Enhanced Trajectory Based Similarity Prediction with Uncertainty Quantification

Jack Lam[1], Shankar Sankararaman[2], and Bryan Stewart[3]

[1,3]*Naval Surface Warfare Center Port Hueneme Division, Port Hueneme, CA, 93043, USA*

jack.lam@navy.mil
bryan.stewart2@navy.mil

[2]*SGT Inc., NASA Ames Research Center, Moffett Field, CA 94035, USA*

shankar.sankararaman@nasa.gov

ABSTRACT

Today, data driven prognostics acquires historic data to generate degradation path and estimate the Remaining Useful Life (RUL) of a system. A successful methodology, Trajectory Similarity Based Prediction (TSBP) that details the process of predicting the system RUL and evaluating the performance metrics of the estimate was proposed in 2008. Two essential components of TSBP identified for potential improvement include 1) a distance or similarity measure that is capable of determining which degradation model the testing data is most similar to and 2) computation of uncertainty in the remaining useful life prediction, instead of a point estimate. In this paper, the Trajectory Based Similarity Prediction approach is evaluated to include Similarity Linear Regression (SLR) based on Pearson Correlation and Dynamic Time Warping (DTW) for determining the degradation models that are most similar to the testing data. A computational approach for uncertainty quantification is implemented using the principle of weighted kernel density estimation in order to quantify the uncertainty in the remaining useful life prediction. The revised approach is measured against the same dataset and performance metrics evaluation method used in the original TBSP approach. The result is documented and discussed in the paper. Future research is expected to augment TSBP methodology with higher accuracy and stronger anticipation of uncertainty quantification.

1. INTRODUCTION

Data driven prognostics acquires historic data to generate degradation path and estimate the Remaining Useful Life (RUL) of a system. In 2008, a new approached known as the Trajectory Similarity Based Prediction (TSBP) methodology was proposed in (Wang T. , 2013), and was successfully demonstrated during the NASA AMES 2008 Prognostics Health Management (PHM) challenge by obtaining the highest score by using a data-driven prognostics method to predict the RUL of a turbofan engine (Saxena & Goebel, PHM08 Challenge Data Description, 2008). While the TSBP is a proven technique, (Wang T. , 2013) does not address imbalanced data (Gouriveau, Ramasso, & Zerhouni, 2013), the effectiveness of different dissimilarity measure (Giusti, 2013), and uncertainty of the model (Dallachiesa, Nushi, Mirylenka, & Palpanas, 2012). These considerations are required to minimize the variation that exists in the data driven prognostics method, and systematically quantify the uncertainty in the RUL prediction.

In (Wang T. , 2013), the author developed a novel RUL prediction method based on the Instance Based Learning methodology called TSBP. In TSBP, the historical instances of a system with life-time condition data and known failure time from the training data are used to create a library of degradation models; these models are then compared against the testing data in order to compute a similarity measure and predict an RUL corresponding to each of the degradation models. The final RUL estimate can be obtained by aggregating the multiple RUL estimates using a density estimation method. While (Wang T. , 2013) focused on the basic TSBP methodology, there are still several areas for improvement.

For example, in (Yu, Yong, Datong, & Xiyuan, 2012), the authors investigated sensor selection as a critical research topic for prognostics. In their research, the authors stated that inclusion of irrelevant or redundant variables during data fusion may lead to over-fitting or less sensitivity of prognostics model, which would lead to adverse prediction performance.

In (Guo, Gerokostopoulos, Liao, & Niu, 2013), the authors proposed to incorporate degradation initiation time into the general degradation path modeling. Their paper argued that there is a “degradation free” period, i.e., degradation starts only after an initiation time and that a product failure is a combined effect of the initiation time and the degradation growth. In (Gouriveau, Ramasso, & Zerhouni, 2013), the authors suggested the need to deal with 1) data whose relative number of instances in each class evolves with time and 2) data whose significance is not known by the user. In (Giusti, 2013)and (Otey & Parthasarathy, 2004), both authors examined the notion of quantifying the dissimilarity between different multivariate time series. Their argument suggested that calculating the Euclidean distance between the centroids of two data sets is ineffective because it ignores the correlations present in the data sets. Finally, in (Dallachiesa, Nushi, Mirylenka, & Palpanas, 2012), the authors summarized the uncertainty in time series and suggested two main approaches to model these uncertain time series. Given all these factors, it can be easily seen that the TSBP method proposed in (Wang T. , 2013) can be reviewed and be improved.

In (Lei & Govindaraju, 2004), authors proposed the use of Simple Linear Regression (SLR) as a similarity measure technique for on-line signature recognition applications in comparison with the traditional approach of computing the Euclidean Distance, while having lower time complexity (O(n)) than Dynamic Time Warping (DTW) (O(n^2)). The SLR method utilized the mean-deviation normalization to circumvent the problem of scaling and shifting, which, in general, impacts the performance of the DTW method. Further, SLR can be adapted to multi-dimensional sequences, where most real-life applications are relevant.

In this paper, we examine the use of SLR and DTW within the TSBP method for similarity prediction and address the various shortcomings of the original TBSP approach that were explained in the previous paragraphs. Further, we test the result on the original dataset (Saxena & Goebel, 2008) and use the original performance evaluation metrics (Saxena, Celaya, Saha, Saha, & Goebel, 2009) against the original TBSP approach described by (Wang T. , 2013). We also compare the results using different density estimation approaches. The TSBP method with SLR and DTW as the similarity measure with the use of the kernel density estimation provide us with more insight into the problem.

The motivation for this work is to improve further the TSBP method by incorporating different similarity measures and develop a better understanding for uncertainty qualification. Although more work is needed to compare the results of TBSP methodology against the state-of-the-art data driven technique used by the industry, our study produced a survey of related areas that can be experimented to serve as an improved TSBP method. The target application is highly complex systems where physical modeling will be difficult and state of the operating condition can be observed. In this case, TSBP method can generate different degradation models against each regime from the different operating condition to generate an aggregation of RUL estimation. Unlike (Wang T. , 2013), this paper 1) anticipates imbalanced data, 2) evaluates the SLR and DTW similarity measures, and 3) incorporates the uncertainty modeling done in (Dallachiesa, Nushi, Mirylenka, & Palpanas, 2012). These capabilities further support the practical feasibility of the proposed method used in real applications. We envision more interest and study in the TBSP approach will drive academic community and industry into maturing the methodology to provide more accurate RUL estimation.

The rest of this paper is organized as follows. In Section 2, we review the multi-regime partitioning and normalization method used in (Wang T. , 2013). In Section 3, we briefly review the techniques for degradation modeling explained in (Wang & Coit, 2007) and (Guo, Gerokostopoulos, Liao, & Niu, 2013). In Section 4, we describe the similarity/dissimilarity measure used in (Dallachiesa, Nushi, Mirylenka, & Palpanas, 2012), (Yu, Yong, Datong, & Xiyuan, 2012), (Giusti, 2013), (Otey & Parthasarathy, 2004), and (Lei & Govindaraju, 2004). In Section 5, we describe uncertainty quantification in RUL estimation and review the density estimation methods. In Section 6, we include the discussion of the performance metrics described in (Saxena, et al., 2008). In Section 7, we review the dataset (Saxena & Goebel, 2008) and describe the procedures for the experiment. In Section 8 and 9 we present results and findings then conclude the paper in Section 10.

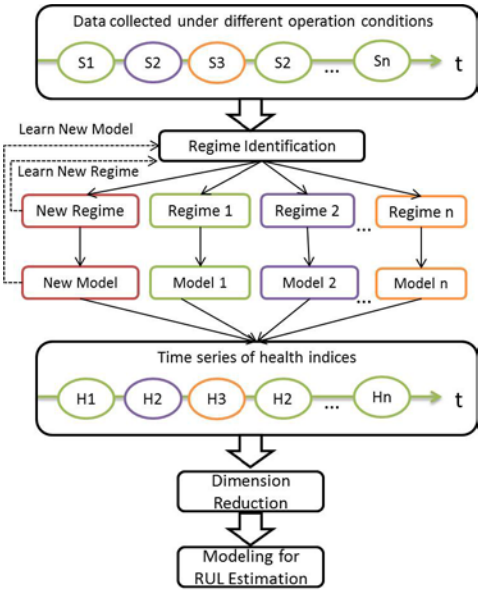

Figure 1. Process for multi-regime health assessment.

2. MULTI-REGIME PARTITIONING AND NORMALIZATION

When a system is operating under multiple operating conditions, the sensor measurements can behave differently in those unique environments, thereby causing difficulty in identifying failure trends. It is beneficial to identify the unique operating conditions or regimes from which sensors can be normalized or features can be extracted. Figure 1shows the high level process for multi-regime health assessment.

To illustrate multi-regime partitioning, the “Turbofan Engine Degradation simulation” data set from (Saxena & Goebel, PHM08 Challenge Data Description, 2008) will be examined. Within this data set, there are 21 sensor measurements and three other measurements that describe the operational conditions the system was operated under. The operating conditions change for each measurement (cycle). Figure 2 shows a select number of sensor measurements for the life time of one particular system.

2.1. Regime Identification

The first step in the process for multi-regime health assessment is to identify the unique, non-overlapping regimes. In this paper, multiple regimes are found using *k*-means clustering. The *k*-means clustering algorithm finds the optimum number of clusters, *k*, where each observation belongs to the nearest cluster’s mean, hence the name *k*-means. Figure 3 shows the results of the *k*-means clustering algorithm on the “Turbofan Engine Degradation simulation” data set. As seen in Figure 3, the data was found to have 6 nicely separated and non-overlapping regimes.

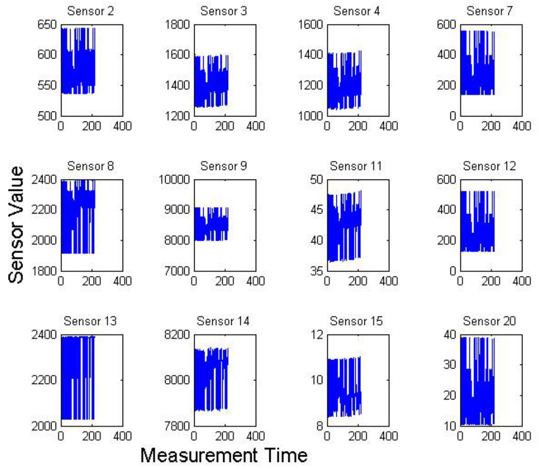

Figure 2. sensor measurement from “Turbofan Engine Degradation simulation” data set.

2.2. Mean-Variance Normalization

The next step is to normalize the sensor data according to the regime the measurement was taken under. This is done by performing mean-variance normalization. Similar to Eq. (1) where p represents the regime the sensor measurement belongs at time instance i.

$$y_i = \frac{{x_i}^p - \mu^p}{\sigma^p} \tag{1}$$

The mean-variance normalized data becomes the time series health indices as depicted in Figure 1. In continuation of the illustration, the progression from Figure 2 to Figure 4 shows a more revealing portrayal of the system behavior once the operating conditions are taken into consideration.

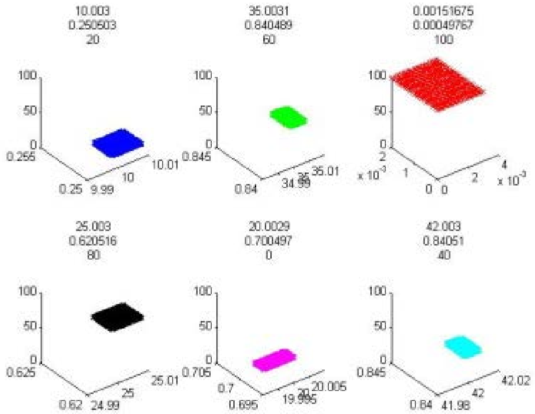

Figure 3. Multi-regime partitioning of the “Turbofan Engine Degradation simulation” data set. This figure represents all the operational condition that was performed.

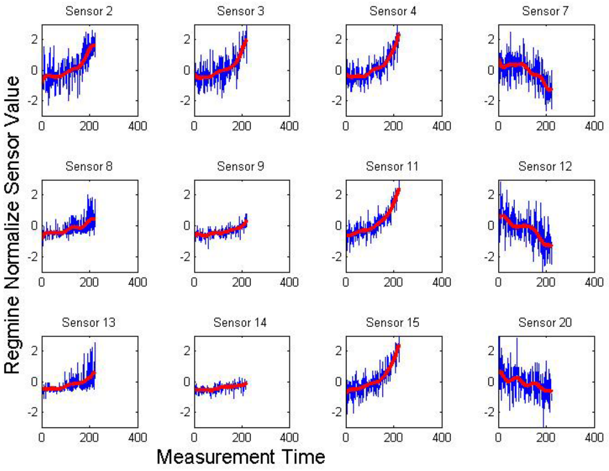

Figure 4. Mean variance normalized data (blue line) from the “Turbofan Engine Degradation simulation” data set for a single system or unit. The red line shows the degradation model for each sensor.

2.3. Variable Weighting and Dimensionality Reduction

At this point, the system data has been prepared and normalized for training the degradation models. However, there are additional techniques that can be used to further emphasize and refine the data to produce more accurate and timely results. Variable/feature weighting is used to emphasis certain sensor measurements over other variable/features and is often used in the feature selection process. In (Wang T. , 2013), an Empirical Signal-to-Noise Ratio (eSNR) is used for variable relevance evaluation. The eSNR is defined as

$$eSNR(s_i) = \frac{var(\underline{s_i})}{var(s_i)} \tag{2}$$

where s_i is a one dimensional time series representing the features of the system evolving over time. Let $\underline{s_i}$ be a smoothed version of s_i filtered by a certain filtering or smoothing algorithm. The idea is that, in the event the global variance (variance of the entire time series) is highly correlated to the local variance (variance within a shorter period of the time series), the smoothed time series will have a much smaller variance compared to the original. Therefore, the feature selection or emphasis can be performed from the ranking of the eSNR. The feature weighting is

$$\tilde{y}_n = y_n \cdot eSNR(y_n) \tag{3}$$

where n represents the n^th feature. This approach effectively de-emphasizes the features with large local variance.

Once the feature has been weighted, the next step is to uncorrelated the features. In this case, (Wang T. , 2013) suggests the use of Principal Component Analysis (PCA). PCA is a common technique used to transform the features into a smaller set of uncorrelated features. The uncorrelated feature will contain minimum redundancy and is important to combat the so-called curse of dimensionality. The method transforms the data into another coordinate system where the first coordinate or principal component (PC) represents the direction of the greatest variance of the original data with the second, third, etc. PC represents decreasing variance of the original data. The transformed features are calculated as

$$z = V_M^T \cdot (\tilde{y} - \overline{\tilde{y}}) \tag{4}$$

Where $\overline{\tilde{y}}$ is the mean of $\tilde{y}$, and V_M consist of the eigenvectors from the covariance matrix of $\tilde{y}$. The top M principal components that make up 90% of the total variance are retained. The resultant PCs form a new time series z for each training and testing instance. An example of variable weighting and dimensionality reduction of the original data can be seen in Figure 5. With the PCA completed, the original data is now ready for Degradation Trajectory Abstraction. The data is Figure 5 show how the system is degrading through time with the red line showing the degradation trajectory abstraction model discussed in the following section.

3. DEGRADATION MODELING/REGRESSION

The degradation models are built from the M Principal Components (PC) extracted from the normalized data as described in Section 2. These models describe the PCs of z as a function of time t:

$$G: z = {}^{l}g(t) + \varepsilon, 0 \le t \le {}^{l}t_I \tag{5}$$

where ε is the noise term and in many cases is modeled as Gaussian. (Wang, 2010)

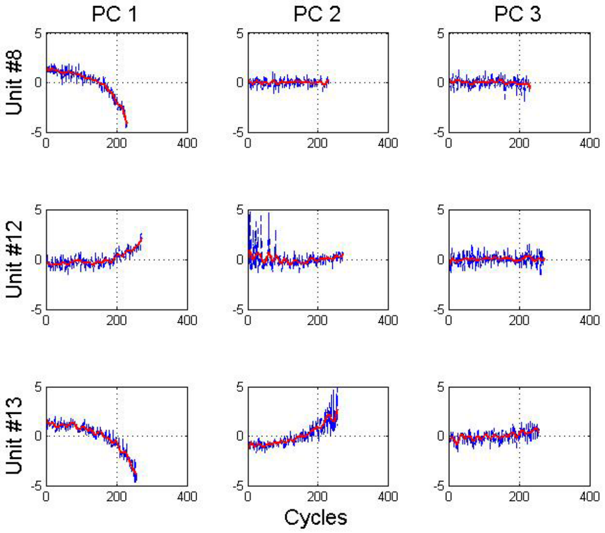

Figure 5. Example Trajectory Abstraction model from the "Turbofan Engine Degradation simulation" data set. The blue line is the variable weighting and dimensionality reduction of the original data, z. The red line is the degradation trajectory abstraction models.

There are many parametric and non-parametric methods that can be used to build the degradation models, all of which should be considered based on their ability to address the global degradation pattern, short-period characteristics, amount of available data, data noise level, and many other influential system characteristics. For this type of RUL estimation, long-term degradation behavior and the operating setting of the system are important, whereas the local fluctuations in the degradation trajectory can largely be considered noise. For these types of applications a smoothing operation of the time series such as a linear interpolation can be used. In (Wang, 2010), an exponential curve fitting, moving average filter and interpolation, Kernel regression smoothing, and relevance vector machines were explored.

Based on the results found in (Wang, 2010) the kernel regression smoothing approach was used for degradation Trajectory Abstraction in this paper; see Eq. (6)-(7).

$$z(t) = \frac{\sum_{i=1}^{E} K_G(t, t_i) \cdot z_i}{\sum_{i=1}^{E} K_G(t, t_i)} \tag{6}$$

$$K_G(x, y) = exp\left(\frac{\|x - y\|^2}{2\rho^2}\right) \tag{7}$$

Where ρ is the kernel width and is a free parameter usually chosen based on the data. An example output is shown in Figure 5 as the red line.

4. REMAINING USEFUL LIFE ESTIMATION

Once all the models have been trained, the testing data will need to be compared to every model and a similarity measure computed. The similarity measure is used to determine which model the system under test is most similar too. This can be done by computing a distance or similarity measure. In (Wang T. , 2013), the Minimum Euclidean Distance with Degradation Acceleration (MED-DA), Minimum Euclidean Distance with Time Lag (MED-TL), and Minimum Euclidean Distance with Time Lag and Degradation Acceleration (MED-TL-DA) was proposed. It was found that the MED-DA performed the best on the CMAPSS dataset evaluated. The remaining of this section, we briefly review MED-DA distance measure and provide an overview of two new similarity/distance measures we propose in this paper: Pearson's Correlation and Dynamic Time Warping.

4.1. Minimum Euclidean Distance with Degradation Acceleration

In (Wang T. , 2013), the Minimum Euclidean Distance with Degradation Acceleration (MED-DA) is the same as computing the Minimum Euclidean Distance between the training and testing models except the MED-DA uses a scaling factor for time dilation. This scaling factor is to accommodate the degradation rate differences between testing and training systems.

$${}^{l}D^2(\lambda) \coloneqq \frac{max(\lambda, \frac{1}{\lambda})}{I} \sum_{i=1}^{I} \sum_{m=1}^{M} \frac{(z_{mi} - {}^{l}g(\lambda \cdot t_i))^2}{2\sigma_m^2} \tag{8}$$

where $\max(\lambda, 1/\lambda)$ is the pentalty term for the difference in degradation rate.

The RUL prediction using this distance measure is calculated as:

$${}^{l}r_I = \frac{{}^{l}t_E}{\underset{\lambda}{arg\,min}\ D^2(\lambda)} - t_I \tag{9}$$

Additionally, in (Wang T. , 2013) it was assumed that the most recent cycles provided more value to the similarity measure than the earlier cycles. Therefore (Wang T. , 2013) used a non-uniform weighting scheme to emphasis the most recent cycles of the system under test. Eq. (8) then becomes

$${}^{l}{D_{DA}}^2(\lambda) \coloneqq \frac{max(\lambda, \frac{1}{\lambda})}{I \sum_{i=1}^{I} \upsilon_i} \sum_{i=1}^{I} \upsilon_i \left(\sum_{m=1}^{M} \frac{(z_{mi} - {}^{l}g(\lambda \cdot t_i))^2}{2\sigma_m^2} \right) \tag{10}$$

where υ_i is the non-uniform weighting of each cycle i.

$$\begin{aligned} \upsilon_i &= exp\left(-\frac{(t_i - t)^2}{2\rho^2}\right) \\ \rho &= \gamma \cdot {}^{l}r_E \end{aligned} \tag{11}$$

The non-uniform weighting is controlled by the spread parameter which is a percentage of the life ${}^{l}r_E$ of the degradation model ${}^{l}G$ and is controlled by the spread ratioγ. In (Wang T. , 2013), through cross-valuation, a spread parameter of 0.3 was found to produce the best results.

Since MED-DA is a squared distance measure, a similarity measure is computed as follows:

$${}^{l}S^{DA} = exp(-{}^{l}{D_{DA}}^2) \tag{12}$$

In (Wang T. , 2013), the best reported performance score on the evaluation set was 0.7534. This score is based the optimum values for the kernel width parameter ρ used for the kernel regression smoothing and spread ratio γ used in the MED-DA similarity evaluation. The optimum parameters were found by a 5-fold cross-validation of the training set where $\rho = 7$ and $\gamma = 0.3$.

4.2. Similarity based on Pearson's correlation

In (Lei & Govindaraju, 2004), a simple linear regression was used to assess the strength of a linear relationship between sequences $X = (x_1, x_2, \ldots, x_n)$ and $Y = (y_1, y_2, \ldots, y_n)$. A goodness-of-fit measures call $R^2$was used and is defined as:

$$R^2 = \frac{\sum_{i=1}^{n} u_i^2}{\sum_{i}^{n} (y_i - \overline{Y})} \tag{13}$$

where u is the error term and $\overline{Y}$ is the mean of Y. R^2 is also called the coefficient of determination. It is interpreted as the fraction of the variation in Y that is explained by X. After further evaluation it is found that R^2 is exactly the square of Pearson's correlation (Lei & Govindaraju, 2004).

$$S^{SLR} = r = \frac{\sum_{i}^{n} (x_i - \overline{X})(y_i - \overline{Y})}{\sqrt{\sum_{i}^{n} (x_i - \overline{X})}^2 \sqrt{\sum_{i}^{n} (y_i - \overline{Y})^2}} \tag{14}$$

As r approaches 1, the linear relation between the two sequences becomes stronger. Therefore the Pearson's correlation of X and Y will have similarity r.

The RUL prediction using this similarity measure is a direct calculation between the test system and the model with the highest Pearson's correlation.

$${}^{l}r_I = {}^{l}t_E - t_I \tag{15}$$

4.3. Similarity based on Dynamic Time Warping

Dynamic Time Warping (DTW) is an alternative approach to determine the distance between two time-series signals where the two temporal sequences may vary in time or speed. It attempts to match two time series by "stretching" and "contracting" subsequences of the series so the difference between the series is minimized. (Giusti, 2013) The distance is then measured as the square root of the sum of the differences between the matched observations.

Technically, DTW (Salvador & Chan, 2007) constructs a warp path between the two time series. A dynamic programming approach is first used to find the warp path and create a cost matrix. A single point in the original time series can be warped to multiple points in the comparing time series. Every cell of the cost matrix is filled and the minimum-distance warp path can be evaluated by reversely following the smallest cost of each move until the original point is reached. If both series were identical, the warp path through the matrix would along the diagonal.

DTW can also adapt a constrained version by incorporating a window size parameter. This parameter limits the number of observations a matching can occur ahead or behind any given observation. It is noted in (Giusti, 2013) that the constrained version may sometimes improve the classification accuracy by avoiding pathological warping.

The RUL prediction using this similarity measure is a direct calculation between the test system and the model with the highest Pearson's correlation.

$$ {}^{l}r_I = {}^{l}t_E - t_I \qquad (16) $$

Since DTW is a squared distance measure, a similarity measure is computed as follows:

$$ {}^{l}S^{DTW} = exp(-{}^{l}D_{DTW}{}^{2}) \qquad (17) $$

4.4. Model Aggregation

All RUL estimates and similarity scores are used to form a hypothesis set and the goal of model aggregation is to use multiple estimates in the hypothesis set and sum them up to create a final prediction. The simplest method of aggregation is to use the similarity-weighted sum, which provides a Point Estimate of the RUL.

$$ r_I := \frac{\sum_{l=1}^{L} {}^{l}S_I \cdot {}^{l}r_I}{\sum_{l=1}^{L} {}^{l}S_I} \qquad (18) $$

This approach is inadequate for uncertainty management in prognostics. A probability distribution or confidence interval for the predicted RUL is desired in order to aid risk-informed decision-making in the context of prognostics and health management. (Wang, 2010)

5. UNCERTAINTY QUANTIFICATION IN RUL PREDICTION

The computation of uncertainty in the remaining useful life prediction is an important, essential, and challenging issue. Since prognostics deals with the prediction of the future behavior of engineering systems, it is necessary to understand that it is almost impossible to make predictions regarding the future. That is why it is important to quantify the various sources of uncertainty in prognostics and quantify their combined effect on the remaining useful life prediction.

Some recent research efforts in (Sankararaman, Daigle, & Goebel, 2014) and (Sankararaman & Goebel, 2013) have been focusing on the topic of quantifying the uncertainty in prognostics and the remaining useful life prediction. At any given instant of time at which prediction needs to be performed, the uncertainty in the RUL prediction depends on three important factors:

- Health state estimate at the time of prediction (initial state)
- Future operating and loading conditions
- Degradation model that predicts health state degradation from the initial state, based on the future operating and loading conditions

It has been demonstrated that the computation of the uncertainty in the RUL, based on the uncertainty in the above quantities is a non-trivial problem and needs to be solved using statistical methods (Sankararaman, 2014). In this context, the goal is to calculate the probability distribution of the remaining useful life prediction continuously as a function of time; note that this probability distribution varies as a function of time and therefore, needs to be recalculated at every time instant. This probability distribution needs to systematically account for the different sources of uncertainty in the aforementioned list of quantities and quantify their combined effect on prognostics and remaining useful life prediction.

Most of the previous efforts have focused on such uncertainty quantification only in the context of model-based prognostics where physics-based models are used to represent health state degradation. Uncertainty quantification and management in the context of data-driven prognostics has not been studied in the detail, and since, different types of data-driven techniques have been used by several researchers, the interpretation, quantification, and management of uncertainty may be different for different data-driven approaches. Hence, uncertainty quantification needs to be discussed in the context of the data-driven approach being pursued, and hence, this paper focuses only on uncertainty quantification in the TBSP approach.

5.1. Uncertainty in Similarity-Based Prediction Technique

In the context of similarity-based prediction, it is first essential to understand the importance of uncertainty quantification. In this methodology, the focus is on finding out the similarity between the desired testing data set and the entire training data set. The remaining useful life of the testing data set can be predicted through some sort of meaningful “interpolation” in the domain of the training data set, where the interpolation procedure attempts to identify where the testing data set lies, with respect to the training data set. An important underlying assumption here is that, at any point of prediction, the future operating conditions and loading conditions in the testing data set can also be interpolated based on that of the training data set; in many practical applications, this assumption may be incorrect and therefore, this method may not be applicable.

Therefore, if there is exact similarity between a testing data set and a particular training data set, then there is no uncertainty regarding the prediction of remaining useful life. This is because the remaining useful life of the desired testing data set is equal to the remaining useful life of the corresponding training data set. This can be easily explained by understanding data-driven learning algorithms such as Gaussian process learning where the variance of the prediction at any training point is exactly equal to zero. Therefore, if the testing point is identical to a training point, the variance of the prediction is zero and hence, there is no uncertainty regarding the remaining useful life. (Note that, the similarity-based comparison is performed only until the time of prediction. There may be significant differences between the testing set and the training set after the time of prediction; such differences lead to uncertainty in the remaining useful life prediction but cannot be quantified without knowledge regarding the future operating/loading conditions of the testing data set.)

Typically, the testing data set may be significantly different from the training data set, and the TBSP approach computes a similarity between the training and testing data set. This similarity measure is simply reflective of the probabilistic weightage that is given to each of the remaining useful life values of the training data set. Therefore, Eq. (18) implies that the remaining useful life is calculated only using a weighted averaging approach, and therefore, is reflective only of the mean behavior. Other statistics of the remaining useful life prediction can also be calculated. For example, the standard deviation can be calculated as:

$$\sigma_r = \sqrt{\frac{\sum_{l=1}^{L'} {}^{l}S_I ({}^{l}r_I - r_I)^2}{\sum_{l=1}^{L'} {}^{l}S_I}} \sqrt{\frac{L'}{L'-1}} \qquad (19)$$

where L' denotes the number of non-zero similarity measures.

Note that the weighted mean and weighted standard deviation are central measures. While such central measures are important, they do not sufficiently capture the information regarding the uncertainty in the remaining useful life prediction. In order to achieve this goal, it is necessary to calculate the entire probability distribution (either in terms of the probability density function or in terms of the cumulative distribution function). This calculation is facilitated through the use of kernel density estimation, as explained later in this section.

5.2. Uncertainty Quantification through Maximum Likelihood Estimation

In (Fonseca, Friswell, Mottershead, Lees, & Adhikari, 2005), the authors describe that the key to the maximum likelihood (ML) approach is to parameterize the probability density functions (PDFs) of the parameters. The uncertainty quantification includes calculating the probability that the measurements occur given the PDF of the parameters.

Suppose that the physical parameters, x, follow a certain probability distribution belonging to a probability distribution family parameterized by Θ (for example the mean, μ, and covariance matrix, Σ). For a given Θ, the output PDF, $f(x|\Theta)$, can be approximated using the uncertainty propagation method. Let the measurements be x_1, x_2, …, x_N. The measurements are assumed to be independent, therefore the measurements likelihood is

$$L(\Theta) = f(x_1, x_2, \ldots, x_N|\Theta) = \prod_{i=1}^{N} f(x_i|\Theta) \qquad (20)$$

The maximum likelihood estimator is value of Θ that corresponds to the maximum of $L(\Theta)$. Note that the maximum likelihood estimate is also a central measure.

Two important changes need to be made in order to adapt this methodology for the purpose of uncertainty quantification in TBSP. First, it is necessary to infer information regarding the uncertainty; such uncertainty can be expressed either in terms of the PDF $f(x)$ or in terms of confidence intervals. Secondly, and more importantly, the PDF $f(x|\Theta)$ corresponds a parametric probability distribution (with parameters Θ), and such a distribution may not be available. So, it may be necessary to use non-parametric distribution and directly estimate the PDF $f(x)$ without employing the use parameters Θ. In this paper, both of these goals are accomplished through the use of a weighted kernel density function that is not only parametric but also can directly compute confidence intervals on the quantity of interest, x in this case.

5.3. Uncertainty Quantification through Kernel Density Estimation

A non-parametric approach for model aggregation is used which is called Kernel Density Estimation or KDE using a

Parzen window method. (Wang T. , 2013) The kernel density approximation is given by:

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x - x_i}{h}\right) \tag{21}$$

where K is the Gaussian kernel function and h is the bandwidth for density estimation. The Gaussian kernel function is defined as:

$$K(u) = \frac{1}{2\sqrt{\pi}}exp\left(\frac{u^2}{2}\right) \tag{22}$$

In (Wang T. , 2013) and in this paper, the KDE method via diffusion with automatic bandwidth selection as proposed in (Botev, Grotowski, & Kroese, 2010) was used.

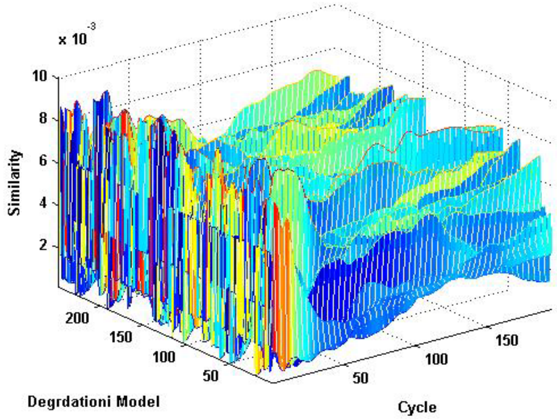

Figure 6. Example of SLR similarity between testing data and all degradation model for each cycle of the test system.

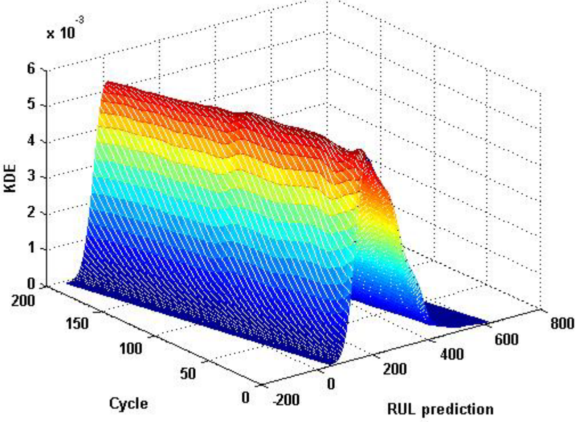

Figure 7. Kernel Density Estimation approach for RUL prediction using model aggregation.

Figure 6 shows an example of the similarity between test data and the trained models for over 200+ cycles. As can be seen in Figure 6, at the beginning the testing unit is very similar to all the degradation models, however as time (cycles) progresses the most similar degradation models can be readily observed. The plot in Figure 7 shows the density estimation of the RUL prediction at each cycle based on the SLR weighted KDE model aggregation.

6. Performance Metrics

The evaluation of the proposed enhancements to TBSP will be based on the work in (Saxena, Celaya, Saha, Saha, & Goebel, 2009): Prediction Horizon, Rate of Acceptable Predictions, Relative Accuracy, and Convergence. A brief description of the metric will be provided in this section but the reader is referred to (Saxena, Celaya, Saha, Saha, & Goebel, 2009) and (Wang T. , 2013) for further information.

6.1. Prediction Horizon

Prediction Horizon (PH) is the time difference between the EoL failure and the time from which the RUL prediction first met the specified performance criteria,i.

$$PH = t_E - t_{i_a} \tag{23}$$

6.2. Rate of Acceptable Predictions

This metric quantifies the prediction quality. This is done by determining whether the prediction falls within a specified percentage of the true RUL for each RUL prediction.

$$AP = Mean(\{\delta_i | t_H \leq t_i \leq t_{EoUP}\}) \tag{24}$$

The specified percentage can be thought of as a cone of accuracy since as the true RUL decreases the accuracy requirement for the prediction become more stringent.

$$\delta_i = \begin{cases} 1 \quad if(1-\propto)r_i^* \leq r_i \leq (1+\propto)r_i^* \\ \quad 0 \quad Otherwise \end{cases} \tag{25}$$

$$\delta_i = \begin{cases} 1 \quad \int_{r_i^* - \propto \cdot t_E}^{r_i^* + \propto \cdot t_E} \pi(r_i)dr_i \geq \beta \\ \quad 0 \quad Otherwise \end{cases} \tag{26}$$

6.3. Relative Accuracy

Relative accuracy quantitatively evaluates the absolute percentage error of a prediction at a time within the prediction horizon, t_H , if the algorithm has met the requirements of the previous metrics.

$$RA = 1 - Mean\left(\left\{\frac{|r_i - r_i^*|}{r_i^*} | t_H \leq t_i \leq t_{EoUP}\right\}\right) \tag{27}$$

6.4. Convergence

Convergence evaluates how fast the prediction performance (any accuracy based metric) improves towards the end life of the instance, if the algorithm has met the requirements of the previous metrics.

$$CG = \left(\frac{\frac{1}{2}\sum_{i=P}^{E_O UP} (t_{i+1}^2 - t_i^2) M_i}{\sum_{i=P}^{E_O UP} (t_{i+1}^2 - t_i) M_i} - t_p \right) \cdot \frac{1}{t_{EoUP} - t_p} \tag{28}$$

6.5. Performance Score

The final evaluation metric or performance score used in (Wang T. , 2013) will be used in this paper. The performance score is a weighted sum of the Rate of Acceptable Predictions, Relative Accuracy, and Convergence.

$$PH = Median(\{{}^kPH\}) \tag{29}$$

$$AP = Median(\{{}^kAR\}) \tag{30}$$

$$RA = Median(\{{}^kRA\}) \tag{31}$$

$$CG = Median(\{{}^kCG\}) \tag{32}$$

Prediction Horizon is the only metric with a unit of time while the others have a value between 0 and 1, where 1 implies perfect. Since PH will be used as a preliminary requirement for the performance of RA, a weighted sum of the other three will be used as the overall performance score.

$$score = w_1 \cdot AP + w_2 \cdot RA + w_3 \cdot CG \tag{33}$$

where $w_1 = 0.6, w_2 = .3, w_3 = .1$. (Wang T. , 2013)

7. DATA SET& EXPERIMENT

To compare the performance of the proposed enhancements to the baseline TBSP in (Wang T. , 2013), this paper will use the same data set and experiment as outlined in (Wang T. , 2013).

The Commercial Modular Aero-Propulsion System Simulation (C-MAPSS) is used in this paper. C-MAPSS is a tool for simulating a realistic large commercial turbofan engine which simulates an engine model of a 90,000 lb thrust class turbofan engine that was written using MATLAB and Simulink. (Saxena A. , Goebel, Simon, & Eklund, 2008) There are four data sets of the run-to-failure data acquired from the C-MAPSS simulation (Saxena & Goebel, 2008). However, only the fourth data set, FD004, was used in (Wang T. , 2013) and will be used in this paper.

The data set FD004, has 2 fault modes, 6 operating condition regimes, 249 training units, and 248 testing units. There are 25 fields in the data set: cycle number, 3 condition settings, and 21 sensor measurements. Though FD004 provides a training and testing set, (Wang T. , 2013) determined that the testing set contained instances with incomplete run-to-failure data and would not be suitable for the performance evaluation method described in Section 6. Therefore, in (Wang T. , 2013) and in this paper the 249 training units are partitioned in to a training set of 150 randomly selected units with the remaining 99 units being used for evaluation.

For the experiment, the regime identification, mean variance normalization, and regression modeling follow the same procedure described in (Wang T. , 2013). For the RUL estimation, the SLR, MED-DA, and DTW are used to determine the similarity between the test system and the degradation models. The RUL of the test system is calculated based on four different approaches: 1) minimum distance (point estimation), 2) model aggregation (point estimation), 3) KDE (probability interval), and 4) MLE - Maximum Likelihood Estimation (confidence interval). In Summary, there are 12 different RUL predictions being evaluating for this paper; each similarity measure will have 2 point estimation, KDE, and a MLE.

8. RESULTS

This paper compares the RUL prediction using the similarity measure MED-DA from (Wang T. , 2013) to the Pearson's linear correlation coefficient and Dynamic Time Warping measures based on the FD004 data of the C-MAPSS data set.

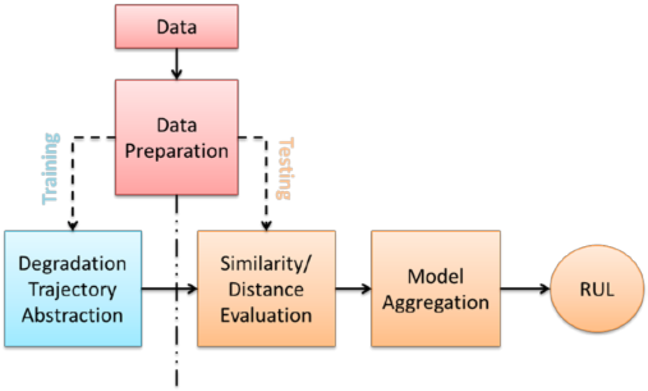
Data
Data Preparation
Training
Testing
Degradation Trajectory Abstraction
Similarity/ Distance Evaluation
Model Aggregation
RUL

Figure 8. TSBP high level process flow (Wang T. , 2013)

The results are quite different than the ones report in (Wang T. , 2013). However, in (Wang T. , 2013) a single trial of 150 randomly selected units were used for training and the remaining 99 were used for testing. In this paper we performed our analysis using 20 independent trials. Figure 9 shows a boxplot of the 20 trial scores as defined in Eq. (33) showing the median performance of the 20 trials with the 25th and 75th percentiles as the edges of the box. The whiskers of the box extend to the most extreme data points not considered outliers, and the outliers are plotted individually.

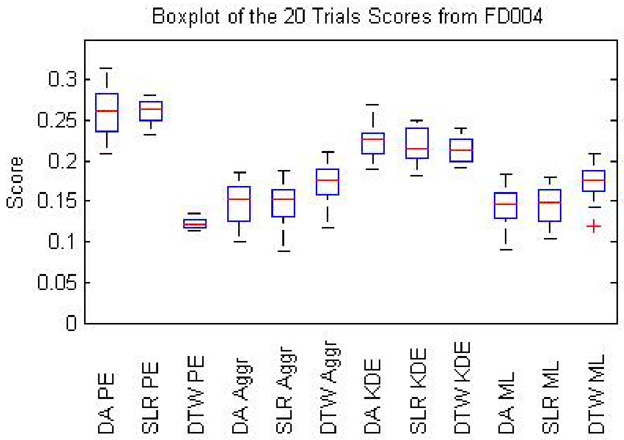

Figure 9. Boxplot of the 20 Trial Scores of 150 randomly selected units used for training and the remaining 99 used for testing.

The results in Figure 9 show that DA and SLR similarity measures are not significantly different for Point Estimation (PE), Aggregated (Aggr), Kernel Density Estimation (KDE), or Maximum Likelihood (ML) predictors. What is interesting is that the DTW measure performed worse than the DA and SLR measure using PE but outperformed them using a ML predictor. It is very difficult to form a conclusion based on the experiment performed by (Wang T. , 2013), because the results will be greatly dependent upon the randomly selected training and testing dataset. Hence, without knowledge of the specific randomly selected training model used for the results in (Wang T. , 2013), it is not feasible to perform analysis of all possible training model configuration to verify results.

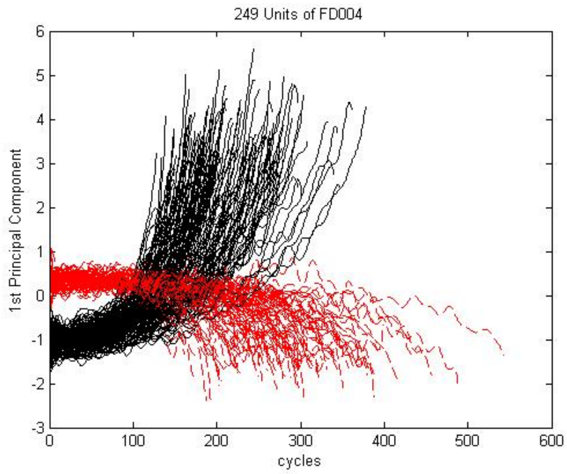

Figure 10. 249 Unit of the FD004 dataset. There were two fault modes identify in the dataset description file and can be clearly seen by the first principal component of the degradation model. Each line is 1st PC for the 249 units in the FD004 dataset.

9. BASELINE EXPERIMENT

Based on the above results we have decided to perform an additional experiment with the intent to baseline these measures and predictors for the FD004 dataset. In the baseline experiment we will make predictions for each of the 249 unit in the dataset. For each unit under test we will use the remaining 248 unit for training. This will allow the experiment to have maximum knowledge of the Fleet but without overlapping the degradation models and unit under test.

There are two fault modes identified in the FD004 dataset and can be seen in Figure 10. For simplicity, we will identify fault mode 1 as the red dashed lines and fault mode 2 as the solid black lines which have 101 and 148 degradation models, respectively. The only computational difference between the similarity evaluating of (Wang T. , 2013) and our baseline experiment is that we use only the degradation models for a given fault mode once it has been identified for the unit under test (UUT). The initial RUL predictions are based on all of the 248 degradation models, however after 30 or less cycles the UUT's fault mode is identified and the similarity comparisons is reduced to 101 or 148 degradation models. Of course at no time will the UUT degradation model be included in similarity computation of the training degradation models.

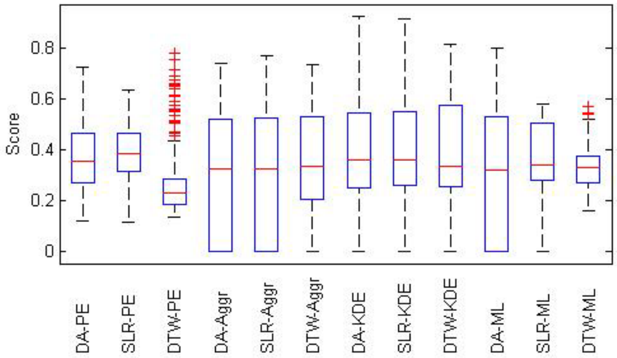

Figure 11. Boxplot of the baseline experiment.

Table 1. Median score performance of RUL similiarity-predictor combinations.

Pointe Estimate	DA-PE	SLR-PE	DTW-PE	Kernel Density Estimation	DA-KDE	SLR-KDE	DTW-KDE
20 Trials	0.2613	**0.2626**	0.1215	20 Trials	0.2267	0.214	0.2126
Baseline	0.2993	**0.3523**	0.1226	Baseline	0.2981	0.3035	0.2664

Model Aggregation	DA-Aggr	SLR-Aggr	DTW-Aggr	Maximum Likelihood	DA-ML	SLR-ML	DTW-ML
20 Trials	0.1528	0.1529	0.1769	20 Trials	0.1464	0.1492	0.1761
Baseline	0.2403	0.2327	0.2605	Baseline	0.2339	0.2808	0.2778

From Table 1, the SLR PE showed the best performance for both the 20 trial experiment adapted from (Wang T. , 2013) and our established baseline experiment. However, these scores are based on point estimation performance metrics (Saxena, Celaya, Saha, Saha, & Goebel, 2009) and do not take advantage of the probability or confidence intervals of the ML or KDE predictors.

10. CONCLUSION

This paper examined alternative approaches to measure similarity in a Trajectory Based Similarity Prediction framework. Additionally, we evaluated a similarity weighted Kernel Density Estimation RUL predictor and similarity weight maximum likelihood RUL predictor. The use of these weighted KDE and ML predictors allows the RUL prediction to be defined over a probability and confidence interval. The two experiments presented show that the point estimation predictor using the Simple Linear Regression measure performed the best for each experiment, but further research will be needed to examine the benefit of the KDE and ML predictors that are not fully evaluated by the performance metrics

Some sources of error and uncertainty for TBSP approach include multi-regime normalization and sensor aggregation through principal component analysis, see Section 2. The regime normalization assumes uniform system degradation within and across the operational regimes which may greatly impact the similarity-predictor performance.

Additionally, it is very difficult and impractical to make predictions of RUL for systems that have an unknown operational profile. It is anticipate that real world systems will have a known operational profile with a desired maintenance free period for a given system. Therefore, predictions and prediction accuracies should be based on a failure occurring within a maintenance free period and known operational profile for certain applications. We envision future research will be focused with these restrictions in mind.

NOMENCLATURE

t_i The time stamp of the i^{th} measurement cycle

z_i The sample of PC vector at the i^{th} measurement cycle

E The index of the End-of-Life measurement cycle for an instance.

P The index of the Start-of-Prediction cycle for an instance.

$EoUP$ The index of End-of-Useful-Prediction cycle for an instance.

r_I The estimated RUL at measurement cycle I

r_I^* The ground-truth RUL at measurement cycle I.

l A left super script applied to any of the above symbols, indicating the symbol corresponding to the l^{th} training instance or degradation model.

${}^{l}G$ The l^{th} degradation model extracted from the l^{th} training instance.

${}^{l}D^2$ Squared distance to the l^{th} degradation model trajectory.

${}^{l}S$ Similarity to the l^{th} degradation model trajectory.

$eSNR(\cdot)$ The Empirical Signal/Noise Ratio computed from 1-D time series data.

α Percentage of RUL prediction error bound, e.g. 0.2.

i_α The index of the first RUL prediction that satisfies the α-bound criteria.

REFERENCES

Botev, Z. I., Grotowski, J. F., & Kroese, D. P. (2010). Kernel density estimation via diffusion. *The Annals of Statistics*, 39(5), 2916-2957.

Dallachiesa, M., Nushi, B., Mirylenka, K., & Palpanas, T. (2012). *Uncertain Time-Series Similarity: Return to the Basics.* University of Trento.

Fonseca, J., Friswell, M. I., Mottershead, J. E., Lees, A. W., & Adhikari, S. (2005). Uncertainty Quantification using Maximum Likelihood: Experimental Validation. *46th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics & Materials Conference.* Austin.

Giusti, G. R. (2013). An Empirical Comparison of Dissimilarity Measures for Time Series Classification. *Brazilian Conference on Intelligent Systems (BRACIS).* Fortaleza.

Gouriveau, R., Ramasso, E., & Zerhouni, N. (2013). Strategies to face imbalanced and unlabelled data in PHM applications. *Chemical Engineering Transactions*(33), 115-120.

Guo, H., Gerokostopoulos, A., Liao, H., & Niu, P. (2013). Modeling and Analysis for Degradation with an Initiation Time. *Reliability and Maintainability Symposium.*

Lei, H., & Govindaraju, V. (2004). Matching and Retrieving Sequential Patterns Under Regression. *IEEE/WIC/ACM International Conference on Web Intelligence.*

Otey, M. E., & Parthasarathy, S. (2004). A Dissimilarity Measure for Comparing Subsets of Data: Application to Multivariate Time Series. *Department of Computer Science and Engineering, The Ohio State University.*

Salvador, S., & Chan, P. (2007). FastDTW: Toward Accurate Dynamic Time Warping in Linear Time and Space. *Intelligent Data Analysis.*

Sankararaman, S. (2014). Significance, interpretation, and quantification of uncertainty in prognostics and remaining useful life prediction. *Mechanical Systems and Signal Processing*, In press.

Sankararaman, S., & Goebel, K. (2013). Why is the Remaining Useful Life Prediction Uncertain? *Annual Conference of the Prognostics and Health Management Society 2013*.

Sankararaman, S., Daigle, M., & Goebel, K. (2014). Uncertainty Quantification in Remaining Useful Life Prediction Using First-Order Reliability Methods. *IEEE Transactions on Reliability*, 603-619.

Saxena, A., & Goebel, K. (2008). C-MAPSS Data Set. NASA Ames Prognostics Data Repository.

Saxena, A., & Goebel, K. (2008). *PHM08 Challenge Data Description*. Denver: 1st International Conference on Prognostics and Health Management.

Saxena, A., Celaya, J., Balaban, E., Goebel, K., Saha, B., Saha, S., et al. (2008). Metrics for Evaluating Performance of Prognostic Techniques. *Prognostics and Health Management (PHM).*

Saxena, A., Celaya, J., Saha, B., Saha, S., & Goebel, K. (2009). On applying the prognostic performance metrics. *Prognostics and Health Management Society*.

Saxena, A., Goebel, K., Simon, D., & Eklund, N. (2008). Damage Propagation Modeling for Aircraft Engine Run-to-Failure Simulation. *Ist International Conference on Prognostics and Health Management (PHM08).* Denver.

Wang, P., & Coit, D. W. (2007). Reliability Assessment Based on Degradation Modeling with Random or Uncertain Failure Threshold. *Reliability and Maintainability Symposium.* Orlando, FL.

Wang, T. (2013). *Trajectory Based prediction for Remaining Useful Life Estimation.* Cincinnati: University of Cincinnati.

Yu, P., Yong, X., Datong, L., & Xiyuan, P. (2012). Sensor Selection with Grey Correlation Analysis for Remaining Useful Life Evaluation. *PHM Society Conference.*