability demonstration testing for repairable systems. *Proceedings of 2011 Annual Reliability and Maintainability Symposium* Lake Buena Vista, FL.
- Homem-de-Mello, T and Bayraksan, G (2015) Stochastic constraints and variance reduction techniques. in M. C. Fu (ed.), *Handbook of Simulation Optimization* 1 edn Vol. 216 Springer-Verlag New York chapter 9, pp. 245–276.
- Kim, H.-J (2008) Moments of truncated Student-t distribution. *Journal of the Korean Statistical Society* **37**(1), 81–87.
- Lawless, J. F (1978) Confidence interval estimation for the weibull and extreme value distributions. *Technometrics* **20**(4), 355–364.
- Leemis, L. M (2006) Lower system reliability bounds from binary failure data using bootstrapping. *Journal of Quality Technology* **38**(1), 2–13.
- Liao, H and Elsayed, E. A (2006) Reliability inference for field conditions from accelerated degradation testing. *Naval Research Logistics* **53**(6), 576–587.
- McKane, S. W, Escobar, L. A and Meeker, W. Q (2005) Sample size and number of failure requirements for demonstration tests with log-location-scale distributions and failure censoring. *Technometrics* **47**(4), 182–190.
- Meeker, W. Q and Escobar, L. A (1998) *Statistical Methods for Reliability Data* John Wiley & Sons New York, New York.
- Ye, Z.-S and Chen, N (2014) The inverse gaussian process as a degradation model. *Technometrics* **56**(3), 302–311.

Insight into cost of defective state using delay-time concept

A. C. J. Santos* and C. A. V. Cavalcante**

* Departamento de Engenharia de Produção, Universidade Federal de Pernambuco, Brazil:

augustocesardejesussantos@gmail.com

** Departamento de Engenharia de Produção, Universidade Federal de Pernambuco, Brazil: cristiano@ufpe.br

Abstract. Delay-time concept has been used in a large number of applications for modelling important aspects of reality, such as the dynamics of failures, which in practice occurs in two stages for the most devices. Concisely, a defect arises in a component or system followed by a subsequent failure after an interval called delay time. The effort to repair a defect is related to the time from when it can be firstly observed until its discovery by an inspection. However, a small number of contributions deal with the cost related to the time in which the system undergoes in the defective state. In addition, a repair action on a system that undergoes in defective state for a short period of time can have different consequences than the same action after a long period of time. Two examples are: a pipeline leakage defect and the repair of a defect in a tooth. In both cases, the consequences vary depending on the time in defective state. Hence, this paper proposes a way to take the defect cost into consideration according to the time in which the system undergoes in this state.

1. Introduction

Modelling a maintenance policy is a way to understand the behavior of a system, equipment or component and to try to establish a clarification or a solution for an aspect that is crucial for the system. Nevertheless, the maintenance models had a late development when compared to models related to operations. Scarf (1997) suggests that the reasons for a late development on maintenance modelling are related to the fact that maintenance actions are carried out on plant, not on product, and also because it is difficult to associate the spending on maintenance and its respective results in production performance.

Currently, maintenance is no longer considered a necessary evil and represents an important function for competitiveness (Pintelon and Parodi-Herz, 2008). In fact, the way that maintenance is organized and structured can highly influence the performance and competitiveness of an organization. Pinjala, Pintelon and Vereecke (2006) indicate that maintenance has now a more important role to achieve cost-effective operations and superior product quality. Due to changes in maintenance perspective along the years and its greater importance in the current industrial scenario, traditional maintenance policies have been continuously studied and applied in order to accomplish better results in terms of cost, availability and other aspects.

Maintenance models have been adjusted to better fit some particular situations. A positive consequence is the creation of new models that have been extensively applied to similar maintenance conditions. As a result, interesting modifications in maintenance models have been suggested to take into consideration specific points that were not addressed on previous models.

This paper proposes to analyse the cost of a defect while a component undergoes in defective state. The model analysed is: delay time model for single component (Christer, 1976, 1982; Christer and Waller, 1984; Wang, 2008, 2012). The basis for the modification is supported by the fact that a penalty cost should be associated to a defect when it implies in a reduction in the performance of a system or the sojourn time of the defect increases the difficulty of the maintenance action. Besides, the modification proposal is in line with practice in many situations and it is not difficult to be applied to real cases, which are important aspects in a maintenance model (Scarf, 1997). In fact, models with a higher complexity and number of parameters are more difficult to be applied to real cases.

The outline of the paper is as follows: the model is reviewed in section 2, the adjustments are suggested in section 3, a model analysis is presented in section 4, possible applications for revised model are presented in section 5, some consideration about application on multi-component system is presented in section 6 and conclusions and future work are discussed in section 7.

2. Reviewing the model

In this section, the delay-time model for single component is reviewed as it was first proposed and the theory related to it is briefly explained. A single component system may be interpreted as a component or a component and a socket that together execute an operational function (Ascher and Feingold, 1984; Scarf and Cavalcante, 2012).

The delay time model was first introduced by Christer (1976) and it considers that the failure is regarded as a two-stage process. A defect arises in the component at some time x , followed by a subsequent failure after an interval h (Baker and Christer, 1994), as shown in Figure 1.

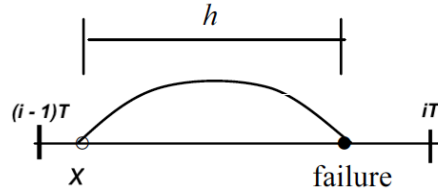


Figure 1. Delay time for a defect (Adapted from Wang, 2008: p. 347).

Clearly, the failure is preceded by a defect that undergoes during the interval h . This interval represents an opportunity to check the system and correct the defect before it turns into a failure. In this sense, the interaction between equipment performance and inspections can be evaluated using the delay-time concept (Wang, 2012). In fact, it is possible to determine an inspection interval that minimizes the total expected cost over an inspection cycle, $C(T)$, which includes: inspection cost (C_i), average cost to repair a defect (C_p) and failure cost (C_f).

First, we assume that the following assumptions are valid (Lee, 1999):

- 1) An inspection takes place every T time units, cost C_i units and requires d_i time units, where $d_i \ll T$.
- 2) Inspections are perfect and the defect present will be identified.
- 3) Defect identified at an inspection will be repaired within the inspection period at an average cost of C_p .
- 4) Failure is repaired as soon as it arises, incurring on average d_f downtime and C_f cost, where $C_f > C_p$ and $d_f \ll T$.
- 5) The initial time x of a defect is uniformly distributed over $(0, T)$, and is independent of h , with defect arises at a rate of k per unit time.
- 6) The probability density function (*pdf*) of delay time h , $f(h)$, is known.

Considering Figure 1, $C(T)$ can be determined considering two possibilities: failure occurs on the interval $(x + h \in [(i-1)T, iT])$ or failure does not occur on the interval $(x \in [(i-1)T, iT])$ and $h > iT - x$. A possible way to model these possibilities, according with the notation used in Figure 1 and considering the assumptions above, is presented in Equation 1:

$$C(T) = \sum_{i=1}^{\infty} [iC_i + C_p] \int_{(i-1)T}^{iT} f(x) R_h(iT - x) dx + \sum_{i=1}^{\infty} [(i-1)C_i + C_f] \int_{(i-1)T}^{iT} f(x) F_h(iT - x) dx \quad (1)$$

Similarly, the lifetime of a component, $L(T)$, can be modelled as suggested in Equation 2.

$$L(T) = \sum_{i=1}^{\infty} iT \int_{(i-1)T}^{iT} f(x) R_h(iT - x) dx + \sum_{i=1}^{\infty} \int_{(i-1)T}^{iT} f(x) \left[\int_0^{iT-x} (x+h) f(h) dh \right] dx \quad (2)$$

The first part of Equations 1 and 2 represents the situation when the failure does not occur on the interval and the second part represents the case in which the failure occurs on the interval. The sum from 1 to infinite number of inspections considers all necessary inspections due to the many defects and failures that can arrive in the failure process of a single component system along time.

The sequence of inspections should be determined in a way that the total expected cost over an inspection cycle is minimized. In many practical situations, the failure cost is much higher than the inspection cost and

the cost to repair a defect item, because it is influenced by some inconvenient brought to the system, such as the downtime period (d_f) from failure to repair action. Consequently, the inspection intervals tend to be determined to avoid failures and their negative influences on system.

This initial model does not consider that the time in defective state can have a penalty cost due to the insertion of possible limitations in performance and productivity. The inclusion of this possibility is commented in the next section.

3. Insight into cost of defective state

In this section, the insight into the cost of defective state is provided. This cost can be modelled as a part of the cost to repair a defect and as a part of the failure cost. As the failure cost is sufficiently high in most cases, the cost of defective state is modelled in this paper only as a part of the cost to repair a defect.

As shown in section 2, the delay time modelling for a single component considers a two-stage process and the failure can occur on the inspection interval ($x + h \in [(i-1)T, iT]$) or out of the inspection interval ($x \in [(i-1)T, iT]$ and $h > iT - x$). For both cases, the cost associated with the time in defective state is not considered. If the defect is found, the repair cost is the same (C_p) no matter the sojourn time of the defect. However, there are situations in which the consequences of a period in this state should be measured in order to better represent the reality, for instance, a leakage defect. In this case, the consequences vary depending on the time in defective state and the more time in this state the more waste of material and penalty cost.

Aven and Castro (2009) propose a discounted cost based on the number of failures of the system and the time spent in the defective state (delay time). In our paper, however, the defect cost per unit time is suggested to be considered in delay time model as an increment in the cost to repair a defect at an inspection.

In this sense, when the failure does not occur on the interval, the defect cost per unit of time can be considered making the cost to repair a defect at an inspection varies as a function of the time that the system undergoes in defective state ($C_p(y) = C_p + C_c(y)$, with $y = iT - x$). Hence, the first part of Equation 1 can be reformulated as follows in Equation 3. Possible formats for $C_c(y)$ can be linear, quadratic, exponential or other according to the specific problem. The next section considers two of them for a model analysis.

$$C(T)_1 = \sum_{i=1}^{\infty} iC_i \int_{(i-1)T}^{iT} f(x)[C_p + C_c(y)]R_h(iT - x)dx \quad (3)$$

When failure occurs on the interval, the defect cost per unit time can be considered making the failure cost varies as function of the time in which the system undergoes in defective state as well. In this paper, we didn't consider this variation. However, we present a way that it would be considered. The failure cost could be modelled as follows: $C_f(h)$ is the sum of the fixed penalty cost relative to the general consequences of the failure, C_{ff} , and the time-related penalty cost in defective state, h , which interferes in variable failure cost, $C_c(h)$. So, $C_f(h) = C_{ff} + C_c(h)$ and the second part of Equation 2 can be reformulated as follows in Equation 4.

$$C(T)_2 = \sum_{i=1}^{\infty} [(i-1)C_i] \int_{(i-1)T}^{iT} f(x) \int_0^{iT-x} f(h)C_f(h)dhdx \quad (4)$$

This modification proposal in delay time model has two important characteristics: (1) it can describe an important and recurrent practice situation, which can be associated with a significant number of real cases; (2) it does not increase the complexity of the model, so that is applicable to reality.

4. Model analysis

This section evaluates the delay time model for a hypothetical problem. It consists in assessing the optimal interval between inspections and the respective optimal cost rate per unit of time for previous and revised models. This analysis considers a linear function ($C_c(y) = cy$) and a quadratic function ($C_c(y) = cy^2$) for representing the cost along the time in defective state.

We consider that $f(x)$ follows a Weibull distribution with shape parameter equals to 3 and scale parameter equals to 1, $f(h) = 2e^{-2h}$, $C_i = \$0.5/\text{inspection}$, $C_p = \$1/\text{repair of a defect}$, $C_f = \$5/\text{failure}$ and c as defined in table 2 (\$/unit time in defective state).

Table 2: Comparing the two models.

	No cost in defective state	Linear cost in defective state $C_c(y) = cy$				Quadratic cost in defective state $C_c(y) = cy^2$			
c	not applicable	1	2	3	4	1	2	3	4
Insp. period (T)	1.191	1.186	1.181	1.177	1.172	1.183	1.176	1.170	1.164
$C(T)$	3,588	3.732	3.877	4.021	4.165	3.665	3.742	3.819	3.896
$L(T)$	1.178	1.178	1.177	1.177	1.176	1.178	1.177	1.176	1.176
$E(T) = C(T)/L(T)$	3.046	3.169	3.293	3.417	3.541	3.113	3.180	3.247	3.314

Table 2 shows that the cost in the defective state directly affects the optimal interval between inspections and the minimal $E(T)$. When the parameter c rises, the interval between inspections is lightly reduced and the $E(T)$ is significantly higher, as expected. The increase in the expected cost is illustrated in Figure 2 for linear case and in Figure 3 for quadratic case.

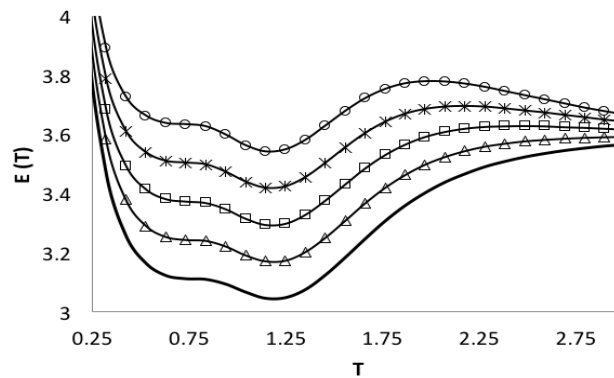


Figure 2. Expected cost per unit time for the defective cost as a linear function and different values of c : $c = 0$ (—); $c = 1$ (Δ); $c = 2$ ($\bullet$); $c = 3$ ($*$); $c = 4$ ($\circ$).

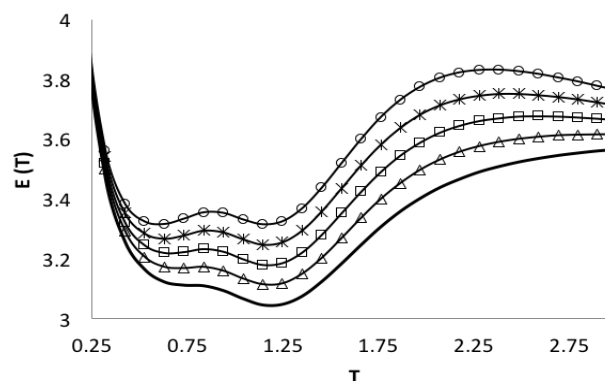


Figure 3. Expected cost per unit time for the defective cost as a quadratic function and different values of c : $c = 0$ (—); $c = 1$ (Δ); $c = 2$ ($\bullet$); $c = 3$ ($*$); $c = 4$ ($\circ$).

Figures 2 and 3 represent the expected cost behavior along the time. As observed, the higher the cost in defective state, the higher is the increase in expected cost. Although the period between inspections does not suffer a significant change for this particular example, the variation in the expected cost may affect the way that maintenance is planned. It gives a more representative view from reality and can offer a more reliable cost prediction for the multiples possible situations related to the maintenance of the system.

Besides, it is important to emphasize that in real cases the defective cost function should be carefully modelled to fit the characteristics of the specific problem and the parameters used in the model should be precisely defined. Considering these aspects, the delay time modelling can be better applied. The next section suggests possible real applications in which the revised model can be applied.

5. Possible applications for revised model

This section presents two practical situations in which the revised model may be applied: leakage defect in pipelines and dental defects. In both cases, the period in defective state considerably affects the features of the system and entails negative consequences.

In the former case, a defect or failure causes high losses of product, energy, capital and can even cause human and environmental damages. Consequently, the inspection on pipelines over regular periods of time is critical (Sorabh, Raghav and Sengar, 2017). The proposal modification presented in Section 3 may be an alternative to model the inspection interval to this specific context.

In the latter case, a dental defect is influenced by the time that the tooth undergoes in defective state, so the longer the period of time in this state, the higher the negative consequences associated with it, such as, decrease in dental health and increase in repair costs. In the context of defective dental restorations, for instance, the repair brings advantages as: less loss and more preservation of tooth structure, reduction of pain, reduction of treatment time, reduced costs to the patient, and increased longevity of the restoration (Hickel, Brühshaver and Ilie, 2013). In this sense, the inclusion of variable defect cost over time in delay time modelling may serve to model, for example, the increase in cost associated with a dental defect over time.

Finally, other practical situations may be characterized and modelled according to the proposals described in this paper. They may be applied in situations in which the time in defective state brings negative impacts that may be accounted in terms of cost, for instance, low performance, environmental consequences, changes in the operating characteristics of the equipment, other penalties, so on.

The next section presents some considerations for possible future work on multi-components systems in order to meet the requirements of particular cases with the same characteristics of a penalty cost associated with the cost in defective state.

6. Extending context and application for multi-component systems

As in the single-component system, there are practical situations that require a multi-component approach. One interesting example is the hammers of a shredding machine. The hammers are components that tend to wear out considerably before fail because of its performance features (Zhou et al, 2016). This wear can decrease efficiency on machine performance and, therefore, a significant portion of the maintenance cost is function of the defective state of each hammer in the system.

A model that would provide a way to represent this situation is presented by Anbar (1976) and suggests that the similar components works in parallel in a system with identical exponential failure distribution. The system works if and only if k out of n units are working. The failures can be discovered only by inspections and when an inspection occurs all failure units are replaced by new ones.

This model would provide a good representation for this particular situation if two important modifications were done. The first one is to insert the possibility of a defect to occur and the second one is to provide a way to account for the cost in defective state. An analysis on delay time modelling for multi-component system may assist in this situation.

7. Conclusion and future work

This paper contributes to the development of maintenance modelling by taking into consideration the cost related to the period in defective state. The proposal modification in delay time model for single component system allows some practical situations to be better modelled. In these cases, the sojourn time in defective state negatively influences the performance of a system. Some examples of real context were presented and other possible related cases can be identified based on their characteristics. Besides, there is not a significantly increase in model complexity, so that the proposal is not difficult to be applied to real cases. In

future work, the revised model may be applied to the practical situations illustrated in Section 6 and the findings compared with the ones obtained by delay time model. Also, a multi-component system analysis is expected to be developed, as suggested in Section 6.

Acknowledgements

The authors are grateful for the support provided by CAPES and CNPq.

References

- Anbar, D. (1976) An Asymptotically Optical Inspection Policy. *Naval Research Logistics Quarterly*, 23, 211-218.
- Ascher, H., Feingold, H. (1984) Repairable Systems Reliability. *MarcelDekker*, New York.
- Aven, A., Castro, I. T. (2009) A delay-time model with safety constraint. *Reliability Engineering and System Safety*, 94, 261-267.
- Baker, R. D., Christer, A. H. (1994) Reviewing delay time-model OR modelling of engineering aspects of maintenance. *European Journal of Operational Research*, 73, 407-422.
- Christer, A. H. (1976) Innovatory decision making. In *Proc. NATO conference on role and effectiveness of decision theory in practice* (Bowen, K. and White, D. J., Eds.)
- Christer, A. H. (1982) Modelling inspection policies for building maintenance. *Journal of Operational Research Society*, 33 (8), 723-732.
- Christer, A. H., Waller, W. M. (1984) Delay time models of industrial inspection maintenance problems. *Journal of Operational Research Society*, 35 (5), 401-406.
- Hickel, R., Brühshaver, K., Ilie, N. (2013) Repair of restorations – criteria for decision making and clinical recommendations. *Dental materials*, 29, 28-50.
- Lee, C. (1999) ‘Applications of delay time theory to maintenance practice of complex plant’, PhD thesis, University of Salford, viewed 7 april 2018, <<http://usir.salford.ac.uk/2161/1/299125.pdf>>.
- Pinjala, S. K., Pintelon, L., Vereecke, A (2006) An empirical investigation on the relationship between business and maintenance strategies. *International Journal of Production Economics*, 104, 214-229.
- Pintelon, L., Parodi-Herz, A. (2008) Maintenance: An Evolutionary Perspective. In: Kobbacy, K. A. H., Murthy, D. N. P. (eds.) *Complex system maintenance handbook*, Springer, 21-48.
- Scarf, P.A. (1997) On the application on mathematical models in maintenance. *European Journal of Operational Research*, 99, 493-506.
- Scarf, P. A., Cavalcante, C. A. V. (2012) Modelling quality in replacement and inspection maintenance. *Int. J. Productions Economics*, 135, 372-381.
- Sorabh, Raghav, V., Sengar, S. (2017) Finite element analysis of rectangular defects in inspection of steel pipelines using magnetic flux leakage technique. *Materials Today: Proceedings*, 4, 2665-2672.
- Wang, W. (2008) Delay time modelling. In: Kobbacy, K. A. H., Murthy, D. N. P. (eds.) *Complex system maintenance handbook*, Springer, 345-370.
- Wang, W. (2012) An overview of the recent advances in delay-time-based maintenance modelling. *Reliability engineering and system safety*, 106, 165-178.
- Zhou, X., Hu, Z., Qin, X., Hua, L. (2016) Study on the stress characteristic and fatigue life of the shredder pin. *Engineering Failure Analysis*, 59, 444-455.

A two-dimensional availability-based warranty policy incorporating preventive maintenance

Chun Su* and Longfei Cheng**

School of Mechanical Engineering, Southeast University, Nanjing 211189, China

* *suchun@seu.edu.cn* ** *longfei_cheng@seu.edu.cn*

Abstract. A two-dimensional (2D) availability-based warranty policy is investigated from the view of the manufacturer. Under the policy, the manufacturer guarantees a negotiated availability for the products during the warranty period. In order to satisfy the availability requirement with rational cost, 2D preventive maintenance (PM) strategy is carried out, which is scheduled every K units of age or L units of usage. This study is to meet products' availability requirement with the lowest cost by optimizing 2D PM strategy. A numerical example is provided to verify the effectiveness of the proposed policy. The results show that compared with age-based or usage-based PM strategy, 2D PM strategy is superior in terms of warranty cost and operational availability.

1. Introduction

Warranty is a type of guarantee that a manufacturer makes to deal with the item which fails to fulfill its intend function throughout the warranty period (Yeh and Fang, 2015). The most common warranty policies include free repair or replacement warranty (FRW), pro-rata warranty (PRW) and renewing free replacement warranty (RFRW) (Darghouth, et al. 2017). However, for complex industrial equipment, such as wind turbine and dump truck, the above policies may not satisfy actual needs in operation and maintenance. The reason is that such devices are usually key elements in the corresponding systems and they may suffer from substantial loss or service interruption due to failure or shutdown. Hence, the availability level of such equipment should be guaranteed from the perspective of the customer, and rational warranty policies should focus on their availability requirement.

This study is to study an availability-based warranty policy for industrial equipment. Under this policy, the manufacturer should not only provide free repair or replacement upon failures, but also guarantee the availability of the warranted equipment to be over a negotiated level during the warranty period. The existing research on availability-based warranty is quite limited (Cheng, et al. 2015; Yan, et al. 2011; Iskandar, et al. 2014). Su and Wang (2016) investigated an availability-based warranty policy in the framework of one-dimensional (1D) warranty. Compared with 1D warranty, 2D warranty can not only protect manufacturers from additional damage caused by customers' heavy usage, but also provide customers with flexible options (Wang and Su, 2016). Therefore, this study investigates an availability-based warranty policy based on a 2D framework, which considers both the age and usage of the warranted products.

In the availability-based warranty policy, one of the most fundamental issues is to guarantee the warranted equipment with a negotiated minimal level of availability and at a reasonable cost. Thus, appropriate PM action is a wise choice, which can effectively reduce the possibility of failure and downtime (Iskandar, et al. 2014).

In most existing references, PM activities are usually performed by considering only one single factor, e.g., age or usage. In this paper, 2D PM strategy is considered, that is, PM actions are scheduled every K units of age or L units of usage, whichever occurs first (Wang and Su, 2016). Compared with age-based or usage-based PM strategy, 2D PM strategy is superior, or at least identical in terms of warranty cost (Su and Wang, 2016). As the major contribution, this study investigates the availability-based warranty policy in the framework of 2D warranty. Meanwhile, imperfect 2D PM strategy is carried out to achieve the prospective availability goal with the lowest cost from manufacturers' perspective.

The remaining of this paper is organized as follows. In Section 2 the mathematical models are established, including product failure, imperfect PM, warranty servicing cost and availability models. The numerical example is provided to illustrate the proposed warranty policy in Section 3. Finally, Section 4 concludes this paper and suggests some directions for future research.

2. Model Formulation

In terms of an item sold with a 2D availability-based warranty policy, the manufacture provides customers with free repair service in a rectangular region as shown in Figure 1 (Note that, this illustration ignores PM intervals). At the same time, the availability level predetermined by both sides of the transaction is guaranteed over the warranty period. The boundary of the warranty implemented by the manufacturer is age limit W or usage limit U , whichever occurs first. In addition, the manufacturer adopts a 2D PM strategy free of charge so as to satisfy the availability requirement with a lower cost, that is, the item is preventively maintained every K units of age or L units of usage, which occurs first.

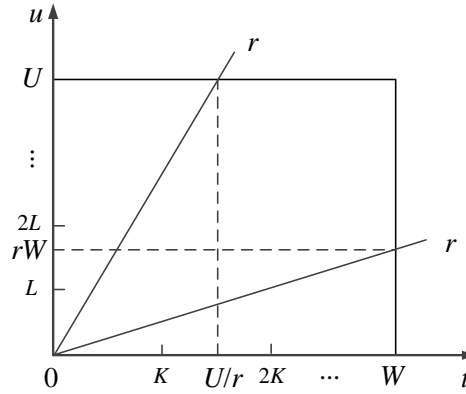


Figure 1. Two-dimensional warranty policy with PM.

In this study, it is supposed that the product is repairable and will deteriorate with both age and usage. In addition, during the warranty period, if a failure happen to the product, it will be minimally repaired so as to restore the product to working status just as that state before it failed.

Up to now, there are three kinds of methods proposed to model the product failure process under 2D warranties, including univariate method, bivariate method and composite scale method (Wu, 2014). Among them, the univariate method treats the usage as a function of the time so that 2D warranties modeling can be solved from a 1D perspective. This method is adopted to model the failures for items due to its simple and wide use.

The operating time and usage of an item can be denoted by $T(t)$ and $U(t)$ at age t . When no failure occurs before t or all failures are repaired with negligible repair durations, $T(t) = t$ (Iskandar and Murthy, 2003). There is a fundamental assumption in the univariate method, that is, the usage rate varies randomly across the customer population but is constant over time for a specific customer. Based the above assumption, the usage rate $R = U(t)/T(t)$ can be expressed as a nonnegative variable with distribution function $G(r) = P(R \leq r)$, $r_{\min} \leq r \leq r_{\max}$, where $r_{\min}$ and $r_{\max}$ are the minimum and maximum values of r , respectively. Note that, the distribution $G(r)$ can be obtained from the past history of similar items or customer reports. Thus, product failures can be modeled by a stochastic process with an intensity function dependent on age and usage. Given a usage rate $R = r$, the conditional intensity function can be expressed as:

$$\lambda(t | r) = \varphi(T(t), U(t)) \quad (1)$$

where $\varphi(\bullet)$ is a non-decreasing function of $T(t)$ and $U(t)$.

Furthermore, the conditional intensity function in this work adopts the simplest first order polynomial form proposed by Iskandar and Murthy (2003), and is given by:

$$\lambda(t | r) = \theta_0 + \theta_1 r + \theta_2 T(t) + \theta_3 U(t) \quad (2)$$

where $\theta_0, \theta_1, \theta_2$ and θ_3 are positive constants.

Conditional on $R = r$, the usage $U(t)$ of an item equals rt at age t .

$$\lambda(t | r) = \theta_0 + \theta_1 r + \theta_2 t + \theta_3 rt \quad (3)$$

In this study, virtual age model proposed by Kijima (1989) is used to model imperfect PM effect, which means that the virtual age of the system is reduced to a certain value after PM actions. Assuming that PM program is carried out at actual age $\tau_1, \tau_2, \dots, \tau_j, \dots$, with $\tau_0 = 0$. It's worth mentioning that the j th PM action only reduces the supplement of system age since the $(j-1)$ th PM. Hence, the virtual age v_j right after the j th PM can be expressed as (Kim, et al. 2004):

$$v_j = v_{j-1} + \delta(m)(\tau_j - \tau_{j-1}) \quad (4)$$

where m is PM effort level with $m = 0, 1, 2, \dots, M$. Note that, $m = 0$ implies that the effect of PM is as bad as old (ABAO), $m = M$ implies that the effect of PM is as good as new (AGAN), and PM is imperfect when $0 < m < M$. $\delta(m)$ is a decreasing function of m with $\delta(0) = 1$ and $\delta(M) = 0$. Here, $\delta(m)$ is modeled as an exponentially function of m , i.e., $\delta(m) = (1+m)e^{-m}$.

According to Huang and Yen (2009), Equation (4) can be recursively derived as:

$$v_j = v_0 + \delta(m)(\tau_j - \tau_0) \quad (5)$$

When $\tau_0 = 0$ and $v_0 = 0$, then

$$v_j = \delta(m)\tau_j \quad (6)$$

Note that, we assume that the level of PM effort m is constant over the warranty period (Huang and Yeh, 2009). In addition, it is supposed that the j th PM activity has the same impact on virtual usage due to the reduction of virtual age under 2D PM strategy. The above assumption is reasonable since the linear relationship $U(t) = rt$ is used in the univariate method (Wang and Su, 2016).

Let $EC(\Omega)$, $EA(\Omega)$ denote the expected warranty servicing cost and operational availability over the warranty coverage respectively, where the decision space Ω represents the parameters of 2D PM and is characterized by $\Omega = (K, L, m)$. Among them, $EC(\Omega)$ consists of the minimal repair cost and the PM cost from manufacturers' perspective and $EA(\Omega)$ is defined by mean uptime (MUT) and mean downtime (MDT) during the warranty period, that is, $EA(\Omega) = \text{MUT}/(\text{MUT} + \text{MDT})$. In addition, faulty items are minimally repaired throughout the warranty period with cost C_f and duration T_f . Each PM action with m level is implemented at a cost of $C_p(m)$, which increases as m increases. Besides, T_p is assumed to be the expected PM duration.

Define $r_1 = U/W$, $r_2 = L/K$ and consider the following two cases under 2D PM strategy: Case a ($r_1 > r_2$) and Case b ($r_1 \leq r_2$). For Case a, the following three subcases are investigated, including (1) $r \leq r_2$; (2) $r_2 < r \leq r_1$; (3) $r > r_1$.

(1) $r \leq r_2$

In this subcase, the warranty expires at time point (W, rW) . Under the 2D PM strategy, the PM actions are scheduled with interval K and when the j th PM activity ends, $\tau_j = j(K + T_p)$. Then the number of PM actions implemented during the warranty coverage can be obtained as:

$$n_1 = \max \{ j \mid j(K + T_p) < W, j \geq 0 \} \quad (7)$$

Note that, after the n_1 th PM actions, the manufacturer would not implement additional PM action and the subsequent interval will end when the warranty terminates. In addition, the occurrence of failures between two successive PMs, which are all minimally rectified and follows a non-homogeneous Poisson process (NHPP). Thus, the conditional expected number of failures can be expressed as:

$$EN_1^r = \sum_{j=0}^{n_1-1} \int_{j\delta(m)(K+T_p)}^{j\delta(m)(K+T_p)+K} \lambda(t|r)dt + \int_{n_1\delta(m)(K+T_p)}^{n_1\delta(m)(K+T_p)+W-n_1(K+T_p)} \lambda(t|r)dt \quad (8)$$

Hence, conditional on $R = r \leq r_2$, the expected warranty servicing cost $EC_1(\Omega)$ can be obtained as:

$$EC_1(\Omega) = EN_1^r C_f + n_1 C_p(m) \quad (9)$$

Furthermore, for this case MDT consists of minimal repair durations and PM durations. When $r \leq r_2$, MDT_1 can be expressed as:

$$\text{MDT}_1 = n_1 T_p + EN_1^r T_f \quad (10)$$

For an item with $r \leq r_2$, the warranty ceases at time W . Therefore, the expected operational availability $EA_1(\Omega)$ is given by:

$$EA_1(\Omega) = \frac{W - \text{MDT}_1}{W} \quad (11)$$

(2) $r_2 < r \leq r_1$

For an item with $r_2 < r \leq r_1$, the warranty terminates at time point (W, rW) and PM activities are carried out with interval L . Therefore, the number of PM actions and the conditional expected number of failures over the warranty period are given by:

$$n_2 = \max \{ j \mid j(L/r + T_p) < W, j \geq 0 \} \quad (12)$$

$$EN_2^r = \sum_{j=0}^{n_2-1} \int_{j\delta(m)(L/r+T_p)}^{j\delta(m)(L/r+T_p)+L/r} \lambda(t|r)dt + \int_{n_2\delta(m)(L/r+T_p)}^{n_2\delta(m)(L/r+T_p)+W-n_2(L/r+T_p)} \lambda(t|r)dt \quad (13)$$

Further, conditional on $R = r$ ($r_2 < r \leq r_1$), the expected warranty cost $EC_2(\Omega)$ and operational availability $EA_2(\Omega)$ can be obtained as follows:

$$EC_2(\Omega) = EN_2^r C_f + n_2 C_p(m) \quad (14)$$

$$EA_2(\Omega) = \frac{W - (EN_2^r T_f + n_2 T_p)}{W} \quad (15)$$

(3) $r > r_1$

Similar to the above analysis, conditional on $R = r > r_1$, the expected warranty cost $EC_3(\Omega)$ and operational availability $EA_3(\Omega)$ can be obtained:

$$EC_3(\Omega) = EN_3^r C_f + n_3 C_p(m) \quad (16)$$

$$EA_3(\Omega) = \frac{U/r - (EN_3^r T_f + n_3 T_p)}{U/r} \quad (17)$$

where n_3 is the number of PM actions, $n_3 = \max \{ j \mid j(L/r + T_p) < U/r, j \geq 0 \}$; EN_3^r is the conditional expected

number of failures, $EN_3^r = \sum_{j=0}^{n_3-1} \int_{j\delta(m)(L/r+T_p)}^{j\delta(m)(L/r+T_p)+L/r} \lambda(t|r)dt + \int_{n_3\delta(m)(L/r+T_p)}^{n_3\delta(m)(L/r+T_p)+U/r-n_3(L/r+T_p)} \lambda(t|r)dt$.

Finally, removing the conditioning on R , the expected warranty servicing cost $EC(\Omega)$ and operational availability $EA(\Omega)$ in Case a, can be obtained as, respectively:

$$EC(\Omega) = \int_{r_{\min}}^{r_2} EC_1(\Omega) dG(r) + \int_{r_2}^{r_1} EC_2(\Omega) dG(r) + \int_{r_1}^{r_{\max}} EC_3(\Omega) dG(r) \quad (18)$$

$$EA(\Omega) = \int_{r_{\min}}^{r_2} EA_1(\Omega) dG(r) + \int_{r_2}^{r_1} EA_2(\Omega) dG(r) + \int_{r_1}^{r_{\max}} EA_3(\Omega) dG(r) \quad (19)$$

Similar to Case a, we also need to consider the following three subcases in Case b. They are: (1) $r \leq r_1$; (2) $r_1 < r \leq r_2$; (3) $r > r_2$. Following the similar procedures, the warranty cost and availability models for Case b can be obtained. Limited the length of this paper, detailed derivation process of Case b is omitted.

Under the availability-based warranty policy, the manufacturer optimizes PM schedules during the warranty period, so as to meet a negotiated availability level A_0 of the equipment with the lowest cost, that is, the optimal PM decision space Ω to minimize $EC(\Omega)$ with an availability constraint. Therefore, the availability-based warranty optimization model is given by:

$$\begin{cases} \min & EC(\Omega) \\ s.t. & EA(\Omega) \geq A_0 \\ & K, L > 0, 0 \leq m \leq M \end{cases} \quad (20)$$

3. Case Study

Considering a repairable automobile component, the manufacturer provides an availability-based warranty service for each product with warranted availability level 0.96, and warranty coverage of 3 years/ 30,000 km. In addition, the manufacture implements 2D PM activities free of charge so as to minimize warranty cost with the availability requirement. It is assumed that the usage rate R follows uniform distribution with $r_{\min} = 0.2$ and $r_{\max} = 1.8$ (10^4 km/year). The values of θ_0 , θ_1 , θ_2 and θ_3 are set to be 0.1, 0.2, 0.7 and 0.7, respectively. The corresponding maintenance time are as follows: $T_f = 0.02$ years, $T_p = 0.004$ years. Besides, set $m = 0, 1, 2, 3, 4, 5$, and the relevant PM cost is $C_p(0) = \$0$, $C_p(1) = \$10$, $C_p(2) = \$30$, $C_p(3) = \$60$, $C_p(4) = \$100$ and $C_p(5) = \$160$ (Cheng, et al. 2015; Wang and Su, 2016).

The grid search in specific regions is performed to optimize PM decisions in this work. For the sake of efficient maintenance management, the search steps of K and L are one month and 10^3 km respectively (Wang and Su, 2016). Under 2D PM strategy, the optimal decisions $\Omega^* = (K^*, L^*, m^*)$, the expected warranty cost EC and expected operational availability EA with varying minimal repair cost C_f are summarized in Table 1. For example, when $C_f = \$150$, PM actions should be scheduled every 12 months or

10, 000 km and the corresponding expected warranty cost and operational availability are $EC = \$548.12$ and $EA = 0.97457 > A_0$. As shown in Table 1, appropriate 2D PM activities carried out by manufacturers can effectively satisfy products' availability requirements during the warranty period. In addition, it is found that K^* and L^* gradually decrease as C_f increases, while m^* tends to increase as C_f increases. It can be interpreted by the fact that the manufacturer tends to improve the level of PM actions or reduce PM intervals, so as to reduce the occurrence of failures and compensate the increase of warranty cost.

Table 1 Optimal PM decisions under 2D PM strategy.

C_f	K^*	L^*	m^*	EC	EA
50	18	15	2	226.96	0.96783
100	12	10	3	405.41	0.97457
150	12	10	3	548.12	0.97457
200	9	8	3	690.35	0.97529
250	9	8	3	817.94	0.97529
300	9	8	3	945.53	0.97529
350	9	8	4	1065.10	0.97810
400	9	8	4	1174.40	0.97810
450	9	8	4	1283.70	0.97810
500	9	6	4	1391.12	0.97809

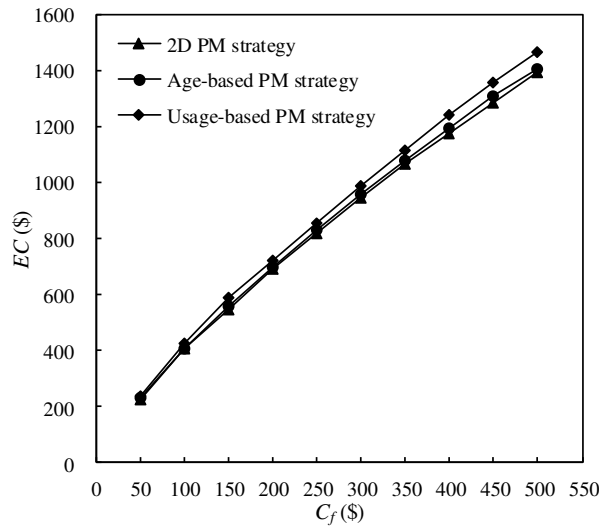


Figure 3. Expected warranty cost under PM strategies.

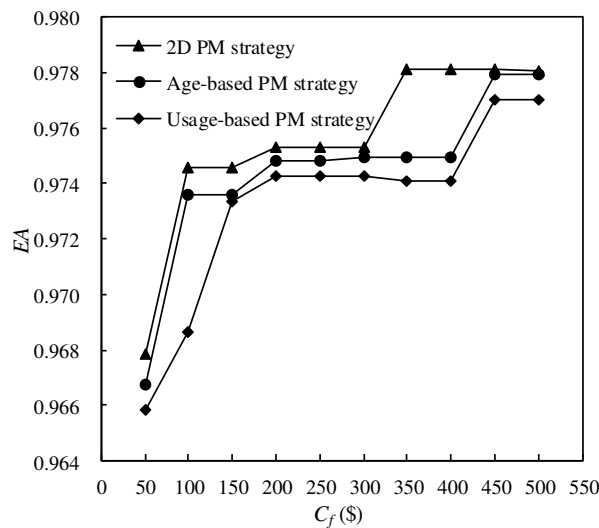


Figure 4. Expected operational availability under PM strategies.

In order to further illustrate the superiority of 2D PM strategy, the results for age-based and usage-based PM strategies can be used as the references. From Figure 3, the expected warranty cost EC under three PM strategies is compared. It is obvious that in terms of warranty cost, 2D PM strategy is superior or at least identical to 1D PM strategy. What's more, as can be seen from Figure 4, the expected operational availability EA for 2D PM strategy is higher in comparison to the other PM strategies. Hence, whether in terms of availability or warranty cost, 2D PM strategy is superior to the age-based and usage-based strategies under the availability-based warranty policy.

4. Conclusion

In this paper, the availability-based warranty policy is investigated in the framework of 2D warranty. From the manufacturer's perspective, rational 2D PM schedules are carried out with a view to satisfying the availability requirement of product and minimizing warranty cost within the warranty period. Numerical example shows that under the availability-based warranty policy, 2D PM strategy is superior to the age-based and usage-based strategies. In future studies, the PM expenditure is shared partially by customers, instead of being paid by manufacturers. The usage rate can also be assumed to be other distributions, e.g., normal distribution or Weibull distribution.

Acknowledgements

This research was supported by the National Natural Science Foundation of China under No. 71671035, the Open Fund of Jiangsu Wind Power Engineering Technology Center, China under No. ZK15-03-01 and ZK16-03-07, the Fundamental Research Funds for the Central Universities of China.

References

- Yeh C.W. and Fang C.C. (2015) Optimal decision for warranty with consideration of marketing and production capacity. *International Journal of Production Research* **53**, 5456-5471.
- Darghouth M.N., Ait-kadi D., and Chelbi A. (2017) Joint optimization of design, warranty and price for products sold with maintenance service contracts. *Reliability Engineering & System Safety* **165**, 197-208.
- Cheng Z.H., Yang Z.Y., Zhao, Z., Wang, Y.B., and Li, Z.B. (2015) Preventive maintenance strategy optimizing model under two-dimensional warranty policy. *Eksploracja i Niezawodnosc - Maintenance and Reliability* **17**, 365-372.
- Yan, N., Rundle, P., and Pecht, M. (2011) Prognostics and Health Management's Potential Benefits to Warranty. In *Proceedings of the Seventh Annual Warranty Chain Management Conference*, 15-17 March 2011, pp. 1-9. San Diego, CA.
- Iskandar, B.P., Husniah, H., and Pasaribu, U.S. (2014) Maintenance Service Contracts for Equipment Sold with Two Dimensional Warranties. *Quality Technology & Quantitative Management* **11**, 321-333.
- Su, C. and Wang, X. L. (2016) A study of availability-based warranty policy. In *Proceedings of the 2016 IEEE International Conference on Industrial Engineering and Engineering Management*, 4-7 Dec. 2016, pp. 1655-1659. IEEE, Bali, Indonesia.
- Wang, X. L. and Su, C. (2016) A two-dimensional preventive maintenance strategy for items sold with warranty. *International Journal of Production Research* **54**, 5901-5915.
- Su, C. and Wang, X. L. (2016) A two-stage preventive maintenance optimization model incorporating two-dimensional extended warranty. *Reliability Engineering & System Safety* **155**, 169-178.
- Wu, S. (2014) Construction of asymmetric copulas and its application in two-dimensional reliability modelling. *European Journal of Operational Research* **238**, 476-485.
- Iskandar, B.P. and Murthy, D.N.P. (2003) Repair-replace strategies for two-dimensional warranty policies. *Mathematical and Computer Modelling* **38**, 1233-1241.
- Kijima, M. (1989) Some Results for Repairable Systems with General Repair. *Journal of Applied Probability* **26**, 89-102.
- Kim, C.S., Djameludin, I., and Murthy, D.N.P. (2004) Warranty and discrete preventive maintenance. *Reliability Engineering & System Safety* **84**, 301-309.
- Huang, Y.S. and Yen, C. (2009) A study of two-dimensional warranty policies with preventive maintenance. *IIE Transactions* **41**, 299-308.

Computational Fluid Dynamics or Gaussian – is there a right way to model gas dispersion?

Amita Tripathi*, Claude Souprayen**, Andy Stanley*** and Naomi Warrilow****

* Fluidyn France, 7 boulevard de la Libération, 93200, Saint-Denis, France, amita.tripathi@fluidyn.com

** Fluidyn France, 7 boulevard de la Libération, 93200, Saint-Denis, France, claude.souprayen@fluidyn.com

*** RAS Ltd, 73 New Crane Street, Chester, CH1 4JE, andy.stanley@ras.ltd.uk

**** RAS Ltd, 73 New Crane Street, Chester, CH1 4JE, naomi.warrilow@ras.ltd.uk

Abstract. At present there are two main approaches to modelling gas dispersion within the process safety risk assessment arena. Historically the Gaussian approach has dominated, however, with the improvement of computing power available in relatively low-cost platforms, Computational Fluid Dynamics (CFD) has begun to challenge that dominance. Both approaches rely on a mathematical description of transport with advection-diffusion equations. However, there are a number of simplifications that are needed for analytical Gaussian solutions to be derived which are not needed for CFD solutions. The Gaussian approach often applies simplified flows over flat terrain, whereas CFD accounts for flow velocities and turbulence and their effects on mixing and can be applied in more complex three-dimensional environments. The questions this paper asks are: Is CFD modelling more mathematically correct than the Gaussian approach? And, Does the emergence of CFD into the mainstream pose an opportunity or threat? To answer these questions this paper will explore the previously mentioned differences between approaches.

1. Introduction

Within the process industries there is a need to understand risk.

When looking to understand the risk due to a release of flammable or toxic gas, models are required to simulate the dispersion. This then allows the hazard footprint and potential consequences to be understood. This in turn informs decisions on safeguards put in place to detect or respond to such an incident, or where it is acceptable or not to locate people. Big decisions are made on the grounds of the results of these models.

A model looks to simulate reality, but there are many constraints that mean simplifications have to be made. Rather than reality, best representations are made, proportionate to the level of risk that needs to be understood, or the level of investment that needs to be made.

For gas dispersion, Gaussian plume models are common place and have been used for many years. They are well known to be idealised and over simplified but have been the best available for most applications. More recently the opportunity to use Computational Fluid Dynamics (CFD) to look at the physics of the dispersing cloud has become available, as the constraints of time and resource are diminishing, with improving computing power. What was once inaccessible for mainstream risk assessment is now within reach.

This paper explores the differences between Gaussian plume models and those generated within CFD, presenting the main numerical techniques, and comparing the pros and cons of each. It will look to conclude which is the more mathematically correct approach.

The paper goes on to describe the current use of these models within the field of risk assessment for high hazard process industries. It will highlight the challenges faced by industry and the regulator in the use of these models.

Having set out the concerns associated with the use of models, this paper will culminate with a proposed way forward, informed by the progress made in this area by other countries.

2. Gaussian vs CFD – Mathematical framework and numerical solution

Dispersion of gases (pure species, then air and compound mixtures) involves their transport with the mean atmospheric flow (including the coherent flow structures) and mixing in ambient air via turbulent and ultimately molecular diffusion. Both Gaussian dispersion models and CFD dispersion models attempt to describe and solve such physical processes within a distinct mathematical and numerical framework. For the flow regimes at play, the mass conservation equation in Eulerian form for each dispersed species can be applied as a mathematical framework for dispersion. It reads:

$$\frac{\partial(\rho y_m)}{\partial t} + \nabla \cdot (\rho \mathbf{U} y_m) = \nabla \cdot D_m \nabla(y_m) + S_m \quad (1)$$

This partial derivative equation (PDE) for the mass fraction y_m of species m expresses from left to right, the local time variation rate of concentration for species m as the result of the advection of mass by velocity (3 component vector $\mathbf{U}$) and mixing with surrounding material under diffusive process with intensity D_m and a potential source input S_m (at leak location). In the above equation all variables (y_m , $\mathbf{U}$, D_m , S_m) are both time and space varying. The gas mixture density, ρ depends on the respective molar masses and fractions of the species. The major differences in Gaussian vs CFD approaches for solving this equation are the simplification vs solution for $\mathbf{U}$ and D_m in the above equation.

In standard Gaussian models, $\mathbf{U}$ and D_m are assumed to be constant (in time) and uniform (horizontally) with a single vertical (imposed) variation with altitude (related to atmospheric stratification). This simplification reduces the above equation to a diffusion equation (in a moving reference frame with velocity $\mathbf{U}$). An analytical solution for a Dirac delta source (in space and time) can be expressed with a Gaussian law function of shifted position ($r + \mathbf{U}t$) for the cloud centre, as for a diffusion process in a random motion/Brownian motion. The cloud from such a source is a gaussian plume with centre line and transverse profiles of concentration showing symmetric gaussian shapes with spreading size, as determined from turbulent intensity estimates in empirical law. The assumptions made here for $\mathbf{U}$ and D_m are not a true reflection of what is observed in industrial applications where there are terrain undulations (hills, cliffs) and obstacles (buildings, processes etc.).

Furthermore, many leak-related dynamical processes operate at short range from the source. Examples are the inertia and momentum from the release (high speed jet flow, explosive release), and the density of the release with either heavier (downward buoyancy) or lighter (or hotter) than air (for positive buoyancy). Therefore, the initial time step (and short range) sequences of dispersion for Gaussian models are often modified with other models termed as integral models: analytical jets or creeping behaviour, with empirical air entrainment rates and heat exchanges.

Again, these integral models are unable to account for the presence of obstacles, heat sources, or varied drag and turbulence coupling effects that exists on industrial sites.

A CFD approach on the other hand solves the full advection diffusion equation (1) along with the relevant fluid mechanic equations (Navier-Stokes Equations) for the atmospheric flow and the turbulence. The extended set of equations are again conservation equations for momentum, total mass, energy (or enthalpy) and turbulence closure intensities and dissipation rate.

Momentum conservation equations

$$\frac{\partial(\rho \mathbf{U})}{\partial t} + \nabla \cdot (\rho \mathbf{U} \mathbf{U}) = \nabla \cdot \tau - \nabla p + S_U \quad (2)$$

Total mass conservation equation

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{U}) = S_\rho \quad (3)$$

Energy Conservation equations

$$Cp \left[\frac{\partial(\rho T)}{\partial t} + \nabla \cdot (\rho \mathbf{U} T) \right] = -\nabla \cdot q + \left[\frac{\partial \rho}{\partial t} + \mathbf{U} \cdot \nabla \rho \right] + \tau : \nabla \mathbf{U} + S_T \quad (4)$$

These equations are universal fundamental (derived from 1st and 2nd principles) equations for the atmospheric flows and so can be applied to any industrial site with various terrain features (vegetation cover, hills, buildings, process units, storage tanks). The numerical solutions of the Navier-Stokes (PDE) equations provide the time and space variations of all necessary scalar and vector fields to be applied to the dispersion equations (ADE).

In the above PDE, the thermodynamic state for the gas mixture is enforced with perfect gas laws $f(P, \rho, T, y_m)$. For the CFD solution to be performed, it needs then to perform a series of steps:

- Design a numerical domain of simulation encompassing the environment and the industrial site with a 3D set-up for all above mentioned features that may perturb the flow and produce enhanced turbulence.

- Build a discrete approximate version (numerical discretisation on a mesh) of the above continuous set of equations.
- Set-up the boundary conditions (on the lateral faces of the domain for the selected meteorological scenario and at the source location of the leak).
- Solve the time and space discrete set of coupled equations for flow (u, v, w, P, T, r) and transport (y_m) for the species.

In the CFD solution, the major physical processes at play that are approximated (empirically) in Gaussian models are explicitly solved in the method, resulting in complex non-homogeneous 3D flows and complex dispersion patterns. Gravity effects (for heavy/light gases) are explicit in S_g for the external forces in momentum equations, pressure build-up in front of obstacles and pressure drop in the lee of buildings, and flow acceleration (corridors). In the discrete version of the Reynolds Average Navier-Stokes (RANS) equations (the standard method for capturing the large scale energy dominant flow structures in leak dispersion applications), a practical limitation on mesh size is applied, so small scale motions are below the grid scale. In solving for k (the turbulent kinetic energy) and ϵ (its dissipation rate), two specific (PDE) equations for turbulent fields are added. Both k and ϵ are local (time and space varying) scalar fields, explicitly solved for the account of turbulence enhancement/dissipation via mechanical (shear) or thermal production. The diffusivity scalar D_m in the dispersion equation (1) is estimated from k and ϵ fields. Thereby, it is neither empirically derived nor uniform in xy-space and time but rather derived from the turbulence fields.

The fluidynTM-PANACHE CFD model includes a dedicated ‘NT’ solver. The NT solver is a pressure-based fully implicit segregated method on meshes that solves all governing equations separately and uses an iterative method for both steady state and transient cases. To couple the pressure and momentum equations in the numerical computations, the Semi-Implicit Method for Pressure Linked Equations-Consistent (SIMPLEC or SIMPLE-Consistent) algorithm (Van Doormaal and Raithby, 1984) is utilized. The buoyancy model is used to parameterize the body force term in the N-S equations. Dispersion of the gaseous pollutants is modelled by solving the standard Eulerian advection-diffusion equation in a computational domain. Residuals on each field (and related discrete equations) are used to check the convergence of the solution during the simulation.

3. Gaussian vs CFD – Case study for dispersion of accidental releases

To demonstrate the difference between the Gaussian and CFD approach for cloud dispersion and subsequent consequence modelling, a case study for an accidental release of chlorine (heavy gas) is described below. The CFD Solution is modelled in fluidynTM-PANACHE dedicated software for atmospheric flows and dispersion at local meteorological scale (domains order of 10km size) and time scales of the order of minutes to hours.

The case study is for a major hazards site storing and handling chlorine. The study considered the scenario of an accidental release from a pipe manifold, in a central position on the plant. Key features included a complex series of obstacles within the site itself and a patchwork of surrounding surface types i.e. open flat vegetation vs residential zones, as shown in Figure 1 below.

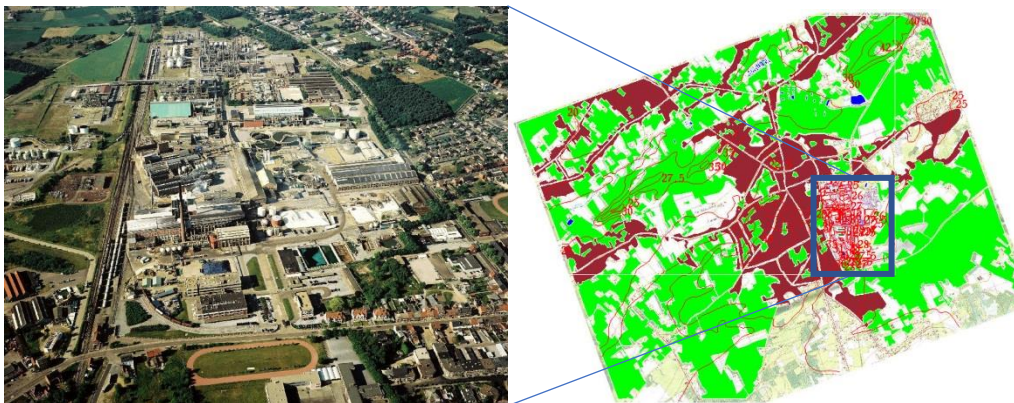


Figure 1: Onsite and offsite features for the accidental dispersion modelling (a) aerial view (b) environment and 3D set-up for CFD Solution.

A 3D representation was made and the atmospheric flows established. Following this, dispersion modelling was carried out using the CFD approach. The results of this modelling are shown in Figure 2 below.

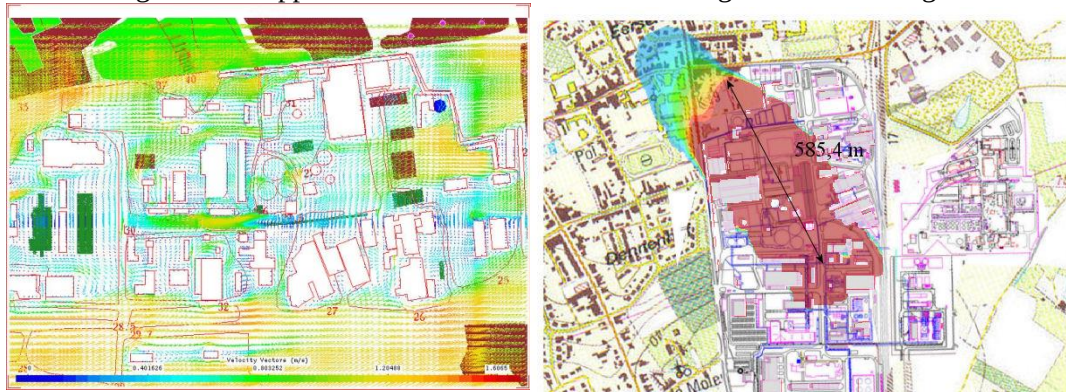


Figure 2: CFD Application for flow (left) velocities vectors and for toxicity map (both at ground level) for a consequence/risk analysis.

By comparison, the same dispersion was performed in a Gaussian (and integral) model (DNV GL Phast). This was compared against the results of the CFD model by overlaying the two on a map, as shown in Figure 3 below. It shows the results from the Gaussian model in the black contour, this model gives a smooth and idealised plume, compared to the red contours from the CFD model which show patches and pockets due to the complexity in the atmospheric flow. This is due to turbulent diffusion as induced by obstacles, flow shear and cloud splitting. In this case, the 1% lethal effect is shown in orange for the CFD model and the black contour for the Gaussian model, it can be seen that the effects of the lethal footprint are quite different.

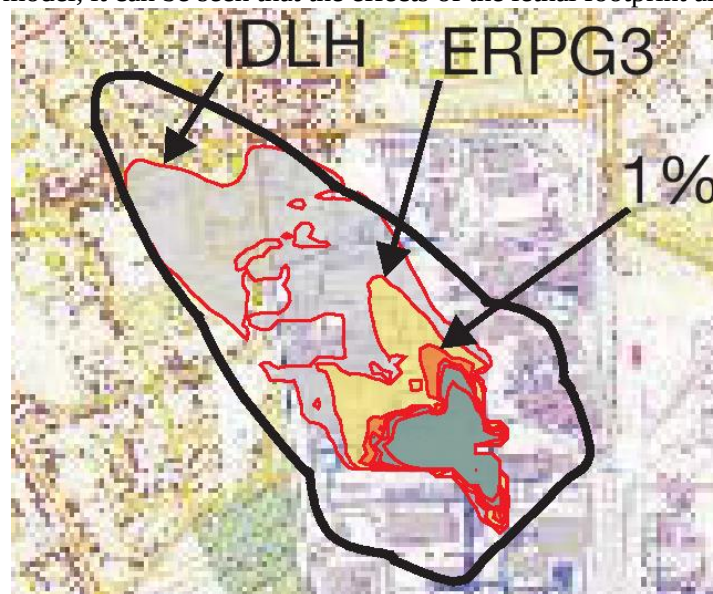


Figure 3: Comparison case between Gaussian model (black thick contour for Lethal effect 1%) and CFD model (red curves for varied thresholds).

The results of the two models if looked at in isolation would result in a different understanding of the risk from a toxic release and which would justify different levels of investment to mitigate such an incident. In this particular case study the hazard footprint is reduced by CFD modelling, but there are many examples where this can result in further reaching hazards.

4. Current Position

CFD is not currently adopted in the mainstream of risk assessment for high hazard sites. Practitioners have been deterred by the complexity and cost, and the regulator is wary that CFD can open up opportunities for error. There are many more variables that can be manipulated, judged incorrectly, or set in error. But are these concerns real, or are we just more comfortable on the well-trodden path?

In the UK, the HSE has a number of concerns with the wide spread use of CFD largely due to the variability which often arises between CFD results. CFD is well suited to models of highly complex situations by providing flexibility across a range of parameters, however, this also leads to a greater scope for error with reliance placed upon the competence of the user. Between different CFD models for the same scenario there can be large differences in the results produced. As HSL has reported in some recent papers on the subject, the differences can be of an order of magnitude. Furthermore, separate users of the same model may produce divergent results, due to level of experience with the programme or a lack of understanding of the mathematic principles behind the code.

The following table looks to consider each of these concerns for both Gaussian models and CFD to compare their relevance.

Table 1: Areas of concern and their relevance to Gaussian models and CFD

Area of Concern	Relevant to Gaussian models	Relevant to CFD
Lack of conformity between models	Little evidence of comparisons	Literature studies suggest large inconsistency
Flexibility of parameters	Simplification may be made to reduce the choice for users, but there are differences between the representative parameters. E.g. representation of Pasquill stability classes varies between commonly used models	Flexibility in choice of parameters, if not set by rules
Software user competence	Often user-friendly interfaces, but can be misused as user can be given false sense of ease of use without understanding limitations or mathematics of the model	Often user interface is complex, so unlikely to be used by analysts that do not appreciate the development of the model
Lack of validation and verification	Validation is often for narrow area of use, however the models are used beyond this	Large amount of effort in validation and verification, however this varies between model providers, in particular in how forthcoming they are with evidence
Few large scale experiments to compare with models	Same experiments available for all models	

As summarised here, many of the challenges facing the use of CFD are not new or specific to its use but are instead universal challenges for any form of modelling.

5. Looking for a better way

1.1. Validation and Verification

Efforts have been made to improve the convergence between CFD modelling results from different models. These include the publication of best practice guidelines which assign the validation and verification requirements for CFD models.

A number of best practice guidelines have been produced. In the United States where a number of guidelines for CFD modelling are in place. Notable amongst these is an ASME standard for the verification and validation of CFD modelling, though this is heat transfer focused (ASME, 2009). Other examples of international guidelines include the Guidelines for Nuclear Safety Evaluation Defined by the Atomic Energy Society of Japan (AESJ) which were created to utilise CFD models for dispersion calculations for nuclear safety evaluation of radiation dose. The guidelines incorporate a verification and validation scheme using experimental data and dictates that all final calculated results are to be modified using an uncertainty factor as suggested in the ASME code.

In the UK the HSE endorses the production of such guidelines, in order to improve conformity in CFD modelling outcomes. They have presented on the topic at various conferences, but they have not issued formal

guidance. There are a large number of guidelines, but there is currently no single point of reference for what good looks like for us to apply in the UK.

1.2. Rules for Modelling

As well as seeking to find common ground on the approach to validation and verification, some good practice is emerging on how to use models.

The Atmospheric Dispersion Modelling Liaison Committee (ADMLC) recently updated guidelines for short range dispersion modelling of releases to atmosphere from industrial sources. The original guidelines were introduced in 1995 and ADMLC believed that modelling techniques have evolved sufficiently to require additional advice (ADMLC, 2004). These guidelines are generic, and apply to all forms of dispersion modelling, acting as a universal guide to ensure regulatory conformity.

Notably, the NAFEMS (National Agency for Finite Element Methods & Standards) has a CFD working group, which includes members from industry, academics and consultants. This group has produced 'How to guides' with CFD best practice guidelines, but these are only made available to members. They also lack input from regulators.

A French working group, created in 2009 at the request of the French Ministry of Ecology, Energy, Sustainable Development and Spatial Planning, investigated gas dispersion modelling for land use planning near high hazard industries. By first assigning the same modelling task to separate modelling groups the study identified variables between the models and investigated methods of regulating certain parameters. The outcome of this work was a definition of best practice for the use for 3D modelling of atmospheric dispersion. This included guidelines for validation and also specific parameters within the modelling relating to mesh, boundary conditions, the definition of a source term and the modelling of turbulence (SAN JOSE & LUIS PEREZ, 2013). There have been a number of other working group studies relating to CFD, but few in relation to atmospheric dispersion and in particular in relation to high hazard industries. Hence, the outcomes of this group were particularly relevant for our consideration. The work of this group is not yet widely adopted and, other than a summary, has not been translated from French.

There is currently no accepted rule set adopted as best practice within the UK. This leaves too much choice in how to represent situations which will introduce variance in the solutions provided by different modellers.

6. The Way Forward

In order to understand and manage risk from high hazard industries it is important that we continuously improve and embrace new technologies to move us forward in our knowledge and understanding. Now that CFD is more accessible in terms of computing power and affordability, it is only worries around application that hinders its application.

It is clear that when looking to understand a complex situation, oversimplification will not be helpful. But we should also recognise that trying to simulate everything may be unnecessary. What we need is a middle ground. An agreement of where extra sophistication is needed or not. And then when we use more sophisticated models, to lock down the variables that don't need to change. By codifying the aspects that should be common to all, or that are less important, we avoid the issue of there being too many variables, so reduce the chance for inadvertent error.

To enable repeatability and comparability between people using these models, we need to agree a set of rules. These rule sets need to be developed for specific applications, such as the one discussed here: atmospheric dispersion.

References

- ADMLC (2004) Guidelines for the Preparation of Dispersion Modelling Assessments for Compliance with Regulatory Requirements – an Update to the 1995 Royal Meteorological Society Guidance. Available from: <https://admlc.com/model-guidelines/>
- ASME (2009). Standard for Verification and Validation in Computational Fluid Dynamics and Heat Transfer. Available at www.asme.org

- Batt, R., Grant, S.E., Lacome, J-M., Truchot, B., Tucker, H. (2016) CFD Modelling of Dispersion in Neutral and Stable Atmospheric Boundary Layers: Results for Prairie Grass and Thorney Island: *17th International Conference on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes*, Budapest, Hungary, 9-12 May 2016
- Franke, J., Hellsten, A., Schlünzen, H., Carissimo, B., (2007) Best Practice Guideline for the CFD Simulation of Flows in the Urban Environment Action COST 732 – Quality Assurance and Improvement of Microscale Meteorological Models, *International Journal of Environment and Pollution*. 44 (1-4)
- Gant, S.E., Tucker, H., (2018) Computational fluid dynamics (CFD) modelling of atmospheric dispersion for land-use planning around major hazards sites in Great Britain, *Journal of Loss Prevention in the Process Industries*
- HSE (2017) RR1099 - Review of consequence model evaluation protocols for major hazards under the EU SAPHEDRA platform. Available at www.hse.gov.uk
- Meroney, R., Ohba, R., Leidl, B., Kondo, H., Grawe, D., Tominaga, Y., (2016). Review of CFD Guidelines for Dispersion Modelling, *Fluids*
- SAN JOSE, R., LUIS PEREZ, J. (2013) Harmonisation of Practices for Atmospheric Dispersion Modelling within the Framework of Risk Assessment. *15th International conference on harmonisation within atmospheric dispersion modelling for regulatory purposes (HARMO 15)*, May 2013, Madrid, Spain. Technical University of Madrid. Madrid, pp.718-723,
- Van Doormaal, J. P., Raithby, G. D. (1984) Enhancements of the simple method for predicting incompressible fluid flows. *Numerical Heat Transfer* 7 (2), pp.147-163

Research on Voice Interaction for Augmented Reality assisted Maintenance

Ranran Wang^{1*}, Chuan Lyu², Shunfeng Mei³, Chunhui Guo⁴, Jie Geng^{5**} and Yu Xiao⁶

^{1*} School of Reliability and Systems Engineering, Beihang University in Beijing of China + wrrbuaa@163.com

² School of Reliability and Systems Engineering, Beihang University in Beijing of China + lc@buaa.edu.cn

³ School of Reliability and Systems Engineering, Beihang University in Beijing of China + meisf125@163.com

⁴ School of Reliability and Systems Engineering, Beihang University in Beijing of China + guochunhui@buaa.edu.cn

^{5**} School of Reliability and Systems Engineering, Beihang University in Beijing of China + wrrbuaa@163.com

⁶ School of Reliability and Systems Engineering, Beihang University in Beijing of China + holomoon@me.com

Abstract. This paper aims at the process of augmented reality maintenance, and studies the voice interaction method to realize a more natural and efficient human-computer interaction between maintenance personnel and virtual objects. Study the application and functional requirements of voice interaction in augmented reality assisted maintenance by analyzing the current interaction status. Determine the information transmission between computer and maintenance personnel and establish a knowledge base for maintenance process. Thus, a suitable semantic-based model of voice interaction mechanism is designed.

In this paper, speech signal and text information are converted by speech recognition and synthesis technology. And the semantic understanding method based on keyword matching is used to analyze the order of maintenance personnel, so as to convey the feedback information to maintenance personnel or make the operating system perform corresponding functions. Finally complete the communication between the maintenance personnel and the virtual objects.

1. Introduction

Augmented reality is a new technology developed on the virtual reality. With the help of computer image and data interaction technology, virtual information can be mapped in real time to real scene, and human perception of real scene can be increased. At present, it is widely used in various fields.

For the assembly system with a large number of components and complex constraints, the primary maintenance personnel have limited efficiency and long training time. Especially in the complex maintenance environment, the mis-operation probability is higher.

But different from virtual maintenance, a typical augmented reality system is user-centric which display virtual information through a headband display or projector and so on. Therefore, human-computer interaction is one of the key technologies, and we seek more possibilities for the communication between maintenance personnel and virtual objects. The traditional human-computer interaction technology is completed by the mouse and the keyboard, and now people want some simple and natural ways of information exchange, such as posture, facial expression, speech in interpersonal communication.

Currently, interactive research in augmented reality focuses more on human gestures, such as the HoloLens produced by Microsoft. Because of the particularity of the maintenance environment, the posture of maintenance personnel is relatively fixed, and their hands cannot completely free. The semantic meaning of the gesture is irrelevant to the movement of the finger, so the user's operation is not natural and inconvenient for memory. The application of voice interaction can overcome these difficulties. With the development of speech recognition and synthesis technology, voice interaction has become an important interactive mode of virtual assembly technology in the field of industrial design and simulation. To use voice guidance can greatly save time between man and machine. The maintenance personnel get rid of the constraint of fixed posture through the voice interaction mechanism. The system can quickly and accurately identify the user's intent and pass it to the corresponding application.

The research of speech interaction involves the application of speech recognition and synthesis technology, the research of semantic understanding, the design of speech interaction mechanism and the construction of speech interaction system. The research process of this paper is shown in the following figure 1.

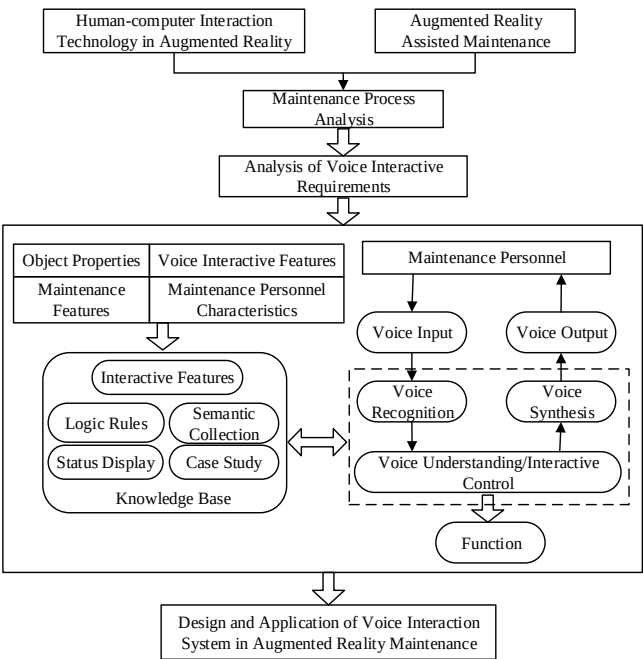


Figure 1. The research process of the thesis

2. Applicability of Speech Interaction and Functional Requirements Analysis in Augmented Reality Maintenance

2.1 The analysis of the interactive features in augmented reality assisted maintenance

The application of augmented reality technology in maintenance field can make up for the deficiency of virtual maintenance technology in the real maintenance scene. It provides more information to users in the complex equipment maintenance, and even manipulate the presentation intuitively. And it provides accurate and real-time help for professional users, and convenience to learning and training for beginner. It is of great significance to shorten the maintenance cycle of equipment and improve the efficiency of maintenance.

The difference between augmented reality maintenance and virtual maintenance is that the former users place in the real environment and have higher requirements of real-time capability. As a result, users have high demands for interactivity, and they hope a more natural way to interact with machines.

At present, interactive research in Augmented Reality focuses on gesture interaction. However, in the actual maintenance, especially in the case of battlefield repair, gestures have a long interaction time and the hands of the maintenance personnel are occupied which make the operation inconvenient. Through the above analysis, it is necessary to study the voice interaction function, and the voice interaction can make up for the shortcomings of the gesture interaction. After the voice interaction system has built, it can be combined with other ways to realize multi-channel interaction.

2.2 Analysis of voice interactive requirements in augmented reality maintenance

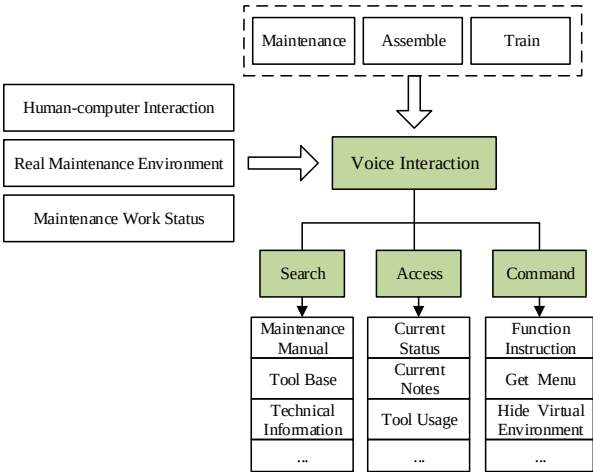


Figure 2. Analysis of voice interactive requirements for augmented reality maintenance

The influence factors in interactive process of augmented reality maintenance are analyzed comprehensively. It including the interactive mode, the real maintenance environment and the working state of the maintenance personnel. And then learn from the voice interaction requirements of virtual maintenance, assembly and personnel training to determine the interactive requirements in the augmented reality maintenance. The results are divided into three parts: information search, resource access and command interaction. As shown in the following figure 2.

- Users can search the maintenance information, such as the current maintenance status, attention, tool usage and other dynamic information. To realize the search function needs the auxiliary support of other technology modules in the augmented reality system, such as scene acquisition module and tracking registration technology. For example, if the maintenance personnel input " What is the next maintenance operation? " by voice, then the system can prompt it according to the current maintenance status.
- Resource access is a variety of technical data required for the maintenance process, including electronic maintenance manual, maintenance tool library and maintenance gesture library. For example, if the maintenance personnel input " Which kinds of maintenance tools will be used in the maintenance process for XX?" by voice, the system will provide the list of tools according to requirements to help the maintenance personnel prepare in advance.
- Command interaction includes function selection, instruction information confirmation and so on. In the operation process, the basic selection and determination function can be realized by voice interaction. For example, the system prompts "Please enter your selection.", then users can say the name or number of the object to be selected directly. Due to the different personal language habits or the diversity of the maintenance environment, it is necessary for the system to confirm the order to ensure correctness. Such as "whether to confirm the exit?".

3. research on voice interaction mechanism in augmented reality assisted maintenance

When analyzing the requirements of voice interaction functions in augmented reality maintenance, it is necessary to study the voice interaction mechanism, and determine the way of information transmission between computer and maintenance personnel. It is the basis for researching the voice interaction method and constructing interactive system. The conversion between command information and voice signal is through speech recognition and synthesis technology, and the communication between maintenance personnel and system is realized through semantic understanding.

Users can send the voice information to the voice recognition module by the hardware such as a microphone, and then extract the characteristic parameters of the identified text information to match the keyword with the knowledge base and the case base. Output the result with the highest matching degree to the dialogue management module, and confirm the completion of the appropriate control commands. The voice interaction mechanism for augmented reality maintenance is as shown in the figure 3.

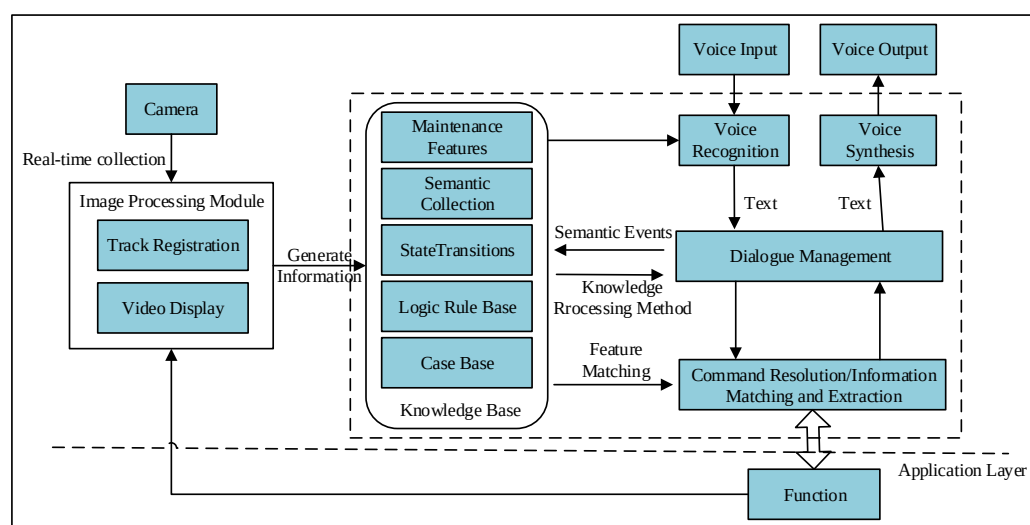


Figure 3. Analysis of voice interactive mechanism for augmented reality maintenance

It mainly includes speech recognition and synthesis module, semantic understanding module, dialogue management module and knowledge base. The knowledge base is the basis for understanding the user's intention, including maintenance features, semantic template library, logic rule base, status information and

so on. All the information needed in the maintenance process comes from the knowledge base. The case base which can record the information in each identification process and simplify the matching process to a certain extent. It can be updated through continuous learning.

Speech recognition and synthesis module relies on speech recognition technology and speech synthesis technology. It can convert the user's voice signal and text information. Through semantic understanding, it can parse the user's instructions and translate the identified text information into a language which can understand by machine. In the voice interaction system, it can convert information into commands understandable by the control module so that the function can be executed and output voice prompt; It can also analysis based on users requirements, and retrieve the information in the knowledge base. Convert it to a natural language text that the users can understand according to certain grammar file, and then converts the text into the voice signal to the user through the speech synthesis technology.

The dialog management module is the brain of the voice interaction system. It mainly includes two steps:

- 1) Remember the current state of the system runtime. Select the appropriate action after the conversation occurs, and switch to the next status according to the dynamic transition diagram.
- 2) Maintain and update the state of the dialogue. The dialogue state is the data representation that the machine can handle, including information that affects the decision making.

Maintenance status during the maintenance is relatively limited, it can directly define the state of the dialogue system based on Finite State Machine (FSM)

4. Speech recognition and semantic understanding

4.1 Semantic recognition and understanding

Speech recognition refers to the technology that the voice system can transform the natural speech signal into a text file in computer-readable or executable command in real time. Speech recognition module usually consists of three parts: speech feature extraction, pattern matching and reference pattern library. It needs to sense the scene actively.

During the maintenance process, maintenance personnel input the maintenance problems encountered by voice into the system. At present, the mature speech recognition technology can convert maintenance personnel commands into words. Using semantic comprehension method based on keyword matching analyze the maintenance personnel' intention, and match the extracted characteristic parameters with the maintenance semantic set. And output the result with the highest matching degree to the dialog management module, so as to realize operation or convey information to maintenance personnel.

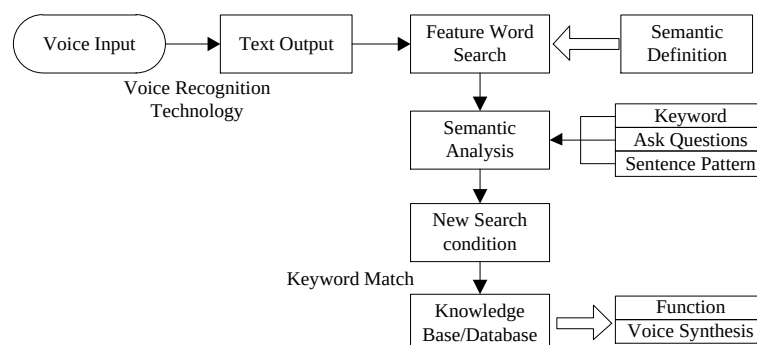


Figure 4. Speech recognition and understanding model

This process focuses on how to complete the semantic understanding of speech signals. First, the semantic definition of the user terms in the knowledge base is made. Due to the existence of a certain structure of Chinese information, we can obtain the characteristic words that can match the knowledge base by analyzing the natural language phrases of the user's expression search target. The characteristic words are extracted from the textual information so as to make a semantic analysis of the key words. The way of asking questions and the key points of the questions form a new retrieval condition , and use the keyword matching method to retrieve the required information in the knowledge base.

4.2 The establishment of the library

The construction of the knowledge base and case base in the semantic understanding model is a key issue. The third chapter introduces the contents of the knowledge base, and the specific analysis will be carried out as follows. They are the core components of the semantic understanding part, which support the user's intention recognition and implementation. The lack of comprehensive knowledge base construction directly leads to the speech interactive system cannot obtain the correct recognition result. Therefore, the knowledge base should be built on the basis of a comprehensive analysis of the requirements of language interaction for augmented reality maintenance. In addition, the structure of knowledge base and case base should be designed to be as complete and clear as possible for retrieval.

Knowledge base includes dictionary library, semantic template library, logical reasoning library and case base. The dictionary library is used to store the different expressions of the voice instruction vocabulary. The semantic template library is used to match the structure of the language when the user semantic parsing is performed, so as to obtain the resolution of the user's intention. Logical reasoning library is used to record the prerequisites and interaction events triggered by maintenance events to support the system to determine the application. The experience database is used to record the matching process and result of each interactive event processed by the system to improve the accuracy of system identification. The case base contains a record of each matching process. It learn each new case to complete the dynamic build by learning components.

The semantic collection provides support for speech recognition, and summarizes the required vocabulary. It completes the construction of the dictionary by integrating the vocabulary encoding, the meaning of the vocabulary, the description of the vocabulary and the description of the vocabulary text by creating a limited collection. Define the knowledge of maintenance feature words and functions used in the actual repair. The complete semantics set is a necessary condition for accurately identifying the content of the user's input voice instruction. On the basis of the complete analysis of the functional requirements of voice interaction, the user terms used in the maintenance process are analyzed and sorted out. Establish proper grammar files according to the features of the maintenance process function instruction and the communication characteristics between maintenance personnel and virtual environment in actual maintenance process. The user terms is integrated into the semantic collection through a certain grammatical format, forming a complete user instruction dictionary, as shown in figure 5.

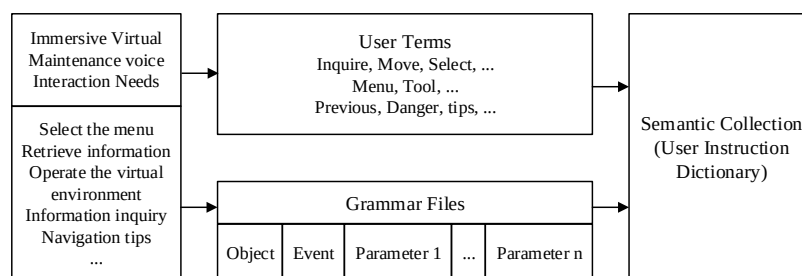


Figure 5. The process of establishing the library

5. Speech synthesis or system execution

This part contains natural language synthesis and speech signal synthesis. Natural language generation is a branch of artificial intelligence and computational linguistics, as opposed to natural language analysis. It divides into textual planning stage and surface form generation stage. Starting from the conceptual level of abstraction, the former is the information what should be told and be expressed. The latter produces text by selecting and executing certain semantics and grammar rules determine how the text is expressed. For the possible interaction of sentences in augmented reality system, corresponding sentence patterns are constructed which includes constants and variables. For example, "The tool used in this step is XX", " Next, please unscrew the screw " and so on. After the semantic understanding is completed and the information to be conveyed is confirmed, the text generator embeds the information as a string into the template instead of the variable to generate a text message that expresses reasonableness.

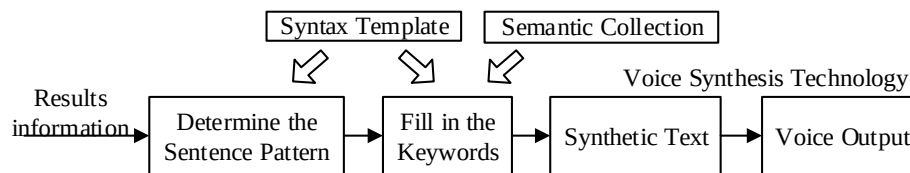


Figure 6. Speech synthesis model

The synthesized text is converted to standard natural spoken language in real time according to the voice signal processing rules. And system convey the voice signal to the maintenance personnel and accept the information feedback of maintenance personnel. Thus complete the interactive information process.

6. Conclusion

The application of voice interaction in augmented reality maintenance can enhance information exchange and give more operability for users, and improve the work efficiency of maintenance personnel. This paper mainly studied the speech interaction method in augmented reality maintenance, and the interaction mechanism is designed to lay the foundation for the realization of the voice interactive system. The main research significance is as follows: Promoting the applicability of augmented reality technology in the maintenance field; Provided a maintenance-oriented human-computer interaction mode; The role of voice interaction in virtual maintenance continues to be played in the field of augmented reality maintenance.

Semantic understanding is the core of interactive process. In this paper, the understanding method of keyword matching is proposed, but there is no specific research on it. This is also the direction of future research. After perfecting the research of above technology in the method, the system of voice interaction can be established. With the help of HoloLens and other hardware devices, it can be applied in the actual maintenance to test and improve.

References

- [1]Guo, Z., Lv, C., Zhou, D., & Wang, Z. (2016). Human-computer interaction in immersive virtual maintenance. *Intelligent Control and Automation* (pp.2567-2573). IEEE.
- [2]Ke, C., Kang, B., Chen, D., & Li, X. (2005). An augmented reality-based application for equipment maintenance. *International Conference on Affective Computing and Intelligent Interaction* (Vol.3784, pp.836-841). Springer-Verlag.
- [3]Hamidia, M., Zenati, N., Belghit, H., Guetiteni, K., & Achour, N. (2016). Voice interaction using Gaussian Mixture Models for Augmented Reality applications. *International Conference on Electrical Engineering*. IEEE.
- [4]Araújo, Tiago, et al. "Aspects of Voice Interaction on a Mobile Augmented Reality Application." (2016)..
- [5]Xincan, Z., Zuo Hongfu|赵新灿, and 左洪福. "Research on Interaction Technique of Augmented Reality Maintenance Guiding System." *China Mechanical Engineering* 19.11(2008):1312-1316.
- [6]Chen, Guiling, et al. "Virtual human control technology for immersed virtual maintenance system in DELMIA environment." *International Conference on Audio Language and Image Processing IEEE*, 2010:1352-1355.

An unpunctual preventive maintenance policy for repairable items sold with a two-dimensional warranty

Xiaolin Wang*, Lishuai Li, Min Xie

Department of Systems Engineering and Engineering Management, City University of Hong Kong, Kowloon, Hong Kong.

*Correspondance author + email address: xlwang28-c@my.cityu.edu.hk

Abstract The optimization of preventive maintenance (PM) strategies for repairable items sold with warranty contracts has received much attention in the literature. However, the existing research implicitly assumes that maintenance actions within the warranty period are punctual. In practice, it is not uncommon that the actual maintenance instants deviate from the scheduled instants. In this paper, an unpunctual imperfect PM strategy, which allows customers to advance or postpone the scheduled PM actions in a tolerable range, is proposed for repairable items sold with a two-dimensional warranty. The objective of this work is to determine the optimal unpunctual PM strategies under the given warranty period so as to minimize the manufacturers total expected warranty servicing cost. It is shown that the unpunctual PM strategy contains its punctual counterpart as a special case and tends to result in slightly higher warranty servicing cost.

1. Introduction

Nowadays, almost all products are sold with warranty contracts due to competition, industrial obligations, and/or customer requirements. However, most manufacturers regard warranty as an overhead item, and they are always under pressure to keep warranty servicing costs staying within limits. Incorporating appropriate preventive maintenance (PM) programs into the warranty policy is one feasible way of reducing warranty servicing cost (Wang and Xie 2018). From the cost-benefit perspective, the optimal trade-off between the reduction of warranty cost and the additional cost incurred by the PM program becomes necessary in order to obtain an optimal PM strategy for the manufacturer.

This paper is interested in investigating optimal imperfect PM strategies for repairable items sold with a two-dimensional (2-D) warranty. A 2-D warranty simultaneously employs two variables, usually age and usage, to characterize the warranty policies. In the 2-D warranty context, there are some references dealing with the optimization of PM strategies; see Wang et al. (2015), Wang and Su (2016), Su and Wang (2016), and Huang et al. (2017), for recent ones. Specifically, Wang and Su (2016) proposed a 2-D PM strategy for items sold with a 2-D warranty, under which the items are preventively maintained every K units of age or every L units of usage, whichever occurs first.

Nevertheless, one of the main drawbacks of the studies above is that they implicitly assume that the PM actions within the warranty period are punctual. In practice, however, the unpunctuality of PM activities that results from the separate nature of PM specification and implementation, is quite common in a B2C (business-to-customer) context (He et al. 2017a). For instance, in an automobile warranty and maintenance setting, it is not uncommon that some automobile owners do not follow the prescribed PM schedule strictly. In the literature, the optimal planning of unpunctual perfect PM (replacement) policies has aroused the interests from a few researchers. He et al. (2017a) studied the optimal age replacement policy with and without minimal repair in anticipation of maintenance unpunctuality. Besides, Zhao et al. (2017) discussed a random age replacement policy in which the unit is replaced before failure at a random point of time.

To our knowledge, there is currently little research dealing with unpunctual imperfect PM strategy in the context of warranty. This paper proposes an unpunctual imperfect PM strategy for repairable items sold with a 2-D warranty, which allows customers to advance or postpone the scheduled PM activities in tolerable ranges. Under this unpunctual PM strategy, the j th PM action should be performed at age $jK \pm \Delta K$, or at usage $jL \pm \Delta L$, whichever occurs first (Note that, K and L are scheduled age- and usage-based PM intervals, respectively). The remainder of the article is organized as follows. Section 2 develops the optimization model for determining the optimal PM schedules under the unpunctual PM strategy. Section 3 conducts a case study to demonstrate the proposed unpunctual PM strategy. Finally, section 4 concludes this paper.

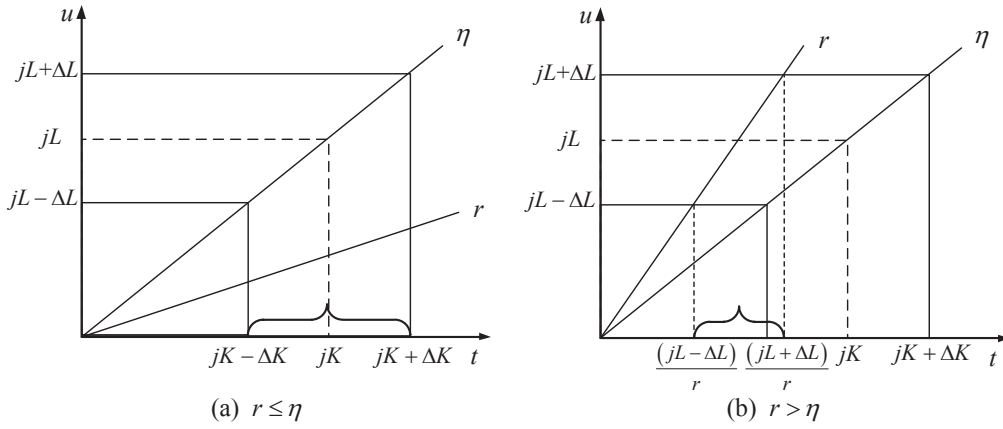


Figure1. Illustration of the tolerable range of the j th unpunctual PM activity.

2. Model formulation

2.1. Two-dimensional warranty and failure modeling

Consider that a repairable item is sold with a 2-D warranty policy, which is characterized by a rectangular region $\Omega_R(W, U) = \{(t, u) : t \in (0, W) \text{ and } u \in (0, U)\}$. Within the warranty region, all item failures are rectified by the manufacturer at no cost to the customer. This free repair warranty expires when either the item age exceeds its limit W or the total usage reaches its limit U , whichever occurs first.

In this study, we look at a repairable item as a black-box system, and use the marginal approach to model its failures in terms of age and cumulative usage (Wang and Xie 2018). For an elapsed time t , the cumulative operating time and total usage of the item can be denoted by $T(t)$ and $U(t)$, respectively. For simplicity, we assume that the item's cumulative operating time equals its age, i.e., $T(t) = t$. An essential assumption adopted here is that different customers have different usage rates but they will not alter their own usage rates during the warranty period (Ye et al. 2013). As a result, the usage rate R can be modelled as a random variable with distribution function $G(r) = P(R \leq r)$, $0 \leq r < \infty$. Conditional on $R = r$, the total usage $U(t)$ of an item at age t can be given by $U(t) = rt$.

Item failures can thus be modeled by a point process with a conditional intensity function $\lambda(t|r)$ that is dependent on both age t and usage $U(t)$. Given that all failures are minimally repaired with negligible repair time and no PM is performed, failures over time occur according to a non-homogeneous Poisson process (NHPP) with conditional intensity function $\lambda(t|r)$. In this study, following Murthy and Wilson (1991), $\lambda(t|r)$ is modeled as a simple polynomial function:

$$\lambda(t|r) = \theta_0 + \theta_1 r + \theta_2 T(t) + \theta_3 U(t) = \theta_0 + \theta_1 r + (\theta_2 + \theta_3 r)t \quad (1)$$

where θ_0 , θ_1 , θ_2 , and θ_3 are positive constants.

2.2. Punctual 2-D PM strategy

In this study, what causes the actual PM instants to deviate from the scheduled instants is the potential unpunctuality of customers returning their items for PM activities in a B2C context. The proposed unpunctual PM strategy allows the earliness or tardiness of customers performing each (originally periodical) PM activity, and the tolerable ranges are $\pm\Delta K$ and $\pm\Delta L$ in the two dimensions of age and usage, respectively. In other words, under the unpunctual PM strategy, the j th PM activity should be performed at age $jK \pm \Delta K$ or at usage $jL \pm \Delta L$, whichever occurs first. Here, we assume that $K = W/(n+1)$ and $L = U/(n+1)$, which is common in practice. For the sake of simplicity, we further assume that $\Delta L/\Delta K = L/K = \eta$. It should be noted that when $\Delta K = \Delta L = 0$, the unpunctual PM strategy reduces to its punctual counterpart.

Figure 1 illustrates the age-based tolerable range for the j th unpunctual PM activity. As can be seen, for a customer with usage rate $r \leq \eta$, the j th PM activity should be performed within the time interval $[jK - \Delta K, jK + \Delta K]$; while for a customer with usage rate $r > \eta$, the j th PM activity should be carried out within the time interval $[(jL - \Delta L)/r, (jL + \Delta L)/r]$.

Under the unpunctual PM strategy, an important issue is how to characterize the maintenance unpunctuality, as different customers may have different unpunctual behaviors. Since the customers' unpunctual behaviors are random, this study models the j th age-based PM instant τ_j for all customers as a uniform random variable defined in a finite support. This assumption is acceptable because 1) it has been adopted in He et al. (2017) and Zhao et al. (2017); and 2) the uniform distribution is often employed for analytical modeling of an unobservable parameter (Kurata and Nam 2013). Nevertheless, an appropriate distribution can be selected and estimated if real datasets are available.

2.3. The optimization model

In this study, the effect of imperfect PM activities on the item reliability is described by the age reduction model in Kim et al. (2004). The amount of age reduction depends on the PM effort level (m) applied. Larger value of m corresponds to greater PM effort and more reliability improvement. Such PM effort is assumed to be imperfect and constrained so that $0 \leq m \leq M$, with $m = 0$ implying no PM action and M indicating the upper limit of the PM effort.

In order to analyze the expected warranty expenses under the unpunctual PM strategy, two cases should be considered, i.e., $r \leq \eta$ and $r > \eta$.

Case 1: $r \leq \eta$. In this case, for a given customer with usage rate $r \leq \eta$, the 2-D warranty period would terminate at time W , with the corresponding usage being rW . Also, the j th age-based PM instant τ_j is a random variable with conditional probability density function (pdf) $z(\tau_j | r \leq \eta) = 1/(2\Delta K)$, $\tau_j \in [jK - \Delta K, jK + \Delta K]$.

Conditioning on the j th PM instant τ_j , the virtual age after the j th PM activity can be obtained as (Kim et al. 2004)

$$v_j = v_{j-1} + \delta(m)(\tau_j - \tau_{j-1}) = \delta(m)\tau_j, j = 1, 2, \dots, n, \quad (2)$$

with $\tau_0 = 0$ and $\tau_{n+1} = W$.

Note that, $\delta(m)$ is a decreasing function of m with $\delta(0) = 1$ and $\delta(M) = 0$. If $m = M$, then the damage during the $(j-1)$ th and the j th PM activities is fully compensated, i.e., $v_j = v_{j-1}$; if $m = 0$, then no PM action is taken, i.e., $v_j = \tau_j$; while in a more general case $0 < m < M$, the item is partially restored, i.e., the PM activity is imperfect. Here, we assume that $\delta(m) = (1+m)e^{-m}$, as in Kim et al. (2004).

We assume that any failure within the warranty period will be rectified through minimal repair with negligible duration, under which the failure rate after repair is nearly the same as that immediately before failure. Hence, the failure process between two successive PM activities is also an NHPP. Then, the conditional expected number of item failures within the warranty period

$$E[N^{\mathcal{W}}(n, m) | r \leq \eta; \tau_1, \dots, \tau_n] = \int_0^{\tau_1} \lambda(t|r)dt + \sum_{j=1}^{n-1} \int_{\delta(m)\tau_j}^{\delta(m)\tau_j + \tau_{j+1} - \tau_j} \lambda(t|r)dt + \int_{\delta(m)\tau_n}^{\delta(m)\tau_n + W - \tau_n} \lambda(t|r)dt \quad (3)$$

Removing the conditioning on τ_j , the conditional number of failures within the warranty period becomes

$$E[N^{\mathcal{W}}(n, m) | r \leq \eta] = \int_{nK - \Delta K}^{nK + \Delta K} \dots \int_{K - \Delta K}^{K + \Delta K} E[N^{\mathcal{W}}(n, m) | r \leq \eta; \tau_1, \dots, \tau_n] \prod_{j=1}^n z(\tau_j | r \leq \eta) d\tau_1 \dots d\tau_n \quad (4)$$

where $z(\tau_j | r \leq \eta) = 1/(2\Delta K)$ and $E[N(n, m) | r \leq \eta; \tau_1, \dots, \tau_n]$ is given by (3).

Case 2: $r > \eta$. For a given customer with usage rate $r > \eta$, the 2-D warranty period would terminate at time U/r , when the corresponding usage reaches its limit U . Moreover, the j th age-based PM instant τ_j is a random variable with conditional pdf $z(\tau_j | r > \eta) = r/(2\Delta L)$, $\tau_j \in [(jL - \Delta L)/r, (jL + \Delta L)/r]$.

Conditioning on the j th PM instant τ_j , the virtual age after the j th PM activity is given by $v_j = \delta(m)\tau_j$, $j = 1, 2, \dots, n$, with $\tau_0 = 0$ and $\tau_{n+1} = U/r$. Then, the conditional number of item failures during the warranty period is

$$E[N^{\mathcal{W}}(n, m) | r > \eta; \tau_1, \dots, \tau_n] = \int_0^{\tau_1} \lambda(t|r)dt + \sum_{j=1}^{n-1} \int_{\delta(m)\tau_j}^{\delta(m)\tau_j + \tau_{j+1} - \tau_j} \lambda(t|r)dt + \int_{\delta(m)\tau_n}^{\delta(m)\tau_n + \frac{U}{r} - \tau_n} \lambda(t|r)dt \quad (5)$$

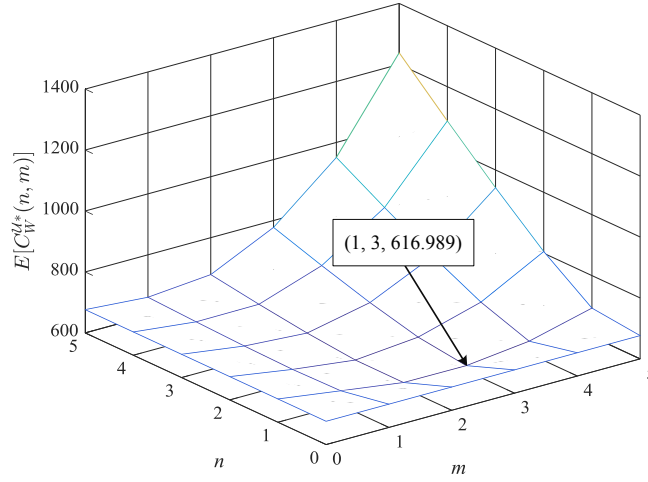


Figure2. Illustration of the existence of the optimal PM strategy.

Similarly, removing the conditioning on τ_j , the conditional number of item failures during the warranty period becomes

$$E[N^{\mathcal{U}}(n, m) | r > \eta] = \int_{\frac{nL-\Delta L}{r}}^{\frac{nL+\Delta L}{r}} \cdots \int_{\frac{L-\Delta L}{r}}^{\frac{L+\Delta L}{r}} E[N^{\mathcal{U}}(n, m) | r > \eta; \tau_1, \dots, \tau_n] \prod_{j=1}^n z(\tau_j | r > \eta) d\tau_1 \dots d\tau_n \quad (6)$$

where $z(\tau_j | r > \eta) = r/(2\Delta L)$ and $E[N(n, m) | r > \eta; \tau_1, \dots, \tau_n]$ is given by (5).

Finally, removing the conditioning on R , the total expected warranty servicing cost under the unpunctual PM strategy becomes

$$E[C_W^{\mathcal{U}}(n, m)] = c_f \int_0^{\eta} E[N^{\mathcal{U}}(n, m) | r \leq \eta] g(r) dr + c_f \int_{\eta}^{\infty} E[N^{\mathcal{U}}(n, m) | r > \eta] g(r) dr + nC_p(m) \quad (7)$$

where c_f is the expected cost of a minimal repair, and $C_p(m)$ is the expected PM cost with level m . Generally, $C_p(m)$ is an increasing function of m , $0 \leq m \leq M$.

When the conditional failure rate $\lambda(t|r)$ is given by (1), then (6) can be expressed as:

$$E[C_W^{\mathcal{U}}(n, m)] = c_f \left\{ \left[\left((1+n\delta(m)) \frac{W^2}{2(n+1)} + \frac{1}{3} (1-\delta(m)) n(\Delta K)^2 \right) \times \right. \right. \\ \left. \left[\theta_2 G(\eta) + \theta_3 \int_0^{\eta} r g(r) dr \right] + \theta_0 W G(\eta) + \theta_1 W \int_0^{\eta} r g(r) dr \right] \\ + c_f \left\{ \left[\left((1+n\delta(m)) \frac{U^2}{2(n+1)} + \frac{1}{3} (1-\delta(m)) n(\Delta L)^2 \right) \times \right. \right. \\ \left. \left[\theta_3 \int_{\eta}^{\infty} \frac{1}{r} g(r) dr + \theta_2 \int_{\eta}^{\infty} \frac{1}{r^2} g(r) dr \right] + \theta_1 U \bar{G}(\eta) + \theta_0 U \int_{\eta}^{\infty} \frac{1}{r} g(r) dr \right] \right\} + nC_p(m) \quad (8)$$

From (8), it is easily known that when $\Delta K = 0$, $E[C_W^{\mathcal{U}}(n, m)] = E[C_W^{\mathcal{P}}(n, m)]$, where $E[C_W^{\mathcal{P}}(n, m)]$ represents the expected warranty cost under the punctual PM strategy.

Under both PM strategies, the optimization problem is to identify the optimal PM parameters, i.e., $n^{\omega*}$ and $m^{\omega*}$, over the warranty period to minimize $E[C_W^{\omega}(n, m)]$, where $\omega = \{\mathcal{U}, \mathcal{P}\}$. Mathematically, it can be expressed as follows:

$$\min E[C_W^{\omega}(n, m)] \text{ s.t. } 0 \leq m^{\omega} \leq M, n^{\omega} \in \mathbb{Z}^+ \cup \{0\}, \omega = \{\mathcal{U}, \mathcal{P}\} \quad (9)$$

3. Case study

In this section, a case study is given to illustrate the implementation of the proposed unpunctual PM strategy. Consider a Chinese truck manufacturer whose trucks are sold with a 2-D warranty of 3 years and 400000 km (He et al. 2017b). Based on warranty data analysis, the usage rates r (in 100000 km/year) of 2016-mode

Table 1. Sensitivity analysis

c_f	Punctual PM				Unpunctual PM											
					$\Delta K = 1$ weeks				$\Delta K = 2$ weeks				$\Delta K = 4$ weeks			
	$n^{\mathcal{P}*}$	$m^{\mathcal{P}*}$	$E[C^{\mathcal{P}*}]$		$n^{\mathcal{U}*}$	$m^{\mathcal{U}*}$	$E[C^{\mathcal{U}*}]$	Ψ	$n^{\mathcal{U}*}$	$m^{\mathcal{U}*}$	$E[C^{\mathcal{U}*}]$	Ψ	$n^{\mathcal{U}*}$	$m^{\mathcal{U}*}$	$E[C^{\mathcal{U}*}]$	Ψ
50	1	2	339.373		1	2	339.375	0.01‰	1	2	339.383	0.03‰	1	2	339.414	0.12‰
100	1	3	646.277		1	3	646.284	0.01‰	1	3	646.305	0.04‰	1	3	646.387	0.17‰
150	2	3	936.574		2	3	936.595	0.02‰	2	3	936.657	0.09‰	2	3	936.904	0.35‰
200	2	3	1208.765		2	3	1208.793	0.02‰	2	3	1208.875	0.09‰	2	3	1209.206	0.36‰
250	2	3	1480.957		2	3	1480.991	0.02‰	2	3	1481.094	0.09‰	2	3	1481.507	0.37‰
300	3	3	1750.306		3	3	1750.368	0.04‰	3	3	1750.554	0.14‰	3	3	1751.298	0.57‰
350	3	3	2012.024		3	3	2012.096	0.04‰	3	3	2012.313	0.14‰	3	3	2013.181	0.57‰
400	3	3	2273.741		3	3	2273.824	0.04‰	3	3	2274.072	0.15‰	3	3	2275.064	0.58‰
450	3	3	2535.459		3	3	2535.552	0.04‰	3	3	2535.831	0.15‰	3	3	2536.947	0.59‰
500	3	4	2790.654		3	4	2790.771	0.04‰	3	4	2791.122	0.17‰	3	4	2792.528	0.67‰

trucks follow a Weibull distribution with scale parameter $\alpha = 1.5259$ and shape parameter $\beta = 2.1062$. The minimum and maximum usage rates are 0.0365 and 5.65, respectively, which are directly obtained through the warranty claims data. Moreover, the parameters in the failure rate function (1) are estimated as $\theta_0 = 0.9558$, $\theta_1 = 0.4603$, $\theta_2 = 0.3427$ and $\theta_3 = 0.6085$. Also, the average repair cost of each failure is approximately \$95. Furthermore, we assume that the PM cost is $C_p(0) = \$0$, $C_p(1) = \$10$, $C_p(2) = \$30$, $C_p(3) = \$60$, $C_p(4) = \$100$ and $C_p(5) = \$160$, as in Kim et al. (2004). We further assume that the manufacturer pre-specifies $\Delta K = 2$ weeks (i.e., 1/26 year) for the unpunctual PM strategy. This means that customers are entitled to have a 4-week tolerable range of maintenance unpunctuality.

Here, simple numerical search approach is adopted to determine the optimal PM parameters, since the number of PM actions, n , and the corresponding PM level, m , are integers. Based on the parameters above, the optimal number and level of PM actions under the unpunctual strategy are solved as $n^{\mathcal{U}*} = 1$ and $m^{\mathcal{U}*} = 3$, respectively, and the corresponding expected warranty cost is $E[C_W^{\mathcal{U}*}] = \$616.989$; while under its punctual counterpart carrying out $n^{\mathcal{P}*} = 1$ PM actions with level $m^{\mathcal{P}*} = 3$ is also the optimal strategy, and the resulting warranty cost is expected to be $E[C_W^{\mathcal{P}*}] = \$616.963$, which is slightly lower than \$616.989. This shows that the unpunctual PM strategy will lead to lightly higher warranty cost than the punctual strategy in this case. To obtain more insights, we perform a sensitivity analysis with respect to c_f and ΔK , as shown in Table 1. Note that, $\Psi = (E[C_W^{\mathcal{U}*}] - E[C_W^{\mathcal{P}*}]) / E[C_W^{\mathcal{P}*}]$ represents the percentage growth of warranty cost of the unpunctual PM strategy in comparison to its punctual counterpart.

As can be seen, for both PM strategies, as the minimal repair cost c_f increases, the optimal number of PM actions, the optimal PM degree, and the total expected warranty costs increase, as to be expected. Moreover, under the unpunctual PM strategy, the total expected warranty cost increases as ΔK increases. This is reasonable since more serious unpunctuality behaviors will result in more warranty servicing cost. Furthermore, it is interesting to observe that the warranty cost growth of the unpunctual PM strategy is very small (less than 1‰) in all cases. This implies that the unpunctual PM strategy may be of practical interest to the manufacturers given its potential of boosting future sales by increasing customer satisfaction. Actually, this unpunctual PM strategy has been adopted by a Chinese automaker (SAIC-GM-Wuling Automobile Company) in their warranty and maintenance practice.

4. Conclusions

This paper investigates an unpunctual PM strategy for repairable items sold with a 2-D warranty, under which the j th PM activity should be performed at age $jK \pm \Delta K$ or at usage $jL \pm \Delta L$, whichever occurs first. The unique characteristic of the proposed unpunctual PM strategy is that it allows customers to advance or postpone the scheduled PM activities in tolerable ranges, which has great potential to increase the customer satisfaction. Case study and sensitivity analysis show that applying the unpunctual PM strategy leads to slightly higher warranty cost than the punctual strategy. Nevertheless, the increase of warranty cost due to the introduction of unpunctual PM strategy is not significant in most cases. In view of its possibility of boosting future sales by increasing customer satisfaction, it is therefore of potential interest for manufacturers to adopt this PM strategy.

In this study, the optimal PM parameters and the corresponding warranty cost under the unpunctual PM strategy is almost the same as those under the punctual strategy. This may be due to our assumption of uniform distribution of unpunctual PM instants τ_i , $j = 1, 2, \dots, n$. Investigating the impact of other distributions on the optimal PM parameters and the corresponding warranty cost will be our future research topic.

References

- He K, Maillart LM, Prokopyev OA (2017a) Optimal planning of unpunctual preventive maintenance. *IIE Transactions*, 49, 127–143.
- He S, Zhang Z, Zhang G, He Z (2017b) Two-dimensional base warranty design based on a new demand function considering heterogeneous usage rate. *International Journal of Production Research*, 55, 7058–7072.
- Huang Y-S, Huang Ch-D, Ho J-W (2017) A customized two-dimensional extended warranty with preventive maintenance. *European Journal of Operational Research*, 257, 971–978.
- Kim CS, Djamaludin I, Murthy DNP (2004) Warranty and discrete preventive maintenance. *Reliability Engineering & System Safety*, 84, 301–309.
- Kurata H, Nam S-H (2013) After-sales service competition in a supply chain: Does uncertainty affect the conflict between profit maximization and customer satisfaction?. *International journal of production economics*, 144, 268–280.
- Murthy DNP, Wilson RJ (1991) Modelling two-dimensional warranties. *Proceedings of the Fifth International Symposium on Applied Stochastic Models and Data Analysis*, pp.481–492.
- Su C, Wang X (2016) A two-stage preventive maintenance optimization model incorporating two-dimensional extended warranty. *Reliability Engineering & System Safety*, 155, 169–178.
- Wang X, Su C (2016) A two-dimensional preventive maintenance strategy for items sold with warranty. *International Journal of Production Research*, 54, 5901–5915.
- Wang X, Xie W (2018) Two-dimensional warranty: A literature review. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, DOI: 1748006X17742776.
- Wang Y, Liu Z, Liu Y (2015) Optimal preventive maintenance strategy for repairable items under two-dimensional warranty. *Reliability Engineering & System Safety*, 142, 326–333.
- Ye Z-S, Murthy DNP, Xie M, Tang L-C (2013) Optimal burn-in for repairable products sold with a two-dimensional warranty. *IIE transactions*, 45, 164–176.
- Zhao X, Al-Khalifa KN, Hamouda AM, Nakagawa T (2017) Age replacement models: A summary with new perspectives and methods. *Reliability Engineering & System Safety*, 161, 95–105.

Decentralisation of reliability data through blockchain

G. Willoughby*, T. Ingram**, T. Byrne***, D. Smith, N Rahman

* Brew House, Wilderspool Business Park, Warrington, WA4 6HL: gwilloughby@reliabilityblockchain.com

** Brew House, Wilderspool Business Park, Warrington, WA4 6HL: tingram@mmeengineering.com

*** Apollo House, Eboracum Way, York, YO31 7RE: tbyrne@geosyntec.com

Abstract. The intent of this paper is to illustrate how the current paradigm of reliability information held by a small selection of companies who sell the information could soon be outdated. We believe that Reliability data ownership could be transferred by a collaborative, decentralized approach, with data sourced from the communities that use the data stored in a blockchain. The advent of blockchain as a shared, immutable ledger, has the potential to be the technology that provides the platform for this. Such a ledger can provide a new way of thinking about reliability data potential: Original Equipment Manufacturers (OEM's) and users contributing data in a transparent manner, clear visibility of underpinning data continually evolving in a live database, self-checking across the community, data beyond the base of the bathtub, quality certification, storage of source information such as Failure Modes Effects Criticality Analysis (FMECAs), direct linkage into the Industrial Internet of Things. This fundamental shift in approach could see a focus of practitioner held information, with trust held within the community. We will present some potential methodologies, and how we are proposing to establish collaborative working groups to establish a standard on this technology. The objective is to provide a platform for the international reliability community that allows for the development of a truly free, trustworthy, reliability resource.

1. Background

The intent of this paper is to present:

- Current Reliability data: considering approaches of how data is used; data sources; and the resulting issues and benefits of these approaches;
- Future of decentralised reliability data repository: presenting a description of decentralisation; benefits of decentralisation; and tools to support decentralisation;
- Implementation of decentralised repository: considerations for implementation of a block chain; challenges of implementation; a forward-looking roadmap from conception;
- Future potential of a decentralized approach: presenting the opportunities from a new way of thinking
- Conclusions: presenting recommendations for the next steps.

2. Current Reliability data

2.1 How data is used

Data is collected for a variety of uses but is most often used to prove or evidence something. OEMs might use data to advertise the capability of their products. End users might use data to evidence the performance of their equipment to regulators or data might be used internally to justify investment or aid decision making. Whatever the reason, reliability data has an integral part to play in almost all elements of the asset life cycle.

This reliability data is typically used to underpin frequency of failures and resulting events, whether that be Availability Analysis, Quantitative Risk Assessment (QRA), Failure Modes Effects Analysis, Failure Analysis, Spares Ranging & Scaling, Multi Criteria Decision Analysis, Accelerated Life Testing, with many more potential uses available. The output of such analyses focusses on a single point result when the reality of operations is much more complicated.

Reliability of equipment is typically categorized into 3 periods: infant mortality in which issues that have not been identified during testing and commissioning result in a high initial failure rate which decreases over time; operational life in which the failure rate remains reasonably constant, due to increased duration and increased data population; end of life phase results in an increasing failure rate as the component approaches or exceeds its intended life. The majority of data sources focus on the portion of "useful life" which is

represented by a constant failure rate, or simplified wear out patterns characterised through distributions such as Weibull.

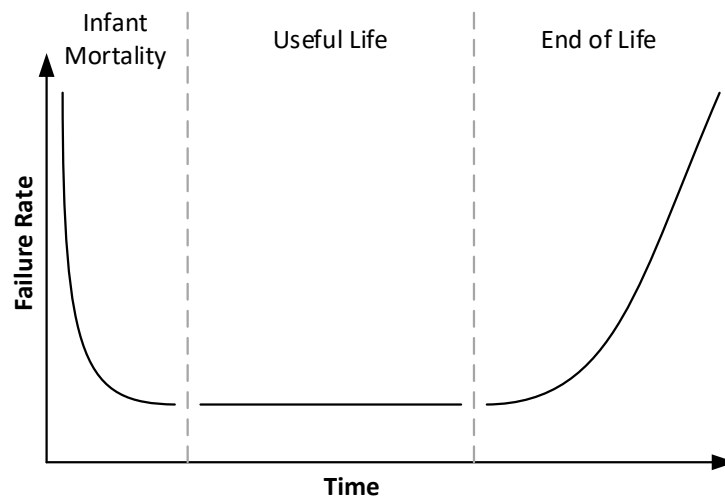


Figure 1. Example “Bathtub curve” with split into 3 typical life stages.

However, many organisations are planning to operate well beyond their design life to strike a careful balance of cost and performance to maintain cost effective production. In order to support accurate predictions operators need a significant amount of intelligence to characterise operational life, and usage for equipment that may be lost 50 years on.

2.2 Data sources

Across industries, approaches differ slightly, but it is typical for data sources to be prioritized, for example: internal operational data; regulatory compiled data (e.g. United Kingdom Health & Safety Executive (HSE)); government sponsored data (e.g. Military Handbooks (Mil-HDK)); group sponsored data (e.g. Non-electronic Parts Reliability Data (NPRD)); sector specific data (e.g. Off-shore and onshore Reliability Data (OREDA)).

Organizations constantly look internally to collect and use data in a more effective way, considering the range and scale of the data they collect through a combination of manual and automated approaches. Such data is directly applicable to the application, and the organization has full visibility of the supporting information. However, internal sources are limited by the available equipment, knowledge of failures, internal confirmation bias, small dataset.

External data sources, such as those mentioned above, provide information from typically a much larger data set, with statistically more confident results. However, the source of the data can raise issues regarding trust and applicability to a given application. For example, government sponsored data such as Mil-Hdbk217F, was last updated in 1995 and data within could be considered out of date when compared to current equipment. Safety related data such as Failure Rate and Event Data for use within Risk Assessments, HSE Planning Case Assessment Guide (PCAG) Chapter 6K (28/06/2012), provides a consistent data source for safety assessments, however its use to underpin operational availability may be questioned as a perception may be that certain failures may reduce operational availability without impacting safety. Paid for standards such as Nonelectronic Parts Reliability Data Publication (NPRD-2016), represent a more varied and independent source. However, the paid for mechanism can also provide uncertainty regarding the validity of the data as new databooks are released, and the provenance of the data can be limited. Sector specific data such as OREDA Handbook 2015, 6th edition – Volume I and II provide useful sector specific data, but may not provide sufficient information or context to enable translation into other sectors or applications.

2.3 Issues and benefits

The referenced data sets provide an average mean time between failure, which then needs to be reflected against the condition of in-service assets in-service. Reliability professionals are experts in translation of the

current asset performance, or design scope, into a prediction of future performance. As such, within the reliability domain, the confidence level of the data is understood. However, as the resulting analyses are used to underpin wider business decisions the understanding of the supporting data can be lost. Inherent weaknesses within the sources such as, dataset size, validity against current & existing equipment, cross sector applicability, and underpinning resolution can limit the confidence and trust outside of the reliability community.

Even if we view these data sources as high integrity, the use of the data may be different to the source, for example risk of a complete failure vs a partial degradation are likely to have different failure rates. Publications such as Failure Mode/Mechanism Distributions FMD-2016 fill some of the gaps through articulation of the distribution of failure modes within the data set. However, there can be a question on the integrity of the data and whether that the usage/environment can translate.

Within the right context each of the data sources mentioned in Section 2.2 can be a valuable asset to support the reliability predictions.

3. Future of decentralised reliability data repository

3.1 What is decentralisation of data?

A comparable example of data decentralisation is the evolution of encyclopedias, which once were limited to printed books, moving onto electronic media. Such publications were a “trusted source” because there was no other information available and had inherent invested value as they were paid for publications. The notion and validity of a crowdsourced data set within the 1990’s would have raised that decentralising such information would result in a lack of trust that may render the data useless. Moving forward to the present day, Wikipedia has made that transition, and disrupted the paid for model, but it is not the only data source. Also the perception of the data is to utilize it with caution, rather than with implicit confidence. Since its creation a significantly information resource has developed within www.wikipedia.com sourced from volunteers with a reasonable level of accuracy.

In a similar manner we propose the transition of reliability data into a shared data set, that self regulates, and provides sufficient granularity, traceability and lineage to make it useful to a wide array of applications and end users.

3.2 Benefits of decentralisation

What if we could work together to produce much larger reliability data sets, ones that not only represented the “as is” state for an individual organisation but a data set that could help to evidence the whole spectrum of performance across industries and applications? The potential to work collaboratively and drive up standards across the board is hard to dispute and has potential benefits for all stakeholders involved in the reliability data domain:

- OEM’s – visibility of how their products perform long term in-service, and how they perform against competitors and see how end users are using their products. This will focus investment in value add research & development and develop better products long term;
- End users – informing equipment procurement decisions, benchmarking of their performance within their sector, and beyond. They can use information to help drive investments in new assets, and potentially benefit from best practice and innovation from other sectors;
- Regulators – provision of a simplified holistic view across relevant sectors which can help organisations to drive up performance in targeted areas.

Expectations of customers, users and regulators are increasing all the time. This results in a need to drive up standards, reduce risk and protect profits. Central to this is the ability to make good decisions based on good data. “Lessons learnt” or “best practice” are terms often used to emphasise the virtues of effective information sharing but is something that is often difficult to achieve even on a small scale because of some of the perceived difficulties in this process. What if we could embrace new technologies that would allow us to share and utilise reliability data in a shared, common, de-centralised network; not for profit, simply to improve asset management performance.

3.3 Tools to support decentralisation;

Numerous data storage types exist, and all have varying levels of complexity. A key factor in developing a decentralised set of data relating to equipment failure will be trust and visibility, due to many contributors in many applications and industry sectors. The trust needs to be formed both through an ability to summarise data providing visibility of the supporting data sources, and fundamentally that no single entity controls the whole data set. On this basis we propose that a blockchain infrastructure could be used to provide a decentralised, secure, and reliable platform.

Experience of conventional repository technologies to date have demonstrate an almost Darwinian evolution, as software is upgraded, copied, changed etc, the data within can suffer random mutations. This can be in the form of corrupt, or missing data, which is typically picked up during or after analysis and can costly or devastating results. The advent of blockchain demonstrates an evolutionary leap, which to date through cryptography, and hash values has been robust, in part through inherent error checking from block to block. Also, if a mutation occurs this is stabilised through the peer to peer nature of the system, and the mutation can be readily removed. As technologies improve and cryptography is cracked, this position may change, however there is sufficient financial backing through the likes of Bitcoin to keep investment in cryptography ahead of the game.

4. Implementation of decentralised repository

4.1 Considerations for implementation of a block chain

In its simplest form blockchain acts as a large database of information from which data can be traced, verified and stored securely. A blockchain is shared across a peer-to-peer network in which multiple versions of the same data are stored across the network, which adds resilience and security.

The technology itself prevents changes from being made by stringing together a series of transactions into a single chain where each record is linked to the previous record using a hash function of the value of the previous and current transaction, a form of digital signature. If changes are made the chain is broken as the hash will change and not be recognised by any other part of the blockchain. This immutable ledger gives credence to the validity of the record of the change or sharing of data and could be used, for example in the energy sector for a peer-to-peer community-based sharing network, allowing for community auditing and transparent functionality of the industry.

Understanding the concept of decentralization is paramount to understanding the implementation of a blockchain network. Several questions must be answered by an organisation or community of people to decide what type of blockchain to implement. Public blockchains usually consist of a large group of nodes (computers) which are connected to the internet to validate transactions ongoing within the blockchain. As a result of this, consensus transactions are agreed by other ways rather than selective endorsement IE, Mining - computers on the network try to solve a complex cryptographic problem to create a proof of work. In contrast private blockchains can be set to only permit known organisations or users to join, enabling the grounds for a permissioned network in which consensus can be achieved through a process called selective endorsement, where known or selected users can verify changes in data. In essence, private blockchains use identity to confirm membership and access privileges so that participants in the network know who they are dealing with, allowing for enhanced auditing and transparency as noted above.

4.2 Challenges of implementation

Information sharing across organisations is notoriously difficult, due to the invested time and money required to collect, verify, manage and analyse data. The inherent requirement to manipulate data into a form that can be shared across electronic platforms causes additional time and financial burden. This is further emphasized when referring to failure related data, as there is additional complexity due to mis representation of failures outside of the organization. However the benefit to all involved (if a common platform can be created) is to drive OEM's to better products, provide more opportunities for informed decision making, ability to truly learn across sectors with trust in a dataset which is large enough to provide statistical confidence in results.

We believe the significantly increase data repository and granularity would outweigh these issues for a significant number of organisations.

A key challenge will be a common understanding of what information (meta data) would be captured in support of the reliability data. For example something as simple as a flag to identify a failure rate that has been influenced in some way by human factors, could provide a new insight into the influence on humans on certain equipment types or industry sectors. This will require a significant amount of cooperation across multiple sectors to develop a set of definitions that work across industries.

4.3 Roadmap

Implementing the blockchain technology requires several steps to be followed as shown below;

- Step 1 – Identify the blockchains purpose;
- Step 2 – Agree on the most suitable consensus mechanism;
- Step 3 – Identify a suitable blockchain platform;
- Step 4 – Understanding the scope of nodes associated with the blockchain;
- Step 5 – Design the blockchain Instance and parameters;
- Step 6 – Build the API's;
- Step 7 - Design the admin and user interface;
- Step 8 – Adding future technology.

5. Future potential of a decentralized approach

There are countless potential benefits of using a blockchain as a decentralised, single source of data for individuals, companies and non-profit organisations operating in multiple industries. The fundamental properties of a blockchain's infrastructure, such as its limitless storage capacity and secure decentralised foundations, make it an ideal solution for the storage of data in a live and manageable repository which can allow for improved data manipulation for the needs of all users.

5.1 Operations Staff and Asset Managers

Operations personnel could be some of the key users of a blockchain based data repository. Access to a blockchain allows operations staff to benchmark and compare performance data internally with their own records, or externally with the data of similar organisations. Comparisons can be based on numerous metrics such as defect rates, production rates, human interaction and other KPI's which will allow for the process optimisation of an organisations operational activities.

5.2 Procurement and Investment Planning

As mentioned above, blockchains have the capability to categorise data from a practically infinite number of decentralised locations allowing for effective data comparison. This could be utilised by purchasers to complete a full market comparison and refine their search criteria to find the best suppliers in the market almost instantaneously.

Currently, many organisations struggle to find suppliers which meet their procurement standards and requirements and invest heavily in new market research strategies. If supplier information was held in a blockchain, one could readily complete a full market analysis while ensuring all the data remains secure in decentralised locations.

Blockchain based market research has the potential to save huge amounts of time and effort for procurement functions, being a means for much easier and effective market research and supply chain management. Thus, companies may choose to refresh their procurement strategies on a more regular basis, since the market will be much more transparent allowing purchasers to refine their demands based on the current market offering.

5.3 Research and Development

Data comparison using the blockchain could be of significant benefit for R&D organisations when searching for new market opportunities for the development of future products and services. Performance data, asset

lifecycle information, new and existing government regulations and policies which are all accessible via one user friendly portal (the blockchain). This could provide R&D organisations with all the necessary information to identify, foresee and exploit opportunities for growth in existing markets. In turn, this could facilitate a rapid growth in research and development and further bridge the gap between R&D and Industry.

5.4 Regulatory Bodies & Government Departments

Regulatory bodies could also benefit from having access to industry data via the blockchain. For example, regulatory bodies assessing the general health of equipment used for manufacturing could use the blockchain to gain access to relevant records of the production performance, maintenance history and other statistical information in order to make more informed decisions. Furthermore, regulatory bodies could use this information as a basis to form future regulations based on current or predictable future transitions in industry.

5.5 Software & hardware developers

Potentially, company hardware and software could directly interphase with the blockchain, allowing for automated transfer of data directly to the blockchain. This will enhance user experience ensuring that all relevant information is up to date, live and ready for use.

6. Conclusions:

This paper has presented some basic concepts of how reliability data is currently sourced by a small selection of companies who sell the information, and how this paradigm could soon be out dated. We believe that reliability data ownership could be transferred into a collaborative, decentralized approach, sourced from the communities that use the data stored on a blockchain. The advent of blockchain as a shared, immutable ledger, has the potential to be the technology that provides the platform for this.

This fundamental shift in approach could see a focus of practitioner held information, with trust held within the community. We will present some potential methodologies, and how we are proposing to establish collaborative working groups to establish a standard on this technology. To provide a platform for the international reliability community to develop a truly free, trustworthy, reliability resource.

To find out more, register interest or become part of the influencing working groups visit www.reliabilityblockchain.com.

References

- HSE Planning Case Assessment Guide (PCAG) Chapter 6K (28/06/2012)
- Mil-Hdbk217F (1995) Military Handbook: Reliability Prediction of Electronic Equipment
- NPRD (2016) Nonelectronic Parts Reliability Data Publication
- OREDA Handbook 2015, 6th edition – Volume I and II
- Failure Mode/Mechanism Distributions FMD-2016
- www.reliabilityblockchain.com

Assurance for sample size determination in binomial reliability demonstration testing

K.J. Wilson, M. Farrow

School of Mathematics, Statistics & Physics, Newcastle University, UK

Abstract Frequently, manufacturers are required to demonstrate that products meet reliability targets. A typical way of doing this is to use reliability demonstration tests (RDTs), in which a number of products are put on test and the test is passed if a critical number or fewer fail. There are various methods for determining the sample size for such tests. Traditionally this was based on the size of a hypothesis test following the RDT. More recently, Bayesian approaches have been proposed based on the idea of risk criteria. However, these approaches do not lead to a single, optimal sample size and conflate the choice of sample size for the test and the analysis to be undertaken once the test has been conducted. In this paper we offer an alternative approach to sample size determination based on the idea of assurance. This approach can overcome each of these issues with risk criteria. Assurance chooses the sample size to answer the question: “What is the probability that the RDT will result in a successful outcome?” We demonstrate the use of assurance for sample size calculations in binomial RDTs and discuss the specification of prior distributions for the design and analysis of the test.

1. Introduction

Hardware products are required to have high levels of reliability, particularly those which provide safety-critical functions within larger systems. It is therefore crucial that the users of such hardware products have confidence in the reliability claimed by the manufacturers. A typical and widely used approach to provide this confidence is through the use of a reliability demonstration test (RDT), in which a number of the hardware products are put on test and the number which fail is observed. If the number of failures is below some pre-defined threshold then the test is said to be passed and the reliability of the product is demonstrated (Elsayed, 2012).

The design of an RDT is therefore given by the number of items put on test and the number of failures which are allowed for the test to be passed. An additional consideration which affects the test plan is the type of hardware product which is of interest. In this paper we consider products which are required to function on demand and can either work or fail, such as a parachute.

Calculating sample sizes for RDTs has been considered at least as far back as the 1960s. A traditional approach was to choose the sample size to give a specified size of a hypothesis test (see, for example, Section 3 of Meeker and Escobar (2004)). Another widely investigated approach is based on the idea of risk criteria, which evaluate the risk to the producer of the product and consumer of the product associated with incorrect conclusions from the RDT (Easterling, 1970; Marz and Waller, 1982; Hamada *et al*, 2008).

Assurance (also known as expected power and average power), has been proposed (Spiegelhalter *et al*, 1994; O’Hagan and Stephens, 2001) as the correct Bayesian method for calculating sample sizes in clinical trials. It is based on the notion that the sample size should be chosen to meet a threshold for the probability that the trial leads to a successful outcome. If the analysis to be conducted following the trial is a frequentist hypothesis test for the superiority of a new treatment over the standard treatment, then the successful outcome would be rejection of the null hypothesis in the test. There is no limitation with assurance that the analysis following the trial be a frequentist test, however.

In this paper we consider the use of assurance for sample size determination for RDTs. We develop an approach based on assurance to select both the number of items to test and the threshold number of failures for a successful RDT. We propose suitable structures for the prior distributions for the unknown model parameters and detail how historical data can be incorporated into the assurance calculation. Following this introduction, we detail our approach to sample size determination using assurance for failure on demand data in Section 2. The following section illustrates the approach developed in the context of an example. The paper concludes with Section 4, which provides a summary of the approach and some areas for further work.

2. Binomial reliability demonstration testing

2.1. Traditional approaches

Suppose that we have n devices which we are to put on test and that Y is the number which will fail the test. Also define π to be the probability that an item survives the test. In reliability demonstration testing, for a given n , we define a maximum allowed number of failures c . If the actual number of failures in the test exceeds c then the test is failed. If not, the test is passed.

Therefore, a reliability demonstration test plan (Hamada *et al*, 2008) is given by the pair (n, c) . The likelihood associated with the test is typically assumed to be binomial

$$Y \mid \pi \sim \text{bin}(n, 1 - \pi). \quad (1)$$

Traditionally, a reliability demonstration test (RDT) would be analysed using a hypothesis test where $H_0 : \pi = \pi_T$, the quantity π_T is the target reliability and $H_1 : \pi > \pi_T$. For a $100(1 - \alpha)\%$ critical level, we would reject H_0 when $\Pr(Y \leq c) \leq \alpha$. That is if $\sum_{y=0}^c \binom{n}{y} (1 - \pi_T)^y \pi_T^{n-y} \leq \alpha$. If we set $c = 0$ then we could choose n to satisfy

$$n \geq \frac{\log(\alpha)}{\log(\pi_T)}. \quad (2)$$

Other approaches have been suggested, often based on the idea of risk criteria. In these approaches, test plans are chosen to keep the probabilities of making the wrong conclusions from the reliability demonstration test small. Classical risk criteria (Tobias and Trindade, 1995), average risk criteria (Easterling, 1970) and (Bayesian) posterior risk criteria (see, for example, Chapter 10 of Hamada *et al* (2008)) have been proposed.

Often in reliability demonstration testing, test plans can involve running large numbers of very reliable items for long periods of time. Meeker and Escobar (2004) proposed using past observations $\mathbf{x}$ in the analysis to reduce n . They named the new tests which incorporated past observations reliability assurance tests (RAT).

2.2. Assurance

If we consider the posterior risk criteria, there is no single optimal (n, c) which results. The approach also conflates two distinct processes: the choice of sample size for the test and the analysis to be undertaken once the test has been conducted. A result of this is that the prior used in the analysis stage is the same as that used in the design stage. This may not be appropriate.

Assurance has been proposed (Spiegelhalter *et al*, 1994; O'Hagan *et al*, 2005; Ren and Oakley, 2013) as the correct Bayesian approach to sample size calculations in medical statistics. In the context of reliability demonstration testing it can overcome each of the potential drawbacks detailed above.

Assurance chooses a sample size based on the answer to the question “what is the probability that the reliability demonstration test is going to result in a successful outcome?”.

The probability of a successful test is given by

$$\Pr[\text{Successful test}] = \int_0^1 f(y \leq c \mid \pi) p(\pi) d\pi \quad (3)$$

$$= \int_0^1 \sum_{y=0}^c \binom{n}{y} (1 - \pi)^y \pi^{n-y} p(\pi) d\pi, \quad (4)$$

where $p(\pi)$ is the prior distribution for π and $f(y \leq c \mid \pi)$ is the cumulative distribution function of y evaluated at c , which in this case is binomial. We would like this probability to be high. We may wish to choose a minimum acceptable value γ .

Let us suppose that we have data from previous tests of the form x_i for $i = 1, \dots, I$ where $x_i \mid \pi_i \sim \text{bin}(n_i, 1 - \pi_i)$, and the probabilities of items surviving the test in each case can be thought of as coming from the same prior distribution $\pi_i \sim \text{beta}(a, b)$, with hyperparameters a, b . For example they could be identical components produced in the same factory being tested at various different locations. Then, let us suppose that our proba-

bility of interest also comes from the same prior distribution $\pi \sim \text{beta}(a, b)$ and that we define a hyper-prior distribution over (a, b) .

The probability of a successful test is now

$$\Pr[\text{Successful test} \mid \mathbf{x}] = \int_0^1 f(y \leq c \mid \pi) p(\pi \mid \mathbf{x}) d\pi \quad (5)$$

$$= \int_0^1 \sum_{y=0}^c \binom{n}{y} (1-\pi)^y \pi^{n-y} p(\pi \mid \mathbf{x}) d\pi. \quad (6)$$

We can sample from the posterior distribution $p(\pi \mid \mathbf{x})$ using Markov Chain Monte Carlo (MCMC) in the following way.

(a) Generate N posterior draws of a, b of the form $a^{(j)}, b^{(j)}$ for $j = 1, \dots, N$.

(b) For $j = 1, \dots, N$ draw $\pi^{(1)}, \dots, \pi^{(N)}$ as

$$\pi^{(j)} \sim \text{beta}(a^{(j)}, b^{(j)}). \quad (7)$$

Using the draws of π from the posterior distribution we can evaluate the probability of a successful test, via Monte Carlo integration, as

$$\Pr[\text{Successful test} \mid \mathbf{x}] \approx \frac{1}{N} \sum_{j=1}^N \left[\sum_{y=0}^c \binom{n}{y} (1-\pi^{(j)})^y (\pi^{(j)})^{n-y} \right], \quad (8)$$

We can use assurance in this way to choose the sample size n for a given value of c . The critical number of failures c is chosen based on the analysis to be carried out following the test. This flexible approach allows the use of a frequentist hypothesis test or a Bayesian analysis to be used to decide on the success or failure of the test. If a Bayesian approach is to be used to analyse the test data, the prior for the analysis can be different to that used here in the sample size determination.

2.3. Cut-off choice

2.3.1. Frequentist approaches

Consider the binomial test in Section 2.1. We would reject H_0 if $\sum_{y=0}^c \binom{n}{y} (1-\pi_T)^y (\pi_T)^{n-y} \leq 0.05$. Therefore we simply select the largest value of c for which this is true.

An alternative to the exact binomial test is to use a Normal distribution to approximate the binomial distribution. In this case the test statistic is $Z = [c/n - \pi_T] / [\sqrt{\pi_T \times (1-\pi_T)/n}]$ and we would choose c to be the largest value for which $Z < Z_{0.05}$, where $Z_{0.05}$ is the 5% critical value of the standard Normal distribution.

2.3.2. Bayesian approaches

Suppose that, in the analysis of the test result, we have a prior distribution for π of $p_A(\pi)$. Note that this may be different from $p(\pi)$. Then we may choose a rule which says that if the posterior probability that $\pi \leq \pi_T$ is small, then the test is passed. That is, the test is passed if $\Pr(\pi \leq \pi_T \mid Y = y) \leq 0.05$, for example.

In this case we would choose c to be the largest value for which $\Pr(\pi \leq \pi_T \mid Y = c) \leq 0.05$. This posterior probability can be calculated as

$$\Pr(\pi \leq \pi_T \mid Y = c) = \frac{\int_0^{\pi_T} f(c \mid \pi) p_A(\pi) d\pi}{\int_0^1 f(c \mid \pi) p_A(\pi) d\pi}, \quad (9)$$

$$= \frac{\int_0^{\pi_T} \binom{n}{c} (1-\pi)^c \pi^{n-c} p_A(\pi) d\pi}{\int_0^1 \binom{n}{c} (1-\pi)^c \pi^{n-c} p_A(\pi) d\pi}, \quad (10)$$

where $f(c \mid \pi)$ is the probability mass function of Y evaluated at c .

This approach produces a reliability demonstration test plan, as the data have been used to calculate the

sample size n only and not to analyse the results of the test. A reliability assurance test plan can also be produced in this way by utilising the data $\mathbf{X} = \mathbf{x}$ here in the choice of c . This can be achieved using MCMC in the same way as when calculating n from known c in the previous section.

2.4. Choice of priors

In order to use assurance to decide on the optimal test plan (n, c) we need a design prior distribution for π , $p(\pi)$, to be used in the assurance calculation and, if a Bayesian analysis is to be undertaken following the test, a second prior distribution on π , $p_A(\pi)$, to be used in the analysis.

The design prior distribution to be used in the assurance calculation, $p(\pi)$ should represent the beliefs of the producer of the items on test. They take the risk associated with the failure of the test and so it is their probability that the test will be a success which will specify the sample size.

The design prior distribution therefore needs to be defined in terms of quantities about which we could reasonably ask an engineer. The beta distribution parameters (a, b) are not suitable. However, we can reparameterise the beta distribution so that $\pi \sim \text{beta}(mp, m(1-p))$, where $p = a/(a+b)$ is the mean and $m = a+b$ is the size of the “prior sample” which the mean is based on. We can then define suitable hyper-prior distributions on (p, m) , such as

$$p \sim \text{beta}(a_p, b_p), \quad (11)$$

$$m \sim \text{gamma}(a_m, b_m). \quad (12)$$

In the analysis following the test, the prior distribution of the producer may be deemed a controversial choice. There could be several different groups of people who will be affected by the decisions made following the success or failure of the test. Therefore, it may be more reasonable in the analysis of the test data to use a prior distribution which is relatively conservative. We call such a prior a sceptical prior.

The sceptical prior should be designed so that there is only a small probability that $\pi > \pi_T$. So, for example, if a $\text{beta}(\alpha, \beta)$ distribution is assumed for $p_A(\pi)$, then we might choose (α, β) to satisfy $1 - \frac{B(\pi_T; \alpha, \beta)}{B(\alpha, \beta)} = \delta$, where $\delta = 0.05$.

3. Example

We consider an example from Marz *et al* (1996). We would like to demonstrate the reliability of an emergency generator in a nuclear power plant. We will compare assurance based on two analysis methods: the exact binomial test and the Bayesian approach using a sceptical analysis prior. In the binomial test we reject H_0 if $p < 0.05$ and in the Bayesian approach we conclude that the test is passed if $\Pr(\pi \leq \pi_T | Y = y) \leq 0.05$.

We choose a beta distribution for the sceptical analysis prior with $\alpha = 13.6$ and $\beta = 2$. This gives $\Pr(\pi > \pi_T) = 0.05$. For the design prior $p(\pi)$ we choose to give (m, p) gamma and beta priors respectively with parameter values chosen to be $(78, 2)$ and $(200, 1)$. This equates to a prior mean with an expected value of 0.975 based on a prior sample with an expected value of 200. We consider a target reliability of 0.975.

Based on these specifications, we can plot the sample size n against the assurance for both of the approaches. This is given in Figure 1.

The binomial test case is the solid line and the sceptical prior case is the dashed line. We see that the binomial test gives the highest assurance for any particular sample size and the sceptical prior the lowest assurance in this case. The lines are decreasing as c remains constant but n increases, and then jump each time c increases. To achieve an assurance of 35% in this case would require $n = (418, 567)$ under the respective methods. We cannot achieve an assurance of greater than about 50% in this case.

However, we have data on the behaviour on demand of generators of the same type at other plants.

The data represent tests of 63 generators in nuclear power plants in the USA. In each case the number of failures on demand of the generator out of the total number of demands was recorded. The number of demands and the proportion of failures in the tests are given in Figure 2.

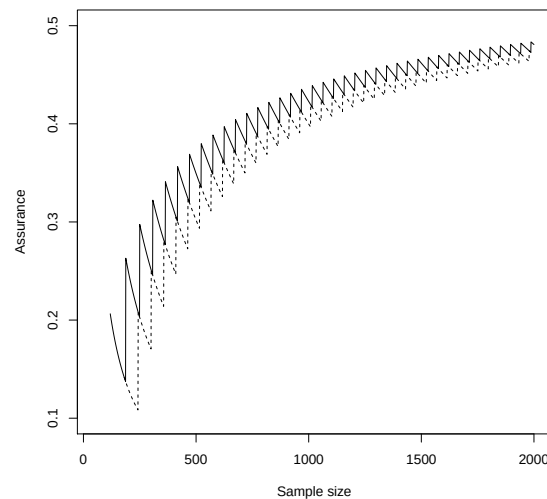


Figure1. The sample size n against the assurance for the binomial test (solid line) and the sceptical prior (dashed line).

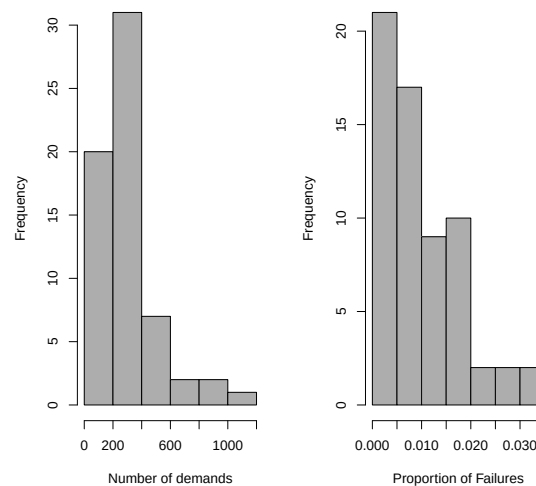


Figure2. The number of demands and proportion of failures of 63 emergency generators in nuclear power plants.

We see that we have some fairly large test data sets and that the failure proportions recorded in the test are very low, typically below 3%.

Using the same assurance prior as above, we generate 10,000 samples of (p, m) from the posterior predictive distribution in `rjags`, having discarded 1000 samples as burn in.

We are able to calculate the assurance for both of the methods as above using this posterior distribution for π . The results for different values of the sample size are given in Figure 3.

We see, by incorporating the data into the analysis, that we are able to provide a reliability assurance test which gives assurances of up to approximately 90%. To ensure assurance of 35% now only requires sample sizes of $n = (119, 243)$ for the analysis methods, both of which are much reduced. Using the sample sizes from the prior calculations would give assurances of $(0.715, 0.725)$ respectively.

4. Summary

In this paper we have considered the problem of sample size determination in reliability demonstration tests for hardware products leading to failure on demand data. The approach we have taken is to use the Bayesian concept of assurance, which chooses the sample size to answer the question “what is the probability that the RDT

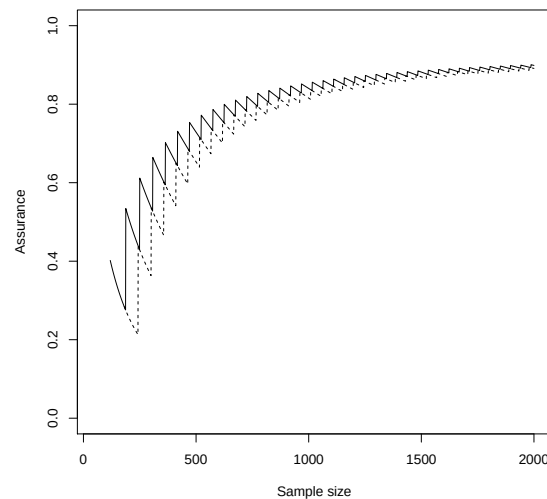


Figure3. The sample size n against the assurance for the binomial test (solid line) and the sceptical prior (dashed line).

will lead to a successful outcome?”. It can be used in conjunction with either frequentist or Bayesian analyses of the data following the test and, unlike typical risk criteria based approaches, separates the specification of prior beliefs in the design of the test from the prior to be used in the analysis following the test. Historical data can be incorporated into the sample size calculation without influencing the analysis of the RDT.

References

- Marz HF, Kvam PH, Abramson LR (1996) Empirical Bayes’ estimation of the reliability of nuclear-power-plant emergency diesel generators, *Technometrics*, 38, 11-24.
- Hamada MS, Wilson AG, Reese CS, Martz HF (2008) *Bayesian Reliability*, Springer.
- Meeker WQ, Escobar LA (2004) Reliability: the other dimension of quality, *Quality technology and quantitative management*, 1, 1-25.
- Tobias PA, Trindade DC (1995) *Applied Reliability*, Van Nostrand Reinhold.
- Easterling RJ (1970) On the use of prior distributions in acceptance sampling, *Annals of reliability and maintainability*, 9, 31-35.
- Spiegelhalter DJ, Freedman LS, Parmar MKB (1994) Bayesian approaches to randomized trials, *Journal of the Royal Statistical Society: Series A*, 157, 357-416.
- O’Hagan A, Stephens JW, Campbell MJ (2005) Assurance in clinical trial design, *Pharmaceutical Statistics*, 4, 187-201.
- Ren S, Oakley JE (2013) Assurance calculations for planning clinical trials with time-to-event outcomes, *Statistics in Medicine*, 33, 31-45.
- Plummer M (2016) rjags: Bayesian Graphical Models using MCMC, *R package version 4-6*, <https://CRAN.R-project.org/package=rjags>
- O’Hagan A, Stevens JW (2001) Bayesian assessment of sample size for clinical trials of cost-effectiveness, *Medical Decision Making*, 21, 219-230.
- Elsayed EA (2012) Overview of reliability testing, *IEEE Transactions on Reliability*, 61, 282–291.
- Martz HF, Waller RA (1982) *Bayesian Reliability Analysis*, Wiley, New York.

Maintenance centered matrix-based universal quality characteristics analysis

Yu Xiao*, Chuan Lv**, Jie Geng, Chunhui Guo, Ranran Wang

*School of reliability and systems engineering, Beijing University of Aeronautics and Astronautics (BUAA), Beijing, 100191, PR China Xiaoy@buaa.edu.cn

**School of reliability and systems engineering, Beijing University of Aeronautics and Astronautics (BUAA), Beijing, 100191, PR China lc@buaa.edu.cn

Abstract. Maintainability, as an important quality characteristic of the product, always been neglected during the design process, which lead to many problems around products' usage. The main purpose of the Maintenance centered matrix-based Universal quality characteristics analysis is, by determining the qualitative and quantitative relationship between the general quality features, to find the data flow, information flow between the common quality features and maintainability. Focusing on the maintenance process in a typical maintenance task model, the maintenance factors and general quality characteristics involved in the maintenance task model are analyzed, and the quantitative or qualitative correlation between the maintenance factors and the general quality characteristics is obtained. Finally, The maintenance matrix model integrated with maintenance-oriented, integrated maintenance elements, design specialties and design elements is available. By combining maintainability factors (observation, inspection, disassembly, replacement, installation, commissioning) with other related design discipline factors, correlating other design elements with maintainability elements, reaching a systematic trade-off centered on maintainability, which take maintenance of all elements into account, into the pre-design process.

1. Introduction

Maintenance is one of the quality characteristics that given to the product, determined in design stage. The main pursuit of today's product maintainability design is to obtaining high-quality products at the lowest cost. Maintainability is not only an economic indicator for the company, but also an urgent need of consumers.

Maintainability, as a formal concept, comes from the repairing problem of the equipment. The analysis of maintenance influencing factors and product maintenance practices show that the maintenance is convenient, fast, and economical, directly related to certain inherent design features of the product itself. As the complexity of equipment increase, reliability design can't always guarantee the stable operation of the product, and the convenient, fast, economic maintenance can make up for the lack of reliability, playing an important role in improving the integrity of the device

In the light of the existing maintenance design theory and methods, maintainability requirements on product design is still a challenging problem, maintenance requirements are difficult to produce a sufficient constraint on product design. Existing maintenance design theory and methods hardly fully meet the needs of device designers. The maintenance design work mode is passive, the development time lags behind, and it is inconsistent with the development of advanced design and manufacturing technologies. It is difficult to apply maintenance design to the current equipment design and development process and other issues.

At the same time, it is difficult to unify the design criteria due to the lack of communication between design departments, resulting in the lack of a more unified parameter standard transfer relationship during the design process. In the design process, little or no consideration is given to the constraints of other quality characteristics, leading to the insufficient of robustness in the design results and inability to obtain a mature and uniform design result.

International advanced aviation companies have established a similar RMS integrated integration system and have shown initial results, but they still need to improve the systematization, comprehensiveness, depth, and efficiency of the work. Parameter passing and shared mutual relationship between the design process, and the impact of factors such as the strength of the relationship still need to be clear.

In the rest of this paper, we first review the state of maintenance analysis, we then introduce our model along with a few others for comparison. After presenting the results and discussion we conclude with direction for future research.

2. Related work

Maintenance analysis methods and other related analysis methods are widely developed since the advantage are well known. Some [1] proposed a virtual prototype simulation of maintenance activities, human factors and accessibility factor analysis. Some focus on failure mode analysis and maintainability simulation model, include [2], [3].

Some [4] start with the maintenance qualitative design factors, taking the interaction of universal quality factors into consideration, propose a maintenance model and maintenance assessment methods. Boeing's ISDP (In-Service Data Program) collects and collates aircraft reliability/maintenance data, monitors aircraft reliability information, analyzes positioning issues, and helps airlines share aircraft quality characteristic information.

In China, AVIC [5] proposed a RMSTS integration system platform combining the correlation between the RMSTS data, achieving the integration of design work, integration of data flow, and seamless integration of interfaces.

3. Methods

We introduce a maintenance-centered model and bring in a matrix to link other quality characteristics together.

3.1 Maintenance task model analysis

By general maintenance task procedure analysis, we conclude a standard maintenance task model, including the correlation between maintainability and other quality characteristics.

Firstly, through the maintenance task input, we extract the maintenance information of a typical maintenance process and obtain the maintenance task model of the general equipment maintenance process. Then, by analyzing the maintenance tasks for the typical components, we make sure the maintenance information of the required maintenance level, maintenance tool requirements, maintenance operator level, and the number of maintenance operations in the maintenance process, so that the maintenance task are refined as a universal procedure.

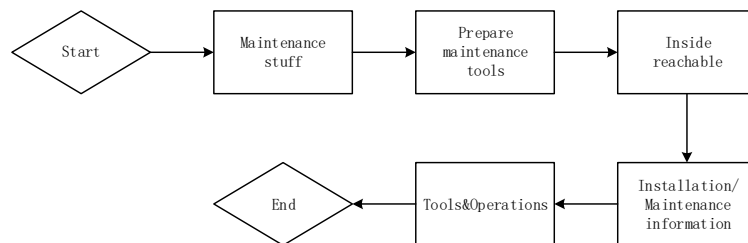


Figure 1. A standard maintenance task model that contains basic maintenance tasks.

These system or component universal maintenance procedure can generally show the operations that maintenance personnel must take at each step. And maintenance motions always involve in some maintenance elements, though the research on the maintenance elements, we can associate maintenance operations with related design disciplines in the design department, shown in table 1. Further more, combining with analysis tools, we can conduct maintenance centered correlation analysis on universal quality characteristics.

Table 1. The specific design departments that are related to the specific maintenance tasks.

Maintenance tasks	Maintenance elements	Design departments
Use of test points, test holes and adjustments	Visibility reachability	Testability
Demolition of fasteners.	Maintenance tools, Detachability	Feature design & distribution design
Use of lifting equipment.	Visibility, Reachability	Maintainability
Demolition of maintenance parts.	Maintenance tools, Detachability, Visibility, Reachability	Maintainability
Movement and replacement of maintenance parts.	Mobility	Feature design
Movement and use of maintenance tools.	Mobility, Interchangeability	Supportability

3.2 Maintenance centered correlation analysis on universal quality characteristics

Each step of the maintenance motion in maintenance operation is related to the product's general quality characteristics design results. Whether the product can be quickly, conveniently, accurately maintained, or whether maintenance personnel can perform maintenance operations safely and conveniently is closely related to the constraint relationship between maintainability and other general quality characteristics at the design stage.

Combining the characteristics of universal quality characteristics of typical system components and the correlations between maintainability, using the principle of maintenance and the correlation data flow between the universal quality characteristics, we can get the design constraint relationship between maintainability and universal quality characteristics.

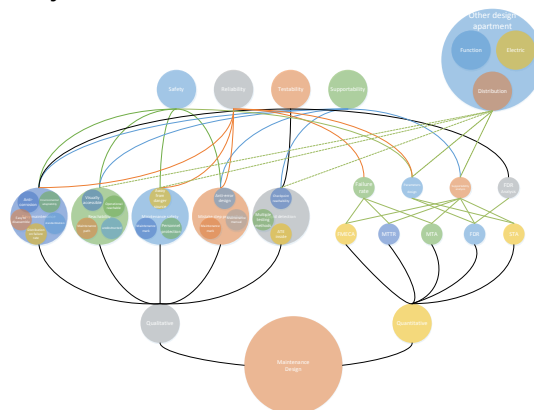


Figure 2. Maintenance centered correlation analysis on universal quality characteristics.

Maintainability is linked to design apartments and universal quality characteristics through qualitative and quantitative factors in Figure 2. The quantitative correlation factors include FMECA, MTTR, FDR, STA, etc. Qualitative factors include easy maintenance, reachability, maintenance safety, mistake operation prevention, rapid detection, etc. Each also contains three to five more detailed items.

3.3 Maintenance centered correlation design matrix

By linking maintenance elements (observation, inspection, disassembly, replacement, installation, commissioning) with related majors in other design fields, we can figure out the correlation between them. Let the maintenance elements as an axis of the 3D correlation matrix, the equipment design elements and the design professions as the other two axes, forming the maintenance centered 3D matrix. We use this kind of matrix to associate professional design elements with maintainability elements to achieve maintainability centered system balance.

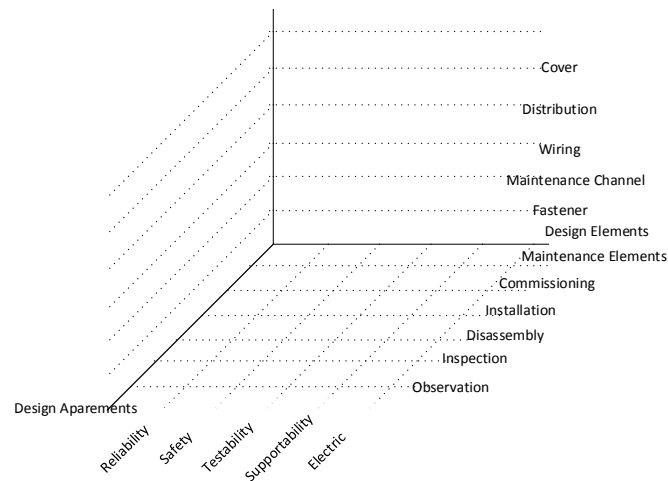


Figure 3. Maintenance centered correlation design matrix, it's a three-dimensional matrix.

The maintenance centered design matrix expresses the relationship between design majors, maintenance elements, design elements, and the correlation of design professions and maintenance elements in a three-dimensional table. As for some specific design element or design professions, it's a two-dimensional table in particular. In this two-dimensional table, what is obtained is an integrated table on specific design elements and maintenance, and the relationship between each design element, design specialty, and maintenance element can be clearly expressed. In the table, strong and weak relationships can be expressed by scores. The score is divided into 1/3/5/7/9. The lower the score, the weaker the correlation, and vice versa.

Table 2. Maintenance centered two-dimensional table.

Design elements Maintenance elements Design departments	Observation	Inspection	Disassembly	Replacement	Installation	commissioning
Reliability						
Testability						
Safety						
Supportability						
Electric						
Function						
Distribution						
.....						

After the expert's rating, the score can be assigned to design departments as constraints to guide product design.

4. Verification

The proposed model has been implemented, a case study on armored vehicle power compartment has been carried out.

First is to establish the standard maintenance task model that contains basic maintenance tasks. After maintenance information collection, the maintenance task model is shown in Figure 4.

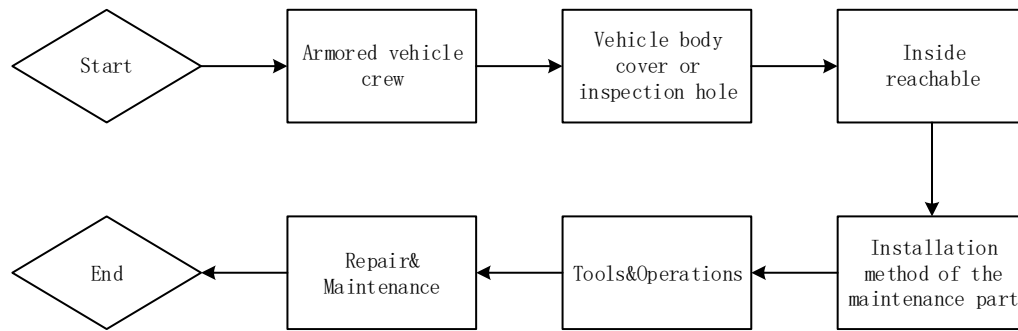


Figure 4. A standard maintenance task model that contains basic maintenance tasks of the armored vehicle.

As different parts of the equipment has the same correlation on the maintenance centered correlation analysis on universal quality characteristics, The result of this step is the same as 3.2 in this paper, so no more discussion here.

The last step is the maintenance centered analysis. By linking maintenance factors (observation, inspection, disassembly, replacement, installation, and commissioning) with related majors in other design fields, other professional design elements are associated with maintainability elements to achieve system balance based on maintainability. Take into account the full range of factors that influence maintenance design and incorporate maintenance design into the preliminary design work.

Table 3. Maintenance centered analysis two-dimensional table.

Maintenance elements Cover Design apartments	Observation	Inspection	Disassembly	Replacement	Installation	Commissioning
Reliability	3	5	3	3	3	5
Testability	9	9	7	3	3	9
Safety	3	3	5	7	7	5
Supportability	7	3	5	7	5	5
Electric	5	7	3	3	5	7
Function	7	5	3	3	5	5
Distribution	9	7	9	7	7	5

After the rating, we use normalization analysis to evaluating the importance of maintenance elements to design apartments under design element ‘cover element’.

It shows that to get better design for cover element, Observation, Commissioning and Inspection are important maintenance elements, and for the design apartments, reliability apartment should mind the design for inspection and commissioning; testability apartment should mind the design for observation, inspection, commissioning; safety apartment should mind the design for commissioning; supportability apartment should mind the design for observation; electric apartment should mind the design for inspection and commissioning; For function apartment inspection and disassembly should take more attention; For distribution apartment is observation. The results are in line with actual situation.

5. Conclusion

This study has shown that maintenance-centered model can help to linking all the design departments together in the early design phase, providing a passable parameter system for the design apartments, products are strong enough and functional healthy.

In future work the model will be extended to an visualization software, and a more intuitive scoring method is necessary for the model.

Reference

- Kraus D, Gramopadhye A K (1999) Team training: role of computers in the aircraft maintenance environment. *Computers & Industrial Engineering*, 36(3):635-654.
- Huagen Wan, Shuming Gao, Qunsheng Peng (2002) VDVAS: An integrated virtual design and virtual assembly system. *Chinese Journal of Image and Graphics*, 7(1):27-35.
- Shuanhang Yan (2014) Research on Aircraft RMSTS Integration Design. *Aviation Science and Technology*, 2014(6):1-4.
- Liao Wu et al (2005) Combat Aircraft Reliability and Maintainability Simulation Model. *Firepower and Command Control* 30(7):50-53.
- Bing Feng et al (2003) Research on Evaluation Method of Maintainability Design of Weapon System. *Journal of Ordnance*, 2003, 24(2):222-225.
- Shuangwei Xu (2007) Ship RMS Modeling, Simulation and Comprehensive Evaluation Technology. National University of Defense Technology.

Dynamic imperfect condition-based maintenance for systems subject to nonlinear degradation paths

Xiujie Zhao*, Zhenglin Liang**, Ajith Parlikad**, Min Xie*

**Department of Systems Engineering and Engineering Management, City University of Hong Kong; City University of Hong Kong Shenzhen Research Institute, email address: xiujizhao2-c@my.cityu.edu.hk; minxie@cityu.edu.hk*

***Institute for Manufacturing, University of Cambridge, Cambridge, UK, email address: zliangcam@gmail.com; aknp2@cam.ac.uk.*

Abstract. This paper presents a maintenance optimization framework for systems suffering from nonlinear continuous degradation. The inspection interval is dynamically determined by the historical system conditions. We model the degradation path as a nonlinear Wiener process with time-varying drift parameter. Techniques to predict remaining useful life (RUL) is utilized to optimize maintenance policy by minimizing the expected cost rate. The effect of imperfect maintenance is assumed to be random in the sense that the maintenance action reduces the systems degradation level by a random proportion described by a beta distribution. Two thresholds on the degradation are determined for the preventive imperfect maintenance and perfect replacement, respectively. We evaluate expected cost rate using Monte Carlo simulation. A dataset from the real-world example is used to provide the pilot parameters as input for the optimization maintenance policies. Afterward, numerical examples are presented to illustrate the proposed method.

1. Introduction

With the rapid development of technologies to monitor and measure system deterioration, condition-based maintenance (CBM) is playing a more important role to maintain systems that suffer measurable degradations (Lu and Meeker, 1993). To facilitate condition-based maintenance, maintenance operations are scheduled based on the online measurements of quality characteristics, such as degradation level, instead of following a fixed schedule or age policy. For modern manufacturing systems, the planning of maintenance based on online status is of great importance to ensure production efficiency and products' quality (Zied et al., 2011; Beheshti Fakhri et al., 2017).

The advance in sensing and inspecting technology significantly facilitates the lifetime prediction via system quality characteristics. Degradation models have drawn significant attention from both researchers and engineers to analyze highly reliable degrading systems. Two common types of degradation models are general path models and stochastic process models (Ye and Xie, 2015; Zhai and Ye, 2017). Under the degradation models, CBMs are the policies to schedule maintenance actions based on a system's actual degradation level which is measured under a certain inspection policy. Stochastic degradation models have been widely used in CBM optimization because of their clear physical explanations and tractable mathematical properties. For an overview, one can refer to Alaswad and Xiang (2017).

Although there exist numerous studies that address the maintenance optimization problems from various aspects, most of them have assumed linear degradation models. However, in real applications, it is found that many degradation paths have nonlinear behaviors. For example, as described in Si et al. (2012), nonlinear degradation characteristics widely exist in electronic devices, such as LED lamps and carbon-film resistors. Further, a very common assumption adopted in literature on maintenance optimization is that inspection is scheduled periodically (Liu et al., 2017; Chen et al., 2015). Nonlinearity in degradation paths makes the periodic inspections inappropriate since the mean degradation increments between two inspections can be drastically different, which brings more risks of system failure. Thus, non-periodic inspections are preferred in this scenario. Do et al. (2015) proposed a proactive maintenance policy with the aid of remaining useful life (RUL), and the inspection interval was dynamically determined based on the system status.

Another concern in maintaining systems is the choice between repair and replacement. More often than not, repairs are imperfect (Pham and Wang, 1996). On the one hand, from the perspective of degradation, an imperfect repair can generally reduce the degradation level though cannot make the system as-good-as-new. On the other hand, replacements are usually conducted with new systems thus reset the system degradation to the initial status that is the same with new systems. A tradeoff is that replacement is better in the extent of

improvement but it is generally more expensive than imperfect repairs. Decision makers need to determine whether to repair or replace a system upon a maintenance action to save more costs.

In this paper, we present a RUL-based dynamic maintenance policy for systems subject to nonlinear degradation paths. Two maintenance thresholds are determined for repairs and replacements. Optimal maintenance policy is obtained by Monte Carlo simulation. The remainder of the paper is organized as follows. Section 2 describes the nonlinear Wiener process for reliability modeling. The optimization of maintenance policy is presented in Section 3 in detail. Section 4 presents a numerical example to illustrate the methods. Finally, Section 5 concludes the paper.

2. Nonlinear Wiener process for degradation modeling

Wiener process is widely used to model non-monotone degradation paths (Li et al., 2017; Ye et al., 2013). A well-adopted expression of drifted Wiener process $\{X(t), t > 0\}$ is given as follows:

$$X(t) = \beta\Lambda(t) + \sigma B(\Lambda(t)), \quad (1)$$

where $\beta > 0$ and $\sigma > 0$ are the drift and diffusion parameters, respectively, and $B(\cdot)$ is the standard Brownian motion. Further, $\Lambda(\cdot)$ is a parametric transformation function that describes the nonlinearity of the degradation path. For example, $\Lambda(t) = t$ implies that the expected degradation path is linear. Note that $\Lambda(\cdot)$ can be in various expressions, such as $\Lambda(t) = t^b$, $\Lambda(t) = \exp(t)$ and $\Lambda(t) = \log at$. Under the current degradation model, the failure time of a system is typically determined by the first passage time (FPT) of $Y(t)$ to a specific failure threshold L . The system failure time, denoted by T , follows an inverse Gaussian (IG) distribution of which the density function is given by

$$F_T(t; \beta, \sigma, b, L) = \Phi \left\{ \sqrt{\frac{L^2}{\sigma^2 \Lambda(t)}} \left[\frac{\beta \Lambda(t)}{L} - 1 \right] \right\} + \exp \left(\frac{2\beta L}{\sigma^2} \right) \Phi \left\{ - \sqrt{\frac{L^2}{\sigma^2 \Lambda(t)}} \left[\frac{\beta \Lambda(t)}{L} + 1 \right] \right\}. \quad (2)$$

3. A two-threshold maintenance policy with dynamic inspection intervals

3.1 Assumptions

Consider a single-component system of which the degradation can be well modeled a parametric nonlinear Wiener process as in Eq. (1). The following assumptions are adopted in this work:

- The failure of the system is assumed to be hidden. In other words, system failures can only be revealed by inspections.
- The initial system degradation is zero. The assumption also applies to systems with non-zero but identical initial degradation levels. Further, when a system is replaced, the new system also has a degradation level of zero.
- Replacement and repairs can only be carried out upon inspections, and their durations can be neglected compared to the system life.
- The degradation rate of the system is not influenced by maintenance at any time.
- The system is discretely inspected at time epochs $T_1, T_2, \dots$, and the inspection intervals are dynamically determined. For convenience, we let $\Delta T_i = T_i - T_{i-1}$, where $i = 1, 2, \dots$ and $T_0 = 0$.

3.2 Imperfect repair with random improvement factor

We propose a random improvement factor to describe the effect of repair actions. Let X_i be the observed degradation level at the time T_i , i.e., the i th inspection. If a repair is conducted on the system at T_i , the system degradation level after the repair, denoted by X_i^+ . We have

$$X_i^+ = \alpha_i X_i, \quad (3)$$

where $0 < \alpha_i < 1$ is the improvement factor of the imperfect repair. To characterize the randomness in the effect of improvement, we assume that $\alpha_i, i = 1, 2, \dots$ follow i.i.d. beta distributions with parameters u and v (Zhang et al., 2015), and the density function is given by

$$f_{\alpha_i}(z; u, v) = \frac{\Gamma(u + v)}{\Gamma(u)\Gamma(v)} z^{u-1} (1 - z)^{v-1}, \quad 0 \leq z \leq 1. \quad (4)$$

When $\mu > v$, the distribution is left skewed and the mean is greater than 0.5. Note that the random improvement factor only applies to repairs, and replacement will reset the degradation level to zero. Figure 1 shows various shapes of the density curves of beta distribution under different parameters.

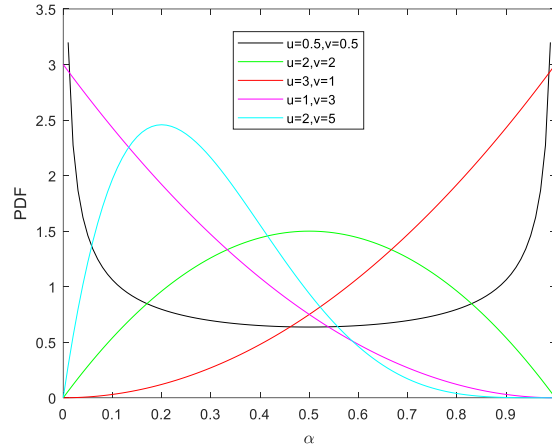


Figure 1. Various shapes of density functions for beta distributions

3.3 Maintenance policy under dynamic RUL-based inspection intervals

As part of the maintenance policy, inspection schedules are of great importance to reduce the risks of system failures and maintenance cost. We adopt the RUL-based inspection policy proposed by Do et al. (2015). Specifically,

$$\Delta T_i = \Delta T_i(Q, X_{i-1}) = \{\Delta t: \Pr(T - T_{i-1} \leq \Delta T | X_i) = Q\}, \quad (5)$$

where

$$\begin{aligned} & \Pr(T - T_{i-1} \leq \Delta T | X_i) = F'_T(\Delta T; \beta, \sigma, b, L - X_{i-1}) \\ & = \Phi \left\{ \sqrt{\frac{L^2}{\sigma^2 [\Lambda(T_{i-1} + \Delta T) - \Lambda(T_{i-1})]}} \left[\frac{\beta [\Lambda(T_{i-1} + \Delta T) - \Lambda(T_{i-1})]}{L} - 1 \right] \right\} \\ & + \exp\left(\frac{2\beta L}{\sigma^2}\right) \Phi \left\{ -\sqrt{\frac{L^2}{\sigma^2 [\Lambda(T_{i-1} + \Delta T) - \Lambda(T_{i-1})]}} \left[\frac{\beta [\Lambda(T_{i-1} + \Delta T) - \Lambda(T_{i-1})]}{L} + 1 \right] \right\}. \end{aligned} \quad (6)$$

$Q \in (0,1)$ is the probability that the system fails between T_{i-1} and T , thus, generally we assign a relatively small value. We define the following types of cost that can be incurred to the system: inspection cost C_I ; preventive replacement cost C_{PR} ; repair cost C_R ; corrective replacement cost per unit of time C_{CR} . The two-threshold maintenance policy has three decision variables to be determined by the planners, Q, M_1 and M_2 , where M_1 and M_2 are the thresholds for repair and replacement. Upon each inspection, if the degradation level is smaller than M_1 , the system will continue to operate until the next inspection; if the degradation level is larger than M_1 while smaller than M_2 , the system is subject to imperfect repair; otherwise, the system is replaced immediately.

The cumulative maintenance cost induced to the system by time, denoted by $C(t; Q, M_1, M_2)$, can be expressed by

$$C(t; Q, M_1, M_2) = C_I N_I(t) + C_R N_R(t) + C_{PR} N_{PR}(t) + C_{CR} N_{CR}(t), \quad (7)$$

where $N_I(t)$, $N_R(t)$, $N_{PR}(t)$ and $N_{CR}(t)$ are the number of inspections, imperfect repairs, preventive replacements and corrective replacement. The asymptotic cost rate can be obtained by

$$CR(Q, M_1, M_2) = \frac{E[C(H; Q, M_1, M_2)]}{E(H)}, \quad (8)$$

where H is the length of the renewal cycle, i.e., the time at which the system is replaced for the first time. We employ Monte Carlo simulation to evaluate the $CR(Q, M_1, M_2)$ for various combinations of Q, M_1 and M_2 , denoted by maintenance policy $\mathcal{M}$. Note that the distribution function in Eq. (6) reflects that $T - T_{i-1} \leq \Delta T | X_i$ also follows inverse Gaussian distribution.

4. Numerical example

In this section, we use a numerical example similar to that in Ye et al. (2013). In the example, the wear of magnetic heads of HDDs is modeled as a degradation characteristic by Wiener process. The degradation path is deemed to be nonlinear. The following preliminary parameters are given for the maintenance optimization:

$$L = 10, \beta = 0.133, \sigma = 0.0627, \Lambda(t) = t^{1.518}, \alpha_i \sim \text{Beta}(3, 2). \quad (9)$$

Additionally, the cost vector is given by $C = (C_I, C_R, C_{PR}, C_{CR})' = (10, 50, 70, 100)'$. For various combinations of Q , M_1 and M_2 , we calculate the asymptotic cost rate to obtain the optimal maintenance policy. Table 1 shows part of the results.

Table 1. Asymptotic cost rate of various $\mathcal{M}$

$\mathcal{M}$			$CR(\mathcal{M})$
Q	M_1	M_2	
0.02	3	5	2.2675
0.02	3	6	2.2688
0.02	3	7	2.2682
0.02	4	6	2.2674
0.02	4	7	2.2680
0.02	4	8	2.2957
0.02	5	7	2.2681
0.02	5	8	2.2964
0.02	5	9	2.7371
0.05	3	5	2.2538
0.05	3	6	2.2532
0.05	3	7	2.2547
0.05	4	6	2.2543
0.05	4	7	2.2505
0.05	4	8	2.2590
0.05	5	7	2.2513
0.05	5	8	2.2631
0.05	5	9	2.5777
0.10	3	5	2.2562
0.10	3	6	2.2588
0.10	3	7	2.2584
0.10	4	6	2.2589
0.10	4	7	2.2543
0.10	4	8	2.2642
0.10	5	7	2.2564
0.10	5	8	2.2617
0.10	5	9	2.4799

From the results, we can observe that the optimal maintenance policy is close to $\mathcal{M} = (Q, M_1, M_2)' = (0.05, 4, 7)'$. To obtain a more accurate solution, a numerical search around $\mathcal{M} = (0.05, 4, 7)'$ is conducted, and $\mathcal{M}^* = (0.073, 4.3, 7.4)'$ is the global optimal solution for the maintenance policy. It is noted that further sensitivity analyses are needed to investigate the robustness of the optimal maintenance policy.

5. Conclusions

In this paper, we present a condition-based maintenance policy with two maintenance thresholds and dynamic inspection policy. Preliminary analytical results are given to facilitate the Monte Carlo simulation to search

for the optimal maintenance policy. A simplistic numerical example from previous literatures is revisited to illustrate the proposed method.

The main challenge in optimizing the maintenance policy lies in the computational burden induced by Monte Carlo simulation. Thus, it is of great interest to derive the analytical form of the asymptotic cost rate for the problem.

References

- Alaswad S, Xiang Y (2017) A review on condition-based maintenance optimization models for stochastically deteriorating system. *Reliability Engineering & System Safety* **157**, 54–63.
- Beheshti Fakher H, Nourelfath M, Gendreau M (2017) A cost minimisation model for joint production and maintenance planning under quality constraints. *International Journal of Production Research* **55**(8), 2163–2176.
- Chen N, Ye Z-S, Xiang Y, Zhang L (2015) Condition-based maintenance using the inverse Gaussian degradation model. *European Journal of Operational Research* **243**(1), 190–199.
- Do P, Voisin A, Levrat E, Iung B (2015) A proactive condition-based maintenance strategy with both perfect and imperfect maintenance actions. *Reliability Engineering & System Safety* **133**, 22–32.
- Li J, Wang Z, Zhang Y, Fu H, Liu C, Krishnaswamy S (2017) Degradation data analysis based on a generalized Wiener process subject to measurement error. *Mechanical Systems and Signal Processing* **94**, 57–72.
- Liu B, Wu S, Xie M, Kuo W (2017) A condition-based maintenance policy for degrading systems with age- and state-dependent operating cost. *European Journal of Operational Research* **263**(3), 879–887.
- Lu CJ, Meeker WQ (1993) Using degradation measures to estimate a time-to-failure distribution. *Technometrics* **35**(2), 161–174.
- Pham H, Wang H (1996) Imperfect maintenance. *European Journal of Operational Research* **94**(3), 425–438.
- Si X-S, Wang W, Hu C-H, Zhou D-H, Pecht MG (2012) Remaining useful life estimation based on a nonlinear diffusion degradation process. *IEEE Transactions on Reliability* **61**(1), 50–67.
- Ye Z-S, Wang Y, Tsui K-L, Pecht M (2013) Degradation data analysis using Wiener processes with measurement errors. *IEEE Transactions on Reliability* **62**(4), 772–780.
- Ye Z-S, Xie M (2015) Stochastic modelling and analysis of degradation for highly reliable products. *Applied Stochastic Models in Business and Industry* **31**(1), 16–32.
- Zhai Q, Ye Z-S (2017) RUL prediction of deteriorating products using an adaptive Wiener process model. *IEEE Transactions on Industrial Informatics* **13**(6), 2911–2921.
- Zhang M, Gaudoin O, Xie M (2015) Degradation-based maintenance decision using stochastic filtering for systems under imperfect maintenance. *European Journal of Operational Research* **245**(2), 531–541.
- Zied H, Sofiene D, Nidhal R (2011) Optimal integrated maintenance/production policy for randomly failing systems with variable failure rate. *International Journal of Production Research* **49**(19), 5695–5712.

A Visualization Automated Analysis and Evaluation Method Based on Virtual Reality

Zhou qidi * and Zhou dong** and Xie yushu*** and Zhou wenqiang****

**School of Reliability and Systems engineering, Beihang University Beijing, China*

+qidizhou@buaa.edu.cn

***School of Reliability and Systems engineering, Beihang University Beijing, China*

+zhoudong@buaa.edu.cn

****School of Reliability and Systems engineering, Beihang University Beijing, China*

+buaaxys1314@163.com

*****School of Reliability and Systems engineering, Beihang University Beijing, China*

+2425771973@qq.com

Abstract. To ensure products with high reliability and quality, maintainability design and verification is an integral part of design process.

As a crucial factor in maintainability design, visibility evaluation is the most important part of maintainability design and verification. However, subjective and expert judgment in evaluation methods are the two major problems limiting the development and application of the visibility design. To solve these problems, this paper proposes a visibility automation analysis and evaluation model based on virtual maintenance. Firstly, two sub-evaluation indicators are designed for visual ability and visual obstruction. Secondly, in responds to the two indicators, two models are established based on visual cone and feature points respectively. Moreover, a comprehensive visual accessibility evaluation model based on line-of-sight interference detection and two models is developed. Finally, a case of an engine coupling fastening bolt visibility assessment is conducted and verifies the effectiveness of the proposed method.

1. Introduction

At present, with the increasing level of automation and intelligence in large-scale equipment equipped with high reliability and quality characteristics, maintainability design and verification have become an indispensable research area in the equipment design process[1]. Currently, virtual reality technology can already be applied in creating a virtual space that includes maintenance personnel, maintenance tools, 3D digital prototypes, and maintenance processes. In this space, it is possible to simulate the equipment maintenance process, provide strong support for the analysis and evaluation of equipment maintenance visibility, detect the lack of equipment maintenance design in advance, and put forward suggestions for improvement [2]. Zhang Yuantao and Wang Zhanhai studied a CATIA-based visual field analysis function, and visualized the target object's visual accessibility by superimposing the left-eye and right-eye cones as the binocular visual area[3 , 4]. Yang Chunhui and others used the quality function development and evidence theory method to construct the evaluation index system of each component, and obtained the quantitative evaluation level of visual accessibility of complex equipment[5]. However, the maintenance personnel currently use the virtual reality technology based on maintenance evaluation method to subjectively evaluate the virtual maintenance process, unable to provide objective data to support the evaluation results, rely on expert experience, and lack of objective, quantitative, and comprehensive evaluation methods. Therefore, how to rely on virtual maintenance technology to study how to achieve accessibility automated analysis and evaluation has become the key to improving the design level of equipment maintenance. This paper firstly establishes a visualized automated analysis and evaluation model, and conducts in-depth study on two sub-evaluation indicators: line-of-sight accessibility and visual occlusion barrier. The visible cone-based visual model and the feature point-based visual occlusion evaluation model are studied. Based on the above, a visual comprehensive evaluation model based on the line of sight interference detection was established to realize visualized automated analysis and evaluation. This part of our specific discussion will be discussed in the second part. The verification of the case will be discussed in the third section. The conclusion will be discussed in the fourth part.

2. Methodology

This section first analyzes the two indicators of the evaluation of visual accessibility design, the line-of-sight accessibility and the visual occlusion barrier. The visual cone-based sight accessibility model and the visual occlusion barrier model based on feature points are proposed separately for the two indicators, on the foundation of which, a visual comprehensive evaluation model based on line-of-sight interferometric detection realizes visualized automatic analysis and evaluation of maintenance operations. The methodology process is shown in Fig 1.

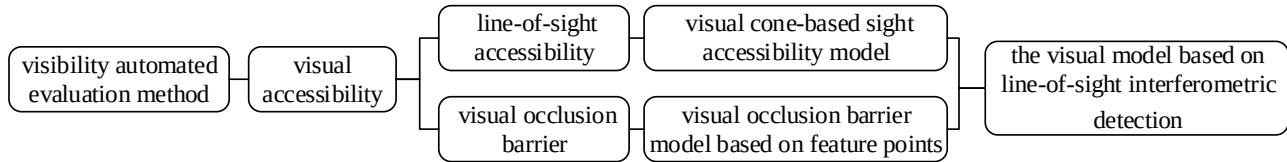


Figure 1. Visibility automated evaluation method research process

2.1 Evaluation

Good visibility requires equipment maintenance personnel to maintain a natural and comfortable view of maintenance components that are not obstructed during maintenance operations[6]. According to the basic principles of visual design and data information in the virtual maintenance environment, the visual accessibility is divided into two aspects: the line-of-sight accessibility and the visual occlusion barrier.

1) Line-of-sight accessibility

Line-of-sight accessibility refers to whether the repair parts are within the field of view of the maintenance personnel during maintenance. Line-of-sight accessibility is the basic requirement for maintenance operations and relates to maintenance safety incidents[7].

2) Visual occlusion barrier

Visual occlusion refers to the condition of maintenance parts that cannot be accurately observed by maintenance personnel because of obstructions caused by maintenance parts or environmental parts during maintenance. Visual obstruction causes maintenance personnel to operate blindly, which may lead to maintenance safety.

2.2 Visual cone-based sight accessibility analysis and evaluation model

In terms of line-of-sight accessibility, coordinate transformation is first carried out. Based on the maintenance personnel's visual field model, the coordinate information of the maintenance component feature points in the virtual environment coordinate system is converted into the visible cone coordinate system. Secondly, by assessing whether the visible feature point is within the field of view of the maintenance personnel, the sight-accessibility design of the maintenance component is evaluated.

2.2.1 The mathematical model of the maintenance personnel's field of vision

According to the provisions of the scope of maintenance, maintenance personnel visual cone can be regarded as an elliptical cone approximately[8], with the middle point of sight of two eyes as the origin O of the coordinates, the vertical direction of the visual field is the x-axis, the horizontal direction of the visual field is the y-axis, and the normal line of sight is the z-axis. In the cone coordinate system O, the parametric Equation(1) is established for the maximum visual field of maintenance personnel:

$$[x_0, y_0, z_0]^T = [0.7002u \sin v, 0.5774u \cos v, u]^T \quad (1)$$

The length of the line of sight is the angle between the line connecting the point on the visible cone and the point of view and the Y axis.

2.2.2 Change matrix

Since the data information in the virtual maintenance environment is based on the world coordinate system of the entire virtual environment, the virtual human visual cone model is established in the virtual visual cone coordinate system, so a homogeneous transformation matrix of virtual world coordinate system A to visible cone coordinate system O is given by Equation(2):

$$T = R(B) \times R(O) \quad (2)$$

R(B) represents the position and attitude transformation matrix of the virtual environment coordinate system A with respect to the virtual maintenance personnel coordinate system B, and Equation(3) is given in combination with the three maintenance postures during maintenance.

$$R(B) = R_x(\theta) \times R_y(\phi) \times R_z(\varphi) \times T = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos \theta & \sin \theta & 0 \\ 0 & -\sin \theta & \cos \theta & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos \phi & 0 & -\sin \phi & 0 \\ 0 & 1 & 0 & 0 \\ \sin \phi & 0 & \cos \phi & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos \varphi & \sin \varphi & 0 & 0 \\ -\sin \varphi & \cos \varphi & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ T_x & T_y & T_z & 1 \end{bmatrix} \quad (3)$$

Wherein, $R_x(\theta)$ indicates that the coordinate system A is rotated by θ degrees around the X axis with respect to the coordinate system B, and $R_y(\varphi)$ indicates that the coordinate system A is rotated by φ degrees around the Y axis with respect to the coordinate system B, $R_z(\phi)$ indicates that the coordinate system A is rotated by ϕ degrees around the Z axis with respect to the coordinate system B. Degree angle, the offset matrix of the origin of the R(O) table coordinate system A in the coordinate system B.

R(O) represents the position and attitude transformation matrix of the maintenance personnel coordinate system B with respect to the visible cone coordinate system O.

2.2.3 Mathematical model based on visual cone for sight line accessibility analysis and evaluation

The coordinates of the maintenance component feature point in the visible cone coordinate system O can be obtained through coordinate transformation. If the content is satisfied in the visible cone Equation(4), the

$$(x_i/0.7002)^2 + (y_i/0.5774)^2 \geq z_i^2 \quad (4)$$

feature point is located in an invisible region; otherwise, it is located in the maximum visual field region.

In the virtual environment, maintenance component information, virtual person information, and virtual environment information can be obtained. The coordinates are transformed into the visual cone coordinate system, and the coordinates of the maintenance component feature points are substituted into the visual cone equation, so that whether or not each point is located in the maintenance personnel can be obtained. Within the visible range. When all the feature points are located in the visible cone of the maintenance personnel, the line-of-sight accessibility of the maintenance part is well-designed, otherwise it is not qualified.

2.3 Visual occlusion barrier evaluation model based on feature points

The evaluation mathematical model firstly extracts the features of the 3D mesh model of the maintenance parts and obtains the collection of the feature points, which is represented by Q_i . The Q_a are the visible feature points of the maintenance parts, and the occlusion of the feature points can be classified into two categories: (1)The feature points Q_b are occluded by maintenance parts themselves, (2) The Q_c are occluded by other maintenance parts in the virtual environment, and the maintenance personnel cannot see it. The Maintenance parts visual obstruction diagram is shown in Figure 2.

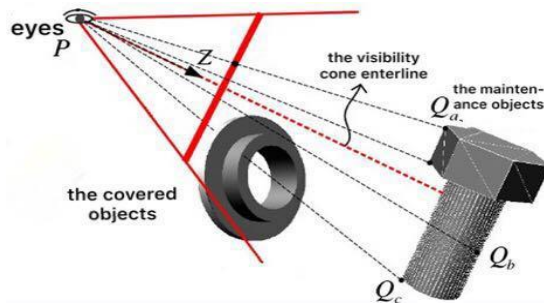


Fig 2. Maintenance parts visual obstruction diagram

2.3.1 Principle of algorithm

(1) Construct the parametric equation of sight. Set (x_0, y_0, z_0) as the starting point of the line of sight. Calculate the direction vector of the line of sight (x_d, y_d, z_d) through the coordinates of the feature point and the coordinates of the point of sight. And the sight (x_0, y_0, z_0) should satisfy Equation(5).

$$x_d^2 + y_d^2 + z_d^2 = 1 \quad (5)$$

The parametric equation shown in Equation (6) shows the line of sight emitted by the point:

$$\begin{cases} x = x_0 + x_d \times t \\ y = y_0 + y_d \times t \\ z = z_0 + z_d \times t \end{cases} \quad (6)$$

(2) AABB Axial Bounding Boxes for Constructing Shading Parts. In various bounding box technologies, the calculation of building the AABB bounding box of the masking component model is very simple, and it is only necessary to calculate the maximum and minimum values of the x-coordinate, y-coordinate, and z-coordinate of each feature point. The feature points of the occlusion part model extracted by the above algorithm are applied to obtain the feature point set FB of the model. Traverse the FB to calculate the maximum, $x_{max}, y_{max}, z_{max}$ the coordinates, the minimum $x_{min}, y_{min}, z_{min}$ the coordinates, respectively, to construct the AABB bounding box of the model.

(3) Determine if the line of sight intersects the AABB bounding box. Equation(7) represents the plane P,

$$AX + BY + CZ + D = 0 \quad (7)$$

substituting the line-of-sight parametric equation, so that Equation(8) can be obtained.

$$t = -b/a \quad (8)$$

When $a=0$, the line of sight is parallel to the plane; if $t < 0$, the corresponding point of intersection is not on the line of light, and the intersection is invalid. Substituting the value into (6) yields the coordinates of the intersection.

(4) Determine if the intersection is on the surface of the bounding box. Although there is an intersection between the line of sight and the bounding box plane, it does not necessarily fall within the rectangular space surrounding the box surface. Determine whether the intersection point is in the axial bounding box. Compare the x_i, y_i and z_i value of the coordinates of the vertex with the maximum $x_{max}, y_{max},$ and z_{max} the coordinates, the minimum x_{min}, y_{min} and z_{min} , and the coordinates respectively. If $x_{min} < x_i < x_{max}$ $y_{min} < y_i < y_{max}$ and $z_{min} < z_i < z_{max}$ are both true, the intersection is in the bounding box.

2.4 Visual accessibility automated analysis evaluation model

The main steps of automated visual assessment of visual accessibility are as follows:

- (1) Acquire virtual human data, virtual prototype data, visual obstruction data, and extract feature data;
- (2) Establish a mathematical model of maintenance personnel's field of vision to determine the sight accessibility of maintenance parts;
- (3) Acquire the current viewpoint position, and construct the line-of-sight parameter equation according to the data of the repair component feature point;
- (4) Construct the AABB bounding box of the visual masking component according to the set of obstacle feature points;
- (5) Iterate through the eight surfaces of the bounding box, check whether the line of sight intersects the plane where the bounding box surface lies, and check whether the intersection point is within the bounding box surface rectangle. If the intersection point is within the range of the surface rectangle, it means that the component blocks the line of sight and stores the feature point in the collection; otherwise, it continues to judge other feature points;
- (6) The above algorithm can be used to judge the occlusion detection result of the viewpoint for a certain feature point. According to the maintenance visual accessibility design criterion, if all visible feature points that are not occluded by itself are not occluded by external objects, the maintenance component visual accessibility is well designed. If the collection L_N not empty, it means that the maintenance part is not visible under the current visual angle, and the visual accessibility design of the maintenance part is unqualified; if the collection L_N is an empty set, the visual accessibility of the maintenance part is well designed. Visual accessibility automated analysis and evaluation process is shown in Fig 3.

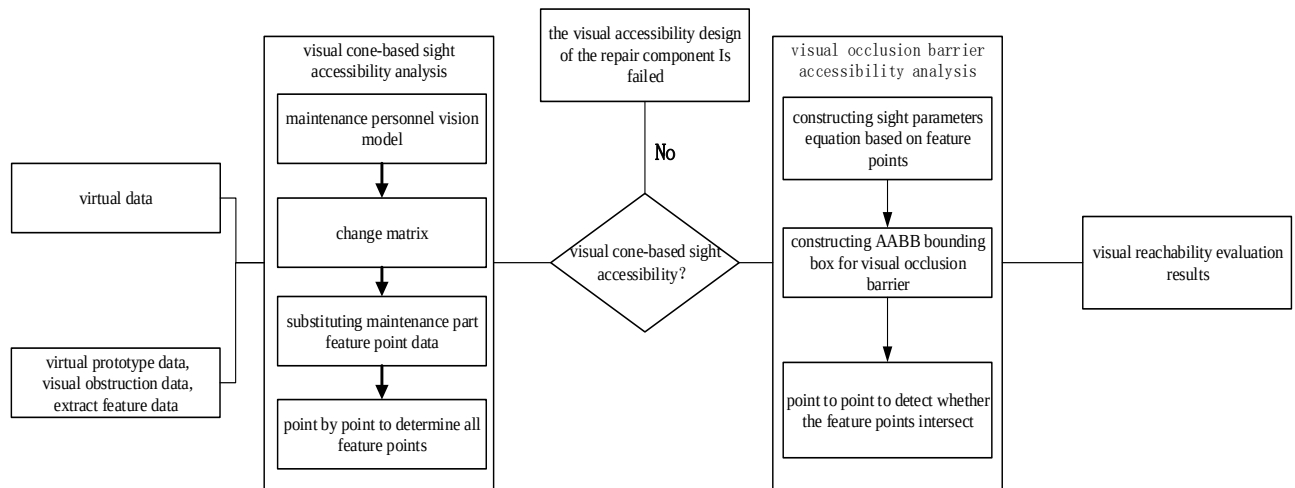


Figure 3. Visual accessibility automated analysis and evaluation process

3. Case analysis

This chapter selects a fastening bolt on an engine coupling as a maintenance component, and uses the shaft sleeve connector as an obstacle. The method of Chapter 2 is used to automate evaluate the visibility of the maintenance operation of a virtual person holding a wrench removal bolt. The DELMIA virtual simulation platform was used to verify the accessibility evaluation results, thus verifying the feasibility of this method.

3.1 Preparatory preparation

The virtual reality simulation platform chosen in this paper is DELMIA software, which is mainly used for mechanical motion simulation and ergonomics simulation in the aerospace field. Using Automation technology, secondary development of DELMIA can be performed in the Visual Basic 6.0 development environment. The virtual maintenance scene consists of a virtual prototype CAD model, a virtual human model, and a maintenance tool model[9]. Virtual maintenance scene is shown in Fig 4.

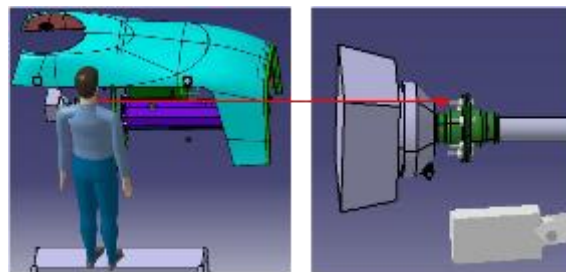


Figure 4. Virtual maintenance scene

Use the DELMIA interface function to collect CAD model data including maintenance parts and environmental part model data around the repair path, as well as to maintain tool size and attitude data. The input program extracts repair part feature point data and environmental part feature point data as well as virtual person related data.

3.2 Visibility automated evaluation method

In the evaluation of visibility, a virtual person root node spatial position coordinates (-585.478, -563.501, -167.99), the avatar's binocular center coordinate data (-567.683, -455.954, 443.943), Virtual person's head back and forth angle 9.952°, the maintenance part feature point data FA the feature point set FB of the shielding part have been obtained.

3.2.1 Visual cone-based sight accessibility model verification

Using the calculation method mentioned above, the DELMIA software is used to obtain the coordinates(x_i, y_i, z_i) of the feature point of the maintenance component in the visible cone coordinate system O by using the second change matrix, and then the repair is determined by using the visual cone accessibility evaluation model based on the visual cone. The distribution of component feature points in the field of maintenance personnel.

Using the MATLAB software for simulation calculations, each feature point coordinate is assigned to the parametric equation of the maximum field of view of the maintenance personnel. The calculation is available: For all the feature points, Equation(9) is established and the calculation result is shown in Figure 5 . From this,

$$(x_i/0.7002)^2 + (y_i/0.5774)^2 < z_i^2 \quad (9)$$

it can be obtained that the maintenance part has a good sight line accessibility design. Verifying in DELMIA software, as shown in Fig. 5(a), the maintenance part is located within the visual field of the virtual person under the current repair station, which verifies the accuracy of the method.

3.2.2 Visual occlusion barrier evaluation model based on feature points verification

Firstly, the visual parameter equation is constructed based on the data of the virtual human binocular center coordinates and the maintenance component feature point data. Secondly, the AABB axial bounding box for visually obstructing the obstacle is constructed based on the mask feature point data, and the eight surfaces of the bounding box are detected for interference. After calculation, as shown in Fig.5(b), no line of sight of the all feature points of the maintenance component interferes with the bounding box. It can be obtained that the maintenance part is not affected by the visual obstruction at the current viewing angle, and the visual accessibility of the maintenance part is well designed. Verifying in DELMIA software, as shown in Fig. 6, the line of sight of the maintenance component is not blocked by the surrounding environment components at the current viewing angle, which verifies the accuracy of the method in this paper.

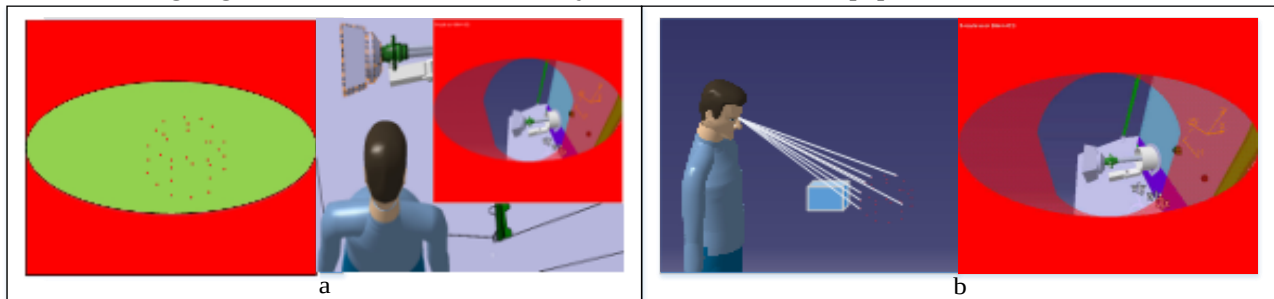


Figure 5.(a) Feature points in maintenance vision distribution and Delmia verification diagram
(b)Visual occlusion evaluation effect and Delmia verification diagram

4. Conclusion

Visibility is a feature that allows the maintenance site to be easily and quickly reached by the design. In order to achieve objective, stereotyped and rapid accessibility analysis and evaluation, this paper constructs a visual cone-based visual line accessibility evaluation model and a visual line interference detection-based visual accessibility assessment model for both visual line reach and visual occlusion. Based on the previous two, a visual automation analysis and evaluation method based on virtual maintenance was proposed. Finally, the feasibility of the method is verified by concrete examples. The disadvantage of this paper is that the AABB axial bounding box is used as the extraction method. This method is simple in structure, simple in collision detection, and requires further research.

References

- [1] Hignett S, Mcatamney L, Rapid entire body assessment (REBA). *Applied Ergonomics*. 2000. 31(2), 201-205
- [2] Abshire KJ, Barron MK, Virtual maintenance real-world applications within virtual environments. *Reliability and Maintainability Symposium 1998*,132-137
- [3] Zhang YT, Ma ZL, Research on visibility analysis method of a refueling vehicle based on CATIA. *Equipment Manufacturing Technology No.6,2010*, 18-19
- [4] Wang ZH, Zhai ZG, Application of DELMIA-Ergonomics in Aircraft Virtual Maintenance. *Mechanical & Electrical Technology(S1)*,2009, 36-53
- [5] Yang CH, Wei J, Yao L, Comprehensive evaluation of complex equipment maintenance accessibility based on quality function deployment and evidence theory. *Computer Integrated Manufacturing Systems*.2009. 15 (11), 2172-2177
- [6] Badler N, Erignac C, Bindiganavale R, Technology for Maintenance Procedure Validation. *Technology for Maintenance Procedure Validation*. 2001

- [7] Xu JX, Li YB, Analysis and study on maintenance visibility based on DELMIA. *Digitization user*. 2013(18),87-89
- [8] Long Y, Yuan J, Zhang ZL, Design and Realization of Virtual Simulation System Based on Stereoscopic Sense of Vision. *Journal of System Simulation*. 2013,274-277
- [9] Li JX, Zheng GL, Research on Application DELMIA System in Aircraft Assembly Simulation . *Aeronautical Manufacturing Technology*. 2008 (11),90-93

Index of Authors

- Abdallah, F., [7](#)
Ahmadini, A.A.H., [1](#)
Akrouche, J., [7](#)
Albano, M., [54](#)
Alberti, A.R., [13](#)
Aljazea, A., [19](#)
Alkali, B.M., [116](#)
Andrews, J., [91](#)
- Becker, M., [54](#)
Bishop, J., [25](#)
Boyd, M.J., [103](#)
Breniere, L., [31](#)
Byrne, T., [172](#)
- Casbeer, D., [66](#)
Castro, I.T., [36](#)
Cavalcante, C.A.V., [13](#), [42](#), [125](#), [131](#), [141](#)
Chatelet, C., [7](#)
Chen, Z., [60](#)
Cheng, L., [147](#)
Coolen, F.P.A., [1](#), [97](#)
- Dantas, P., [125](#)
Dantas, P.P.L., [42](#)
Delmas, A., [48](#)
Di Orio, G., [54](#)
Do, P., [109](#)
Doyen, L., [31](#)
- Farrow, M., [178](#)
Fouladirad, M., [79](#)
- Gaudoin, O., [31](#)
Geng, J., [184](#)
Grall, A., [79](#)
Guo, C., [160](#), [184](#)
- Hajj Chehade, H., [7](#)
- Ingram, T., [172](#)
Iung, B., [109](#)
- Jantunen, E., [54](#)
Jiang, R., [60](#)
- Kalyanam, M., [66](#)
Knight, G., [73](#)
Kollias, S., [85](#)
- Landesa, L., [36](#)
Larrinaga, F., [54](#)
Leontidis, G., [85](#)
Li, L., [166](#)
Liang, Z., [190](#)
Liao, H., [135](#)
Luo, M., [19](#)
Lv, C., [184](#)
- Lyu, C., [160](#)
- Ma, J., [79](#)
Malo, P., [54](#)
McCarthy, M., [25](#)
McGinty, J., [85](#)
Meesit, R., [91](#)
Mei, S., [160](#)
- Najem, A., [97](#)
Naybour, M., [103](#)
Nguyen, H.S.H., [109](#)
Niblett, M., [25](#)
Niculita, O., [116](#)
- Pachter, M., [66](#)
Parlikad, A., [190](#)
Pereira, E., [116](#)
Pohl, E.A., [135](#)
- Rahman, N., [172](#)
Remenyte-Prescott, R., [103](#)
Ribeiro, L., [125](#), [131](#)
Rodriguez, T.F.O., [42](#)
Ruiz, C., [135](#)
- Sallak, M., [7](#), [48](#)
Santos, A.C.J., [141](#)
Scarf, P.A., [13](#)
Schon, W., [48](#)
Serna, A., [36](#)
Silva, A.L.O., [13](#)
Smith, T., [172](#)
Souprayen, C., [153](#)
Stanley, A., [153](#)
Su, C., [147](#)
- Tripathi, A., [153](#)
- Vu, H.C., [109](#)
- Wang, R., [160](#), [184](#)
Wang, X., [166](#)
Warrilow, N., [153](#)
Willoughby, G., [172](#)
Wilson, K.J., [178](#)
Wu, S., [19](#)
- Xiao, Y., [160](#), [184](#)
Xie, M., [166](#), [190](#)
Xie, Y., [195](#)
- Zhao, L., [48](#)
Zhao, X., [190](#)
Zhou, D., [195](#)
Zhou, Q., [195](#)
Zhou, W., [195](#)