Performance Metrics in the Perspective of Prognosis Uncertainty

Bruno P. Leão [1], Takashi Yoneyama[2]

[1] *GE Global Research - Brazil Technology Center, Rio de Janeiro, RJ, 21941-615, Brazil*
leao@ge.com

[2] *Instituto Tecnológico de Aeronáutica, São José dos Campos, SP, 12228-900, Brazil*
takashi@ita.br

ABSTRACT

The subject of uncertainty in failure prognosis, including the importance of estimating and managing it, is a recurring topic in PHM literature. Considering that the prognosis task comprises forecasting, this could not be any different. However, prognosis performance metrics proposed in literature are usually concerned with measuring adherence to requirements, but not the adequate representation of the true uncertainty that arises from various sources in a prognosis problem. This paper presents statistically sound means for evaluating the performance of prognosis methods in the perspective of comparing the true uncertainty to its estimates. This provides a useful yet simple framework for failure prognosis performance evaluation.

1. INTRODUCTION

Failure prognosis is a subject which draws increasing attention from academia and industry over the years. Improvements in computational and sensing capabilities has led to unprecedented possibilities for employing data analytics tools in creating value for asset users and maintainers. Reliable estimates of remaining useful life (RUL) of equipment can yield benefits not only for maintenance but also for logistics, spare parts management and equipment operation. However, various challenges arise during the development and validation of prognosis solutions. Many of these challenges are associated with the intrinsic uncertainty associated with the prognosis task. Prognosis comprises forecasting, and future operational and ambient conditions are usually difficult to estimate in advance. Besides that, the estimation of the degradation state and its trend cannot usually be performed in a reliable way when uncertainty is not properly taken into consideration. Because of these factors, uncertainty is an important topic in failure prognosis literature.

In one of the seminal works that discussed about uncertainty in failure prognosis (Engel, Gilmartin, Bongort, & Hess, 2000) the authors describe the role of uncertainty and the importance of its estimation as part of the prognosis task. The important trade-off between the precision of prognosis estimates and the probability that they will capture the actual failure (sometimes referred to as Engel's paradox) and discussions on the true versus the estimated uncertainty in RUL are also present in the referred paper. Other authors, such as Orchard, Kacprzynski, Goebel, Saha, and Vachtsevanos (2008), discuss about uncertainty in the context of specific prognosis methods. More recently, Celaya, Saxena, and Goebel (2012) recalled the topic of estimated versus true prognosis uncertainty considering also a sample application. In the referred paper, the authors highlight that there is part of prognosis uncertainty that is intrinsic to the problem under consideration and should be properly accounted for and represented. This reinforces the claim that the prognosis task should not be aimed at estimating RUL with minimal uncertainty but rather at making the estimated uncertainty as close as possible to the true one.

Another topic of active research in the failure prognosis field is the definition of proper metrics for performance evaluation. Vachtsevanos, Lewis, Roemer, Hess, and Wu (2006) presented one of the first compiled lists of prognosis performance metrics. This and other seminal works present the metrics with focus on accuracy and precision. More recently, Leão, Yoneyama, Rocha, and Fitzgibon (2008) and Saxena, Celaya, Saha, Saha, and Goebel (2010) proposed more elaborate metrics that consider also other aspects such as design tradeoffs and convergence of estimates. The work presented herein is based on the method described by Leão, Gomes, Galvão, and Yoneyama (2010); Leão and Yoneyama (2011) for prognosis performance evaluation using the Probability Integral Transform (PIT). This approach is presented here as an adequate and statistically sound method for evaluating the quality of prognosis results in terms of fitting the true uncertainty as described in the aforementioned references.

The remainder of this paper is organized as follows: Section 2 presents a review on prognosis performance metrics focused on the evaluation of uncertainty; Section 3 comprises a discussion on PIT-based performance evaluation methods and their adequacy for comparing the estimated uncertainty to the true uncertainty in prognosis problems; Section 4 presents a sample application of the proposed framework based on real data; Section 5 is the conclusion.

2. PROGNOSIS PERFORMANCE METRICS

In order to adequately evaluate results yielded by failure prognosis methods, proper figures of merit must be employed. The definition of adequate performance metrics for prognosis has been, and continues to be, an active research topic. Recent works in the literature (Saxena et al., 2010) have tried to establish an acceptable set of metrics for this purpose.

In the so called offline prognosis performance evaluation, a dataset consisting of a number of run to failure time series is employed for obtaining RUL estimates using the prognosis under evaluation. Vachtsevanos et al. (2006) compiled one of the first sets of metrics intended for prognosis performance evaluation. Early works on metrics have focused on accuracy as a function of the difference between the expected value and the true value of the RUL, and precision as a function of standard deviation or variance of estimates. Such precision related metrics may be considered to be the first proposed means for evaluating the uncertainty associated to failure prognosis.

Improvements on prognosis performance metrics associated to uncertainty evaluation were proposed along the years. For instance Leão et al. (2008) presents metrics and a plot inspired in the receiver operating characteristics (ROC) curve that allow the evaluation of the aforementioned tradeoff represented by Engel's paradox. More recently, Saxena et al. (2010) proposed improvements in accuracy/precision evaluation by the application of more general location and spread measures. Motivation for that comes from the fact that RUL is not necessarily Gaussian. The same work also presents novel metrics to evaluate the convergence of estimates, i.e. the reduction in uncertainty that is expected to occur as the time of failure approaches. Therefore, Saxena et al. (2010) presents better definitions and formalisms for important concepts explored earlier by, for instance, Engel et al. (2000) and Vachtsevanos et al. (2006) concerning uncertainty evaluation.

The aforementioned metrics illustrate the current practice in failure prognosis uncertainty evaluation. Such metrics are useful for comparing methods against requirements or amongst themselves. However, to the best of the authors' knowledge, none of the existing means of performance evaluation can be used to assess how well a method captures true prognosis uncertainty. Such evaluation means can only be used for performing relative assertions taking requirements and other methods as references. The prognosis performance evaluation solution presented in this work can be used to overcome such limitation. Another noticeable aspect of current practice in prognosis performance evaluation is that a large set of very specialized metrics must usually be employed in order to evaluate a prognosis solution. This specialization makes it difficult to have an overall picture of which method performs better than the other. The framework proposed in this paper yields a single metric that quantifies the quality achieved in capturing the underlying uncertainty in the prognosis problem. Therefore, such single metric provides comprehensive information about the overall quality of prognosis solutions.

3. PIT BASED PROGNOSIS PERFORMANCE EVALUATION

Performance metrics described in Section 2 are useful for evaluating the fulfillment of requirements and comparing different prognosis methods. However, none of the referred metrics can be used to evaluate how well the estimated RUL PDFs represent the true uncertainty of the prognosis task under consideration. Direct comparison between estimated and true RUL PDFs is usually not possible, because the latter cannot be obtained. However, recent works in literature (Leão et al., 2010) describe means for indirectly performing this kind of comparison. This may be accomplished through the use of the Probability Integral Transform, which yields a simple, yet comprehensive, means for evaluating prognosis methods. Explanations on how PIT works and how it can be employed for performance evaluation of prognosis results are presented in the following Sections.

3.1. The Probability Integral Transform

The PIT is a method for transforming random variables (RVs) associated to arbitrary probability distributions to a RV with uniform distribution in the range [0,1]. This is accomplished through the use of the cumulative distribution function (CDF). Defining X as a continuous scalar random variable, with $X \sim \pi$, the CDF of X is defined by equation 1:

$$F(x) = \int_{-\infty}^{x} \pi(\xi) d\xi = P(X \leq x) \tag{1}$$

where x is a value in the support of the probability distribution associated to X. Let a new RV Z be defined by equation 2:

$$Z = F(X) \tag{2}$$

This new random variable is uniformly distributed over the range [0,1], i.e. $Z \sim U(0, 1)$. This result was originally presented by Rosenblatt (1952). The same function $F(\cdot)$ can be

used to transform samples, i.e. samples x_i drawn from X can be transformed into samples z_i from Z by means of such function, as presented in equation 3.

$$z_i = F(x_i) \tag{3}$$

The function $F(\cdot)$ is hereafter called PIT. It can be employed for performance assessment in various types of problems where probability distributions are estimated. The following Section presents how this can be accomplished.

3.2. PIT for Performance Evaluation

Literature related to other fields of knowledge presents the application of PIT for performance evaluation. In order to employ this method, the following conditions must hold true:

- the problem under consideration must comprise the estimation of a probability distribution of an outcome;
- measurements of the actual value of this outcome must be available for performance evaluation of proposed solutions.

In the context of performance evaluation of failure prognosis solutions, the outcome of interest is the RUL. Therefore, the prognosis methods under evaluation must yield each RUL estimate in the form of a probability distribution. The actual value of the outcome in this case is the true value of the RUL, which is readily available when run-to-failure datasets are employed for evaluating prognosis performance.

PIT based performance evaluation is a well established method in financial analysis. It was originally proposed by Diebold, Gunther, and Tay (1998) in this context, motivated by the increase on the use of PDF estimates in financial forecasting. State estimation is another field of knowledge where the use of PIT for performance evaluation has been proposed (Chen, Lee, & Mehra, 2007).

In order to use PIT for performance assessment, each estimated PDF ($\hat{\pi}$) yielded by a method under evaluation is used to create a corresponding estimate of the $F(\cdot)$ function. Such $F(\cdot)$ estimates are referred hereafter as $\hat{F}(\cdot)$. $\hat{F}(\cdot)$ can then be obtained from the definition of the CDF as presented in equation 4.

$$\hat{F}(x) = \int_{-\infty}^{x} \hat{\pi}(\xi) d\xi \tag{4}$$

In a failure prognosis context, each estimated PDF is associated to one RUL estimate. Therefore, the probability distribution associated to each RUL estimate yields one $\hat{F}(\cdot)$ function.

Function $\hat{F}_i(\cdot)$ obtained from each estimated PDF is in turn used to transform the corresponding actual measurement of the outcome associated to $\hat{\pi}_i$ as presented in equation 5. Subscripts i were added to $\hat{F}_i(\cdot)$ and $\hat{\pi}_i$ in order to make it clear that a new estimate $\hat{F}_i(\cdot)$ is produced for each estimated PDF $\hat{\pi}_i$. In equation 5, x_i is the actual measurement of the outcome and $\hat{z}_i$ is the value yielded by the transformation. Therefore, application of PIT to a set of m PDF estimates yields a set of $\hat{z}_i$ values, $i = 1, 2, ..., m$.

$$\hat{z}_i = \hat{F}_i(x_i) \tag{5}$$

Concerning performance evaluation of failure prognosis, $x_i, i = 1, 2, ..., m$ are the true RUL values (e.g. one for each run-to-failure datasets). Each $\hat{\pi}_i$ is an RUL estimate in the form of a probability distribution obtained from the prognosis solution under evaluation. Each of these estimates yields a function $\hat{F}_i(\cdot)$ which is in turn used to transform the corresponding true RUL value x_i. In the case where a fixed RUL is considered for the evaluation, then m is equal to the number of run-to-failure datasets available. In such case, m values $\hat{z}_i$ are obtained by transforming true RUL values according to equation 5 using $\hat{F}_i(\cdot)$ functions.

Recalling the definition of PIT described above, if the estimated PDFs adequately represent the uncertainty in the data, then the set of transformed values $\hat{z}_i$ should resemble i.i.d. samples of the distribution $U(0, 1)$. This is the property of PIT that makes it possible to assess how well the method under evaluation could capture the true uncertainty in the problem. The greater the resemblance of the set of $\hat{z}_i$ values to i.i.d. samples drawn from a uniform distribution, the better the proposed PDFs represent the true uncertainty in the data. In a failure prognosis context, if a fixed RUL is considered in the evaluation (e.g. prognosis predictions are performed 10 time/usage units prior to failure for all run-to-failure datasets), then m run-to-failure datasets are associated to m true RUL values x_i, all equal to each other (e.g. all of them equal to 10 in the aforementioned example). m RUL estimates are obtained using the prognosis solution under evaluation, one for each run-to-failure dataset. Therefore, there are m estimates $\hat{\pi}_i$ (each one corresponding to an RUL estimate in the form of a probability distribution) that yield m functions $\hat{F}_i(\cdot)$, which in turn result in the same number of $\hat{z}_i$ values. The following steps summarize the process for obtaining the set of $\hat{z}_i$ values when evaluating performance of a prognosis solution:

1. obtaining m RUL probability distribution estimates $\hat{\pi}_i$ using the prognosis solution under evaluation;
2. using the m RUL probability distribution estimates to create m functions $\hat{F}_i(\cdot)$;
3. using the m functions $\hat{F}_i(\cdot)$ to transform the corresponding ground truth RUL values x_i (if a fixed RUL value is considered, then all functions transform this fixed value, i.e. x_i is the same for every i).

At the end of these steps a set of m values $\hat{z}_i$ is produced. The resemblance of this set of values to samples drawn from $U(0,1)$, according to the PIT result, will indicate how well uncertainty was captured by the prognosis solution. Therefore, one more question must be answered: how can this resemblance be checked? Next Section presents solutions for accomplishing this task.

One final remark about the procedures described in this Section must be made. Although a fixed RUL is referred throughout this Section as an option for performing prognosis evaluation, the proposed method may also be employed with the combination of different RUL values. For instance, instead of evaluating performance at 10 time/usage units prior to failure, performance at 9,10 and 11 time/usage units prior to failure could be considered together. This would yield three times the number of $\hat{z}_i$ values obtained by using a single true RUL value, i.e. m would be three times higher despite the fact that the number of run-to-failure datasets employed for evaluation would be the same. The combination of results obtained using multiple true RUL values is possible since $\hat{F}_i(x_i)$ produces i.i.d. uniform samples regardless of the true RUL value considered, supposing the prognosis solution adequately captures uncertainty in the problem.

3.3. PIT Based Performance Metrics

In order to quantify the resemblance of the $\hat{z}_i$ set to $U(0,1)$ samples, the empirical CDF (ECDF) of $\hat{z}_i$ points may be employed. This ECDF may be compared to the ideal case, which is the actual $U(0,1)$ CDF. Such comparison may be performed using an average point-to-point absolute difference between the ECDF and the ideal CDF. The prognosis quality index (q) is calculated based on this difference, according to equation 6, yielding a value between zero (worst case) and one (best case).

$$q = 1 - \frac{2}{M}\sum_{j=1}^{M} |abs_j - ord_j| \tag{6}$$

In the referred equation, abs_j and ord_j are respectively the abscissa and ordinate of the calculated ECDF points and M is the number of points used for representing the ECDF. The M value is a characteristic of how the ECDF is built and must not be confused with m, which is the number of $\hat{z}_i$ values. It is important to adequately build the ECDF, since all performance evaluation as proposed in this work depends on such curve. One standard means for building the ECDF is using the Kaplan-Meier estimate (Kaplan & Meier, 1958). Using such method, the M value is fixed for a certain set of $\hat{z}_i$ values, i.e. it cannot be freely chosen.

The value 2 in equation 6 was added so that the metric ranges from 0 to 1. The q metric is derived from the PIT related metrics proposed in (Chen et al., 2007). Other means for comparing the resemblance of the $\hat{z}_i$ set to $U(0,1)$ samples may also be employed, including well established statistical methods such as the Kolmogorov-Smirnov test (Papoulis, 1991). However, this form of the q metric is proposed here because it is a simple and intuitive means for quantifying such resemblance.

Besides the q metric, a graphical evaluation of the quality of uncertainty representation in RUL estimates may also be performed using the plot of ECDF curves corresponding to prognosis solutions and the ideal case $U(0,1)$ CDF curve at the same coordinate axis. Figure 1 presents a fictitious sample of such a plot where two ECDF curves, corresponding to two different prognosis solutions, are presented together with the ideal CDF. In this case, visual inspection indicates that the solution associated to ECDF number 2 yields better results than the one associated to ECDF number 1, i.e. ECDF number 2 is closer to the reference $U(0,1)$ CDF. Such a plot will be called prognosis performance plot (PPP) hereafter.

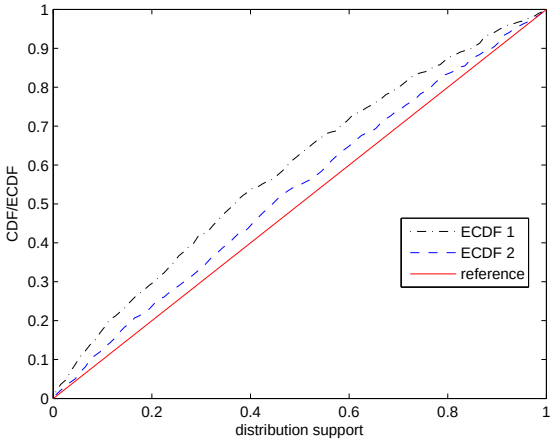

Figure 1. Sample prognosis performance plot.

Figure 2 presents sample PPP curves that correspond to simple anomalies in the estimated RUL PDF. These examples were artificially generated using a Gaussian distribution as the ground truth from which failure times were drawn. This Gaussian distribution represents the true uncertainty associated to the prognosis problem. Gaussian distributions intentionally modified to present anomalies in uncertainty estimation, which in this example correspond to anomalies in standard deviation, were employed as RUL estimates. Such modified distributions correspond to RUL estimates that could be yielded by prognosis methods. It can be noticed from the figure that both over-estimation and under-estimation of uncertainty, which in this case correspond respectively to over-estimation and under-estimation of standard deviation, are penalized in terms of deviation from the ideal CDF curve. It must also be clear that Gaussian curves and standard deviation anomalies were employed in this illustrative exam-

ple only for the sake of simplicity. The proposed PIT based framework can deal with any kind of continuous RUL probability distribution estimates and can capture any kind of anomaly in uncertainty estimation, be it associated to the distribution mean, variance, or even higher order statistical moments.

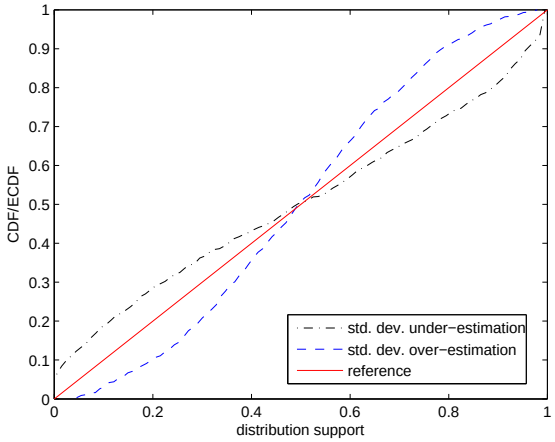

Figure 2. Sample uncertainty related anomalies indicated by PPP curves.

The aforementioned tools (q metric and PPP plot) provide an answer to the question presented in the end of Section 3.2. It is now possible to present additional steps that complement the ones presented in the end of the referred Section, yielding a complete evaluation of prognosis solutions. Recalling the steps presented above, after the end of the third step, a set of m values $\hat{z}_i$ was produced. Two more steps are added:

4. producing an ECDF from the set of $\hat{z}_i$ values by means of a proper method such as Kaplan-Meier estimate;
5. comparing the obtained ECDF to the reference $U(0,1)$ CDF using a suitable means such as the q metric or the PPP plot.

The next Section presents additional tools that may contribute to a more comprehensive evaluation of the performance of failure prognosis solutions. Such tools are based on Hypothesis Testing methods.

3.4. Hypothesis Testing

Checking if the set of transformed points $\hat{z}_i$ can be associated to a RV distributed according to $U(0,1)$ may be interpreted as a hypothesis test (HT). The goal of HT is to evaluate if a certain hypothesis (the null hypothesis) may or may not be rejected in favor of an alternative hypothesis (Papoulis, 1991). In the context of the prognosis problem, the null hypothesis states that the $\hat{z}_i$ set of values can be considered to be i.i.d. samples drawn from $U(0,1)$. Recalling the definitions presented in Section 3.2, this null hypothesis states that the prognosis method under evaluation adequately captures the true uncertainty in the problem. The alternative hypothesis would then state that the method does not adequately capture this uncertainty.

The critical value approach to hypothesis testing (Papoulis, 1991) is presented here as one type of method that may be employed in the context of PIT based prognosis performance evaluation. This approach was originally proposed in this context by Leão and Yoneyama (2011). In order to employ this method, four steps must be followed:

1. defining the null hypothesis and the alternative hypothesis;
2. calculating a test statistic using the sample data;
3. determining the critical value, based on: the probability distribution of the test statistic, supposing the null hypothesis holds; and a user defined significance level, which is the probability of making a Type I error;
4. comparing the test statistic calculated in step 2 to the critical value obtained in step 3.

If the test statistic is more extreme (this may be greater or lower depending on the test statistic) than the critical value, the null hypothesis is rejected. Otherwise, it cannot be rejected.

In the PIT based prognosis performance evaluation context, step 1 in the list was presented above. The test statistic employed here (step 2) is the q metric defined in Section 3.3.

Step 3 requires the probability distribution of the test statistic supposing the null hypothesis holds true. In the PIT context, this means the probability distribution of the q metric when the inputs for its calculation are actual i.i.d. random samples drawn from $U(0,1)$. To the best of the authors' knowledge, no analytical form of such probability distribution exists. However, such distributions may be estimated numerically using Monte Carlo simulation. The shape of these probability distributions depends on the number m of $\hat{z}_i$ estimates (equation 5) available for performance evaluation. The probability distributions required to evaluate critical values need only to be generated once for each value of m, and the same results may be employed for any application.

In order to estimate the probability distribution for a certain value of m using Monte Carlo, first a large number of samples of size m must be drawn from $U(0,1)$. The metric q is then calculated for each sample and the empirical probability distribution of this set of calculated values is used as the probability distribution estimate. For illustration, empirical probability distribution estimates of q were generated for $m = 10, 30, 50, 100, 1000, 10000$. One hundred thousand samples of size m were generated for building each curve. The resulting ECDFs are presented in Figure 3. As can be noticed from the figure, the greater the value of m, the steeper

Table 1. Critical values of q for 5% significance level.

m	**critical values**
10	0.616
30	0.786
50	0.834
100	0.883
1000	0.963
10000	0.989

the slope of the ECDF curve. A dashed line indicating the 5% significance level is also presented in the figure. Other values of significance level could be employed as well. q values obtained from the intersection between the ECDF curves and the significance level line are the critical values. Table 1 presents the corresponding critical values. In this example of 5% significance level, an actual sample drawn from $U(0, 1)$ has 5% probability of yielding a q value lower than the critical values. Recalling the steps for performing the HT, this concludes step 3. Step 4 follows as a simple comparison between the results of the q metric resulting from step 2 and the critical value resulting from step 3. If the q metric is lower than the critical value then the hypothesis that the prognosis method captures the actual uncertainty in the problem is rejected.

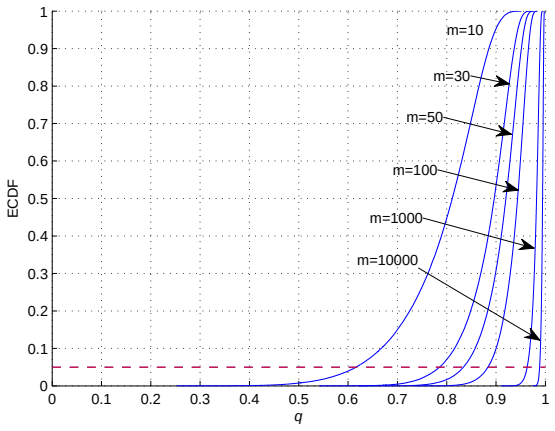

Figure 3. ECDFs for q using different values of m.

The proposed HT approach may be used in the definition of requirements associated to the quality of estimation of true prognosis uncertainty. The results are also useful for understanding the evaluation capability gains yielded by considering additional test datasets in the prognosis performance assessment. This is a unique feature in prognosis performance evaluation which can potentially be employed for evaluating the cost-benefit relation associated to collecting additional data for such assessment.

3.5. Additional Considerations

The use of PIT-based methods for performance evaluation is well suited to failure prognosis when comparing to the aforementioned fields of knowledge where similar approaches have been employed. Some points that corroborate to this fact are the following (Leão et al., 2010):

- Ground truth information required for using PIT (i.e. actual times of equipment failure) is readily available from run-to-failure data. In other fields, such as state estimation, ground truth is usually known only in simulation or laboratory experiments.
- In a prognosis context, the variable of interest (RUL) is one-dimensional. This means that the procedures presented here for PIT calculation can be directly applied. Application of this framework to other fields of knowledge will often require the extension of this analysis to multi-dimensional problems. This commonly translates to extra burden/cost for assessing performance and compromises or limitations to such assessment.

The use of PIT-based prognosis performance evaluation can potentially be even more flexible when the resulting points $\hat{z}_i$ are further transformed using the inverse of a standard Gaussian CDF. This additional step was originally proposed by Berkowitz (2001) in the econometrics literature. Equation 7 presents this transformation, where G^{-1} is the inverse of the CDF corresponding to the standard Gaussian (average equals to zero and variance equals to one) and $\hat{y}_i$ are the resulting transformed points. This additional transformation provides means for employing the set of statistical techniques available for evaluating Gaussian distributions, which is broader than that available for uniform distributions. This is one extension of the PIT-based performance evaluation framework which can be further explored in future work.

$$\hat{y}_i = G^{-1}(\hat{z}_i) \tag{7}$$

Although PIT-based metrics provide a comprehensive way of evaluating prognosis performance, additional metrics can also be used for complementing it, depending on the purpose of the evaluation. For instance, testing a RUL PDF curve estimated based on reliability measures, i.e. based solely on the time of failure of a fleet of components, should provide good results on a PIT based evaluation. This happens because such RUL estimate adequately represents the uncertainty in the time of failure dataset. A precision metric could be employed in combination with the PIT based evaluation when comparing different prognosis algorithms, in order to favor solutions that, besides adequately capturing uncertainty, also present lower dispersion. The q index may also be used, for instance, in combination with convergence metrics such as those presented by Saxena et al. (2010), in order to evaluate how the quality of the estimation of true prognosis uncer-

tainty improves as the time of failure approaches.

The following Section presents a sample application of the proposed evaluation framework to an actual failure prognosis problem.

4. SAMPLE APPLICATION

A sample application based on failure prognosis methods employed for a real world problem is presented here to illustrate the use of the proposed performance evaluation framework. The problem under consideration is the low gas pressure failure in an aircraft crew oxygen system. Forty-two run-to-failure datasets based on aircraft field data were employed for this analysis. Such datasets consist of time series of condition indicators (CIs) associated to the referred failure mode. Each CI data point corresponds to one flight leg. Greater CI values indicate greater degradation. Failure is declared when such a CI exceeds a pre-defined deterministic threshold. Figure 4 present samples of the datasets employed in this study.

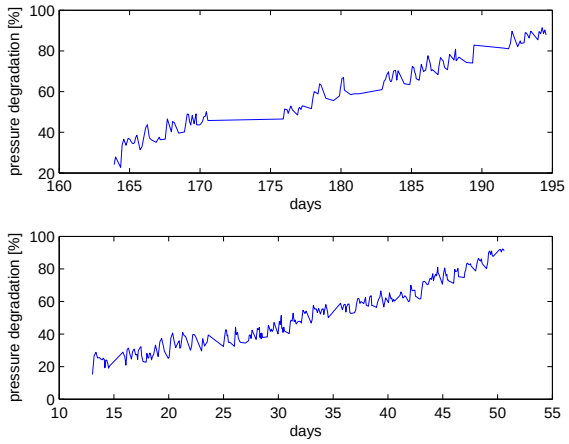

Figure 4. Sample CI datasets employed in the study.

Concerning the prognosis method, a degradation evolution model is employed which is yielded by the analogy between the degradation trend evolution and the tracking of an object's trajectory. The application of this type of model for failure prognosis was originally proposed by Batzel and Swanson (2009). Such model is described by equation 8, where d_k is the degradation value for the k-th flight leg, $\dot{d}_k$ is the degradation rate and $\ddot{d}_k$ is the time derivative of the degradation rate, Δt_k is the time difference between flight leg k and $k-1$, $\mathbf{v}_k$ and q_k are respectively the process noise vector and the measurement noise, y_k is the measurement, which corresponds to the CI value at time instant k.

$$\begin{pmatrix} d_k \\ \dot{d}_k \\ \ddot{d}_k \end{pmatrix} = \begin{pmatrix} 1 & \Delta t_k & 0.5(\Delta t_k)^2 \\ 0 & 1 & \Delta t_k \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} d_{k-1} \\ \dot{d}_{k-1} \\ \ddot{d}_{k-1} \end{pmatrix} + \mathbf{v}_{k-1}$$
$$y_k = d_k + q_k \tag{8}$$

The Kalman Filter was employed for state estimation. Two different methods were employed for yielding RUL probability distribution estimates:

- A Monte Carlo (MC) approach with 500 samples;
- An Unscented Transform (UT) approach, according to Leão and Yoneyama (2011).

One RUL probability distribution estimate was obtained for each run-to-failure dataset, twenty flight legs prior to failure, i.e. a fixed true RUL was considered. Figure 5 presents one sample result of RUL probability distribution estimates yielded by each method for one of the run-to-failure datasets. The normalized histogram yielded by the MC approach is used as an RUL probability distribution estimate. Recalling the performance evaluation framework described above, m in this case is equal to forty-two (forty-two run-to-failure datasets with one prognosis estimate for each dataset) and all $x_i, i = 1, 2, ..., 42$, are equal to twenty. For each of the proposed prognosis methods, RUL probability distribution estimates are used to obtain $\hat{F}_i(\cdot)$ functions. These functions are used to obtain $\hat{z}_i$ values for each method according to equation 9. Therefore, in the end of this process, forty-two $\hat{z}_i$ values are yielded for each method. These values are then employed to build an ECDF for each prognosis method, which in turn is used to calculate a corresponding q value and build a PPP plot.

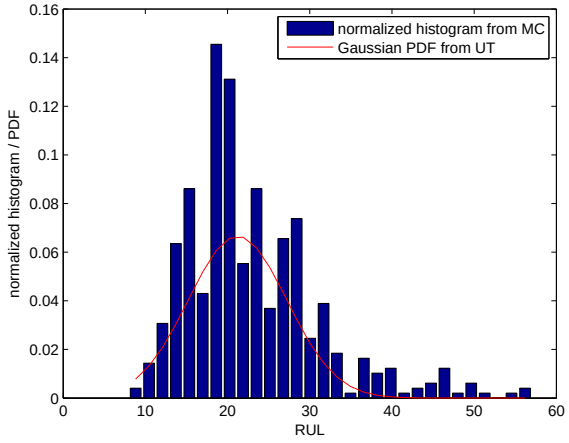

Figure 5. Sample RUL probability distribution estimates obtained from MC and UT approaches for one of the run-to-failure datasets.

Table 2. Metrics for comparing the prognosis methods employed in the sample application.

method	q	avg. abs. error	error std. dev.
MC	0.91	10.4	13.1
UT	0.72	28.4	60.3

$$\hat{z}_i = \hat{F}_i(20), i = 1, 2, ..., 42 \quad (9)$$

Figure 6 presents the PPPs for both methods. Recalling the definitions presented in Section 3.3, the closer the PPP is to the reference, the better. Corresponding q values for the MC and UT approach were respectively 0.91 and 0.72. Table 2 presents these results compared to two standard accuracy and precision metrics: average absolute error and error standard deviation. Error in this case is defined as the difference between the mean of the RUL probability distribution estimate and the true RUL value. Both the q values and the PPP curves indicate a better performance of the MC approach in representing the actual uncertainty in the prognosis problem. Values obtained for accuracy and precision metrics corroborate these results.

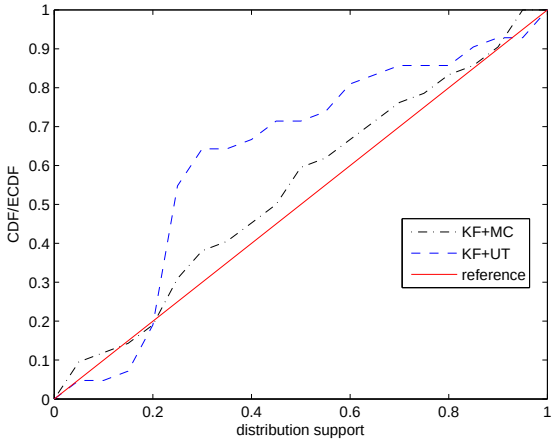

Figure 6. PPP curves for the prognosis algorithms.

5. Conclusion

This paper presented a discussion on how to evaluate the quality of failure prognosis methods in the view of the true prognosis uncertainty. Prognosis performance evaluation literature presents various examples of useful metrics for assessing different aspects of the problem. However, until recently, none of such metrics could answer the following question: "How well does a prognosis method capture the true uncertainty in my problem?" A PIT based performance evaluation framework, originally proposed by Leão et al. (2010); Leão and Yoneyama (2011), is presented here as an adequate means for obtaining such answer. This framework provides ways for assessing the quality of prognosis methods in ways that are not possible by employing other metrics described in literature. The evaluation of how well prognosis methods capture the true uncertainty in a problem represents useful information to the users about how much they can rely on such methods for their decision making.

Related opportunities for future work include the extension of the proposed framework, either by combination with other metrics, such as the convergence metrics proposed by Saxena et al. (2010), or by using new tools such as the inverse Gaussian CDF transformation proposed by Berkowitz (2001). Further applications of PIT-based prognosis performance evaluation to real world problems should also be pursued in order to better demonstrate the usefulness of the proposed approach.

References

Batzel, T. D., & Swanson, D. C. (2009). Prognostic health management of aircraft power generators. *IEEE Transactions on Aerospace and Electronic Systems*, *45*, 473–483.

Berkowitz, J. (2001). Testing density forecasts, with applications to risk management. *Journal of Business and Economic Statistics*, *19*(4), 465-474.

Celaya, J. R., Saxena, A., & Goebel, K. (2012). Uncertainty representation and interpretation in model-based prognostics algorithms based on kalman filter estimation. In *Proceedings of the international conference on prognostics and health management.*

Chen, L., Lee, C., & Mehra, R. K. (2007). How to tell a bad filter through monte carlo simulations. *IEEE Transactions on Automatic Control*(52), 1302-1307.

Diebold, F. X., Gunther, T. A., & Tay, A. S. (1998). Evaluating density forecasts, with applications to financial risk management. *International Economic Review*, *39*(4), 863-883.

Engel, S. J., Gilmartin, B. J., Bongort, K., & Hess, A. (2000). Prognostics, the real issues involved with predicting life remaining. In *Proceedings of ieee aerospace conference.*

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, *53*, 457-481.

Leão, B. P., Gomes, J. P. P., Galvão, R. K. H., & Yoneyama, T. (2010). How to tell the good from the bad in failure prognostics methods. In *Proceedings of ieee aerospace conference.*

Leão, B. P., & Yoneyama, T. (2011). Improvements on the offline performance evaluation of fault prognostics methods. In *Proceedings of ieee aerospace conference.*

Leão, B. P., & Yoneyama, T. (2011). On the use of the unscented transform for failure prognostics. In *Proceedings of ieee aerospace conference.*

Leão, B. P., Yoneyama, T., Rocha, G. C., & Fitzgibon, K. T. (2008). Prognostics performance metrics and their relation to requirements, design, verification and cost-benefit. In *Proceedings of the international conference on prognostics and health management.*

Orchard, M., Kacprzynski, G., Goebel, K., Saha, B., & Vachtsevanos, G. (2008). Advances in uncertainty representation and management for particle filtering applied to prognostics. In *Proceedings of the international conference on prognostics and health management.*

Papoulis, A. (1991). *Probability, random variables, and stochastic processes*. McGraw-Hill.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *The Annals of Mathematical Statistics*, *23*(3), 470-472.

Saxena, A., Celaya, J., Saha, B., Saha, S., & Goebel, K. (2010). Metrics for offline evaluation of prognostic performance. *International Journal of Prognostics and Health Management*, *1*(1).

Vachtsevanos, G., Lewis, F. L., Roemer, M., Hess, A., & Wu, B. (2006). *Intelligent fault diagnosis and prognosis for engineering systems* (1st ed.). Hoboken: John Wiley & Sons, Inc.

BIOGRAPHIES

Bruno P. Leão holds a bachelor's degree in Control and Automation Engineering (2004) from Universidade Federal de Minas Gerais (UFMG), Brazil, a master's degree in Aeronautical Engineering (2007) from Instituto Tecnológico de Aeronáutica (ITA), Brazil, and a D.Sc. degree in Electronics Engineering and Computer Science (2011) also from ITA. He is currently a Lead Scientist with GE Global Research at the Brazil Technology Center in Rio de Janeiro, where he performs research in the fields of Prognosis and Health Management and Air Traffic Management. He was formerly with Embraer S.A. in Brazil from 2005 to 2012. During five years he has been with the PHM research group at Embraer developing innovative PHM solutions for aircraft systems. Before that, he has worked as a Systems Engineer in the fields of Flight Controls and Automatic Flight Controls.

Takashi Yoneyama is a Professor of Control Theory with the Electronic Engineering Department of ITA. He received the bachelor's degree in electronic engineering from Instituto Tecnológico de Aeronáutica (ITA), Brazil, the M.D. degree in medicine from Universidade de Taubaté, Brazil, and the Ph.D. degree in electrical engineering from the University of London, U.K. (1983). He has more than 250 published papers, has written four books, and has supervised more than 50 theses. His research is concerned mainly with stochastic optimal control theory. Prof. Yoneyama served as the President of the Brazilian Automatics Society in the period 2004-2006.