



TECHNICAL ARTICLE—PEER-REVIEWED

Reducing Dimensionality of Multi-regime Data for Failure Prognostics

Oguz Bektas · Amjad Alfudail · Jeffrey A. Jones

Submitted: 30 August 2017 / Published online: 23 October 2017
© The Author(s) 2017. This article is an open access publication

Abstract Over the last decade, the prognostics and health management literature has introduced many conceptual frameworks for remaining useful life predictions. However, estimating the future behavior of critical machinery systems is a challenging task due to the uncertainties and complexity involved in the multi-dimensional condition monitoring data. Even though many studies have reported promising methods in data processing and dimensionality reduction, the prognostics applications require integration of these methods with remaining useful life estimations. This paper describes a multiple linear regression process that reduces the number of data regimes under consideration by obtaining a set of principal degradation variables. The process also extracts health indicators and useful features. Finally, a state-space model based on frequency-domain data is used to estimate remaining useful life. The presented approach is assessed with a case study on turbofan engine degradation simulation dataset, and the prediction performance is validated by error-based prognostic metrics.

Keywords Failure prognostics · Multi-dimensional data · Dimensionality reduction · Remaining useful life estimation

O. Bektas (✉) · J. A. Jones
Warwick Manufacturing Group, University of Warwick,
Coventry CV4 7AL, UK
e-mail: o.bektas@warwick.ac.uk

J. A. Jones
e-mail: j.a.jones@warwick.ac.uk

A. Alfudail
Mechanical Engineering Department, Umm Al-Qura University,
Al Taif Road, Makkah 24382, Saudi Arabia
e-mail: asfudail@uqu.edu.sa

Introduction

Maintenance strategies have witnessed substantial changes over the years. The paradigm has shifted from classic breakdown repairs to more complex and sophisticated condition monitoring strategies, which avoid unnecessary tasks by taking actions only when there is evidence of abnormal system behaviors [1]. Due to the increase in the variety and number of assets with more complex designs, the maintenance strategies must respond changing expectations and increasing awareness in an attempt to achieve high plant availability and reliability in operations.

Prognostics can make contributions into these changing expectations by providing dynamic maintenance planning strategies for critical engineering systems. They can provide improved reliability and reduced costs for operation and maintenance of complex systems. As a steadily growing subject, prognostics have advanced expertise in various disciplines [2]. Many breakthroughs in remaining useful life estimation can be found in complex engineering systems such as electronics [3, 4], batteries [5, 6], actuators [7], turbofan engines [8, 9] and NASA's launch vehicles and spacecraft systems [10].

In general, a typical prognostic method modeled for the complex systems depends on measured condition monitoring data and provides simplified representations of complex datasets. Considering that the operations are generally performed in multiple regimes, data processing becomes a major issue confronting the prognostic users. Since it is very unlikely to evaluate the operational and environmental conditions, a systematic framework for data processing is required to account for the uncertainties in prognostics [11]. Such a data processing can analyze the uncertainties in condition monitoring data for a better

understanding of the system’s damage propagation in upcoming operational and environmental conditions.

The main objective of this paper is to develop a conceptual prognostic framework to overcome the issues presented by noisy and multi-dimensional data. Multiple linear regression is used to model the relationship between different explanatory regime variables and a monotonic response variable by fitting an equation to monitoring data. This process returns the coefficient estimates for a multiple linear regression of the responses which can be further used to calculate the response variables of all different operational trajectories. A state-space model is then proposed to use these response variables for the multi-step ahead remaining useful life (RUL) predictions.

Motivation and Problem Statement

For a degradation process that is predicated on system aging and monotonic damage accumulation and manifests itself in the physical composition of the system, it should be possible to correlate sensor behavior with signs of aging to estimate the remaining useful life of systems [12]. However, the multi-dimensional data caused by multiple operational regimes could not provide useful information to measure the monotonic damage accumulation. Further applications are needed to provide useful information for remaining useful life predictions.

To identify multi-dimensional data in a degradation process, let

$$\mathbf{X}_i = [x_{i1}, x_{i2}, \dots, x_{it}, \dots, x_{iT_i}] \tag{Eq 1}$$

be the set of features for the *i*th unit in a dataset and

$$x_t = [x_{t,1}, x_{t,2}, \dots, x_{t,q}] \tag{Eq 2}$$

be the *q*-dimensional feature vector extracted from the raw data collected from a system [9].

The preprocessing of such raw data is an essential step to any study relying on any type of data-driven techniques. To obtain a meaningful wear level index for prognosis, a data processing approach is applied for feature extraction, data cleaning and feature selection. The characteristics of raw data and system conditions are first extracted. Then, any useless and misleading outliers caused by noise during operations are removed. This practice first deals with the issues relating to organizing the multi-dimensional data to reduce data redundancy and improve regime integrity. The noisy sensors operating under different regimes are standardized to each other, and so the common behavior of sensors can be observed and investigated. Next, a wear level index can provide comparable and actionable information about the common population health, as well as track degradation progress and performance over time.

Signal Processing and Dimensionality Reduction

The “turbofan engine degradation simulation dataset” used in this paper was provided by the Prognostics CoE at NASA Ames and made publicly available [13]. Engine degradation simulation was carried out using C-MAPSS software, and four different scenarios were simulated under different combinations of operational conditions, regimes and fault modes (see Table 1). Several sensor channels in the datasets characterize the fault evolution. It is expected from users to develop their algorithms using training sets and make the remaining useful life estimations by using test sets provided in the package.

All four datasets are formed of multi-various time series, which are assembled into training and test subsets. The start of each variable is set in normal operational conditions with an unknown case-specific initial wear level which is considered normal [13].

Training time series operates in full operational periods which terminates at a failure point due to the wear. On the other hand, the test subsets are ended at a certain point before the engine reaches the system failure. The challenge is to predict the remaining useful life between the end of each test set and to validate the results with the actual failure point which was given separately by a vector corresponding to true RUL values of the test data [14].

Each measurement in both data subsets is a snapshot data which are taken during a single operational cycle. Although the measurements are not named, it is known to users that they correspond to different variables [13].

Datasets with single and multiple operational regimes are used in this paper. It is observed that some sensors behave differently in different datasets. The raw measurements are highly noisy and scattered values with different value ranges in each single series.

Table 1 Turbofan engine degradation simulation dataset characteristics

Dataset: FD001	Dataset: FD003
Train trajectories: 100	Train trajectories: 100
Test trajectories: 100	Test trajectories: 100
Conditions: ONE (sea level)	Conditions: ONE (sea level)
Fault modes: ONE (HPC degradation)	Fault modes: TWO (HPC and fan degradation)
Dataset: FD002	Dataset: FD004
Train trajectories: 260	Train trajectories: 248
Test trajectories: 259	Test trajectories: 249
Conditions: SIX	Conditions: SIX
Fault modes: ONE (HPC degradation)	Fault modes: TWO (HPC and fan degradation)

Multiple Linear Regression

The raw values of selected time series, which are inconsistent with each other, need a feature extraction transformation of the multi-regime data in the high-dimensional space to a space of a single wear level dimension. This transformation can reduce the dimensionality of the time series from their original scales to a notionally common scale that will include meaningful information for prognosis.

Feature extraction and dimension reduction can be combined in one step by using “multiple linear regression” model which performs a mapping of the multi-regime data to a lower-dimensional space in such a way that the variance of the measurements in the low-dimensional finding is maximized. Multiple linear regression calculates the relationship between different explanatory variables and a target variable by fitting a linear equation to observed data [15, 16]. This model is based on:

$$Y = X\beta + \epsilon \quad (\text{Eq 3})$$

where Y is a $n \times 1$ vector of values of the target variable, X is an $n \times p$ matrix of observed responses and β is a $n \times 1$ vector of coefficient estimates for a multiple linear regression of the responses.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_p x_p \quad (\text{Eq 4})$$

More complex models may include multiple observations.

$$y_i = \beta_0 + \beta_{1x_{i1}} + \beta_{2x_{i2}} + \dots + \beta_{px_{ip}} \text{ for } i = 1, 2, \dots, k \quad (\text{Eq 5})$$

With this equation, the multiple linear regression model can be formulated in the following form.

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1n} \\ 1 & x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nn} \end{bmatrix} \quad (\text{Eq 6})$$

In fitting the multiple linear regression model, the coefficients are calculated by the methods of least squares.

$$\hat{\beta} = (X'X)^{-1} X'y \quad (\text{Eq 7})$$

Synthetic Wear Level Index Estimation

To assign the target variable, a mathematical model for the synthetic data has been established. This makes it possible to model a useful prognostic output for raw measurement data. Since the exact behavior of degradation change is known, the coefficient variables can be calculated with

regard to the operational setting, and the differences caused by the noise.

The generalized time-varying health index equation can be used as a synthetic wear level index to yield supervised classifications for C-MAPSS datasets [14].

$$h(t) = 1 - d - \exp(at^b) \quad (\text{Eq 8})$$

where d is an arbitrary point in the wear space, a and b are model parameters and t is time. This health index can be used for various phenomena within a system.

With reference to this function, the following equation for synthetic wear level index is formulated [9].

$$sWI_t = 1 - \exp\left(\frac{\log(0.05) \times (1-t)}{0.95 \times t}\right) \quad (\text{Eq 9})$$

where t is the time unit and l is the length of time series representing the full sets of operations. This function forces exponentially to increase wear levels. In Fig. 1, the wear levels with different operational length measures are shown.

The datasets FD002 and FD004 consist of a set of operational regimes, but the degradation trends can be clearly seen after the readings at each regime are selected and monitored separately. In order to increase the performance of the multiple linear regression model, the readings at each regime order can be clustered and the dimensionality reduction is applied into these clustered readings [9].

In order to standardize the entire dataset into a common scale, only the calculated coefficients from a reference trajectory are applied in Eq. 5. Considering the initial wear levels and the failure threshold points, the dimensionality of different raw sensor measurements is organized and all trajectories in the same dataset are normalized (see Fig. 2)

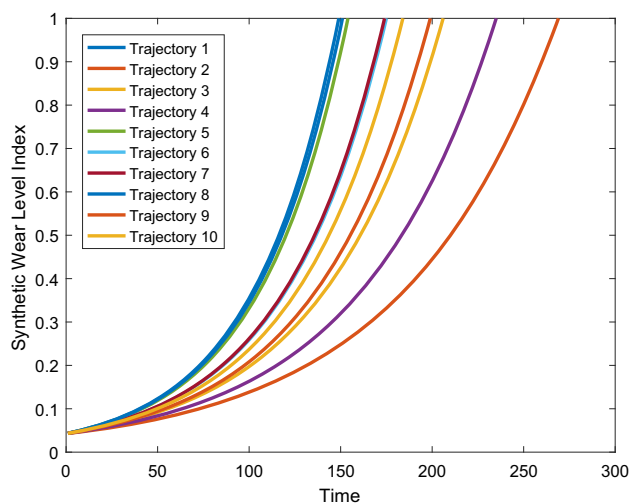


Fig. 1 Synthetic wear level measurements (based on the first ten trajectories from FD002)

Failure Prognosis

Training trajectories demonstrate full operational life time of engines, and failure occurs at a certain point which is accepted as the threshold level for wear growth. Test subsets, on the other hand, end some time prior to failure occurrence. This means that there is an unknown time to failure and also that there are no real data to train remaining step.

In the lack of future data steps, the state-space modeling predicts the future behavior of test subsets with a direct connection from reference training trajectories. It is necessary to train the model with a training subset and then convert to the estimation mode to make multi-step ahead remaining useful life estimations by including only the external test trajectories.

State-Space Estimation Model

The proposed multi-step ahead prediction algorithm estimates a continuous-time state-space model of order nx using the frequency-domain data, the recurrence relation of wear level index. The function generates a state-space model object with identifiable parameters [17].

A state-space model with input u , output y and error term (disturbance) e is represented by the following equation in continuous time.

$$\frac{dx_t}{dt} = Ax_t + Bu_t + Ke_t \tag{Eq 10}$$

$$y_t = Cx_t + Du_t + e_t \tag{Eq 11}$$

where A, B, C, D and K are state-space matrices, and x_t is the vector of nx states.

Considering the discrete time, the state-space estimation model takes the following form.

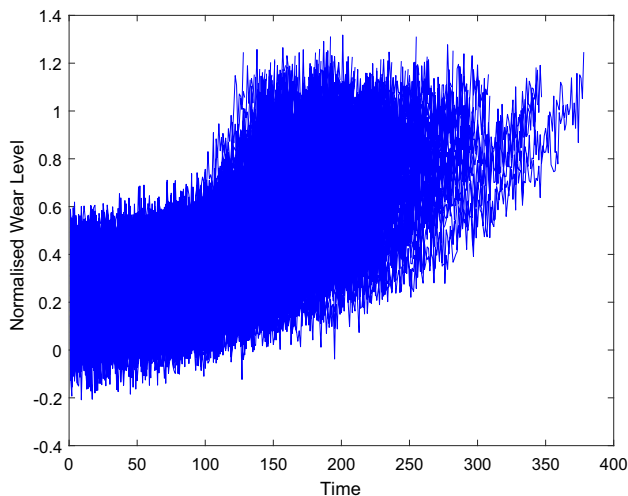


Fig. 2 Wear level index of trajectories

$$x_{(k+1)} = Ax_k + Bu_k + Ke_k \tag{Eq 12}$$

$$y_k = Ck_t + Du_k + e_k \tag{Eq 13}$$

This model matches the measured wear level index. When the future behavior of the wear level as a state in the model is concerned, an arbitrary state of the identified model can be transformed so that the state can make multi-step ahead predictions.

After the dimensionality of data is reduced to a single wear level index for each trajectory, it is expected that the wear growth model can be applied to learn the pattern from historical data and to estimate remaining useful life time until the pattern exceeds threshold point. Although the state-space model can accomplish the training for the cleaned vectors, it cannot produce predictions of multi-step long-term time series when exponential growth is present as in the case in Fig. 2. The exponential series is transformed so that the model can perform well. Then, each further series is defined as a function of the preceding values [18].

The corresponding formula for the recurrence relation of the exponential growth is

$$x_{r(i)} = \begin{cases} 1 & \text{for } i = 1 \\ x_{f(i)}/x_{f(i-1)} & \text{for } i \geq 2 \end{cases} \tag{Eq 14}$$

where x_r corresponds to the recurrence relation of the wear index which will be used for state-space modeling and forecasting.

Since the wear level index is noisy, it is required to be simplified into a form that is suitable for the recurrence relation. It is observed that when the wear level is fitted and recursively defined, the algorithm can perform effectively in terms of prediction performance.

The proposed model matches the wear level index between the test trajectories and the corresponding part of the training trajectories, but the model is interested in the recurrence relation which is a state in the model. After the data are recalculated as shown in Fig. 3, the wear index patterns take a stationary form rather than being non-stationary. The proposed model has an arbitrary state that can be transformed so that the stationary state has meaning, in this case the recurrence relation of the exponential wear index.

The model then transforms the state coordinates in order to generate a multi-step ahead predictor expressed in the same state coordinates as the original training model so that the model's state corresponds to the time dependent trajectory cycle size. The key point is to rely on actual, direct measurements of the recurrence relation of the matching training trajectory. In practice, the predictor state of the matching trajectory x_n is transformed into the multi-step ahead prediction state z_n . After the multi-step ahead

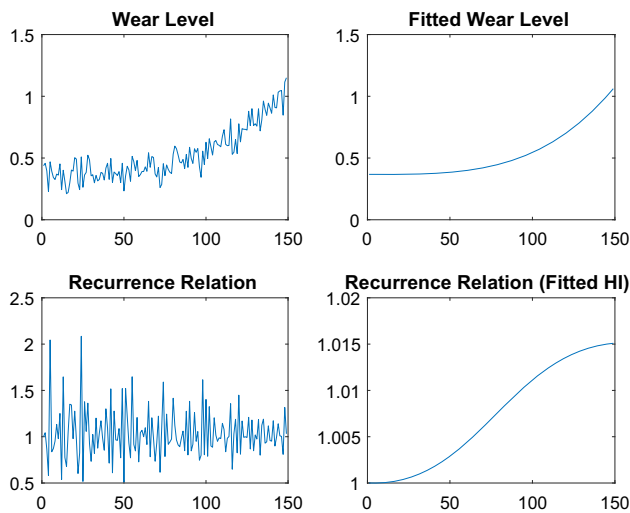


Fig. 3 Wear level and recurrence relation

predictor is expressed in the desired state coordinates, it has a single input, the measured system output, and a single output, the predicted system output. This predictor function is simulated to estimate the system output and system state of the matching training trajectory. When the predictor state of the test trajectory is applied into that function, the estimated output of the test data with measured and known values can be achieved. In Fig. 4, the blue curve shows the recurrence relation of the original recurrence relation. The red curve is also the recurrence relation, but it is derived from a fitted wear level index and it is used in the model to increase the performance of the model. The yellow curve shows the forecasted response for 200 hours beyond the measured test data’s time range.

Subsequently, the final data vectors should be reinstated to the initial exponential form, after the forecasted values are received and the predictor function makes multi-step predictions with these series.

$$x_{p(i)} = \begin{cases} x_f(end) & \text{for } i = 1 \\ x_{p(i)} \times x_{p(i-1)} & \text{for } i \geq 0 \end{cases} \quad (\text{Eq 15})$$

RUL estimation of the model corresponds to a unique number of cycles in each instance. However, all the calculations from multiple training inputs are required to describe the relative likelihood of the remaining useful life variable to take on a given value. Therefore, the final estimation for each instance varies to the other calculations. Figure 5 illustrates multiple predictions for a single test trajectory. Each prediction here is estimated from a different predictor function trained with a different matching training trajectory. In order to minimize the prediction risk, the final remaining useful is accepted as the mean of the top matching predictor calculations.

Remaining useful life is calculated as the time interval between the end of the test subset and the point where the

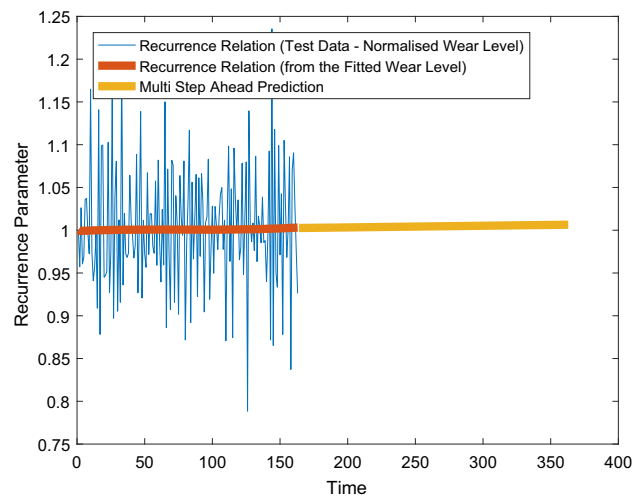


Fig. 4 Multi-step ahead prediction of recurrence relation

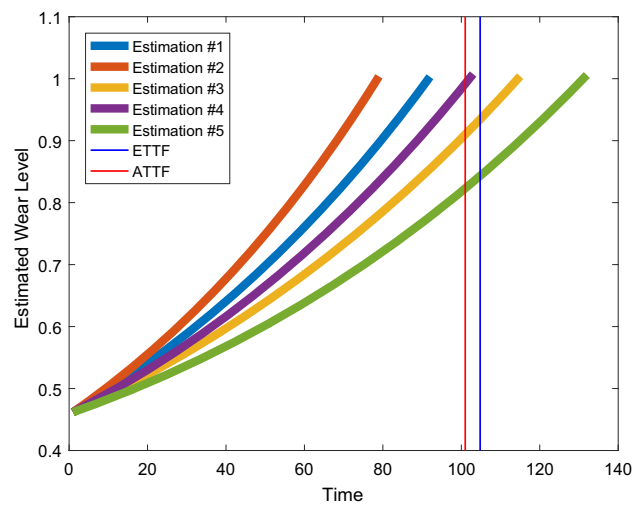


Fig. 5 Multi-step ahead prediction (after reinstating to exponential form)

prediction value exceeds the value of training subset target vector. In Fig. 6, the future wear growth at gas turbine performance is shown. This calculation can be made as much as the number of compatible training subsets. In other words, the model can give more detailed and more accurate results as the amount of operational data increases.

Results and Discussion

The C-MAPSS turbofan dataset provides a separate vector of true remaining useful life values for the test data series. According to their true RUL values, the performance evaluation metrics based on the estimation performance can be applied. The measurements have signified their

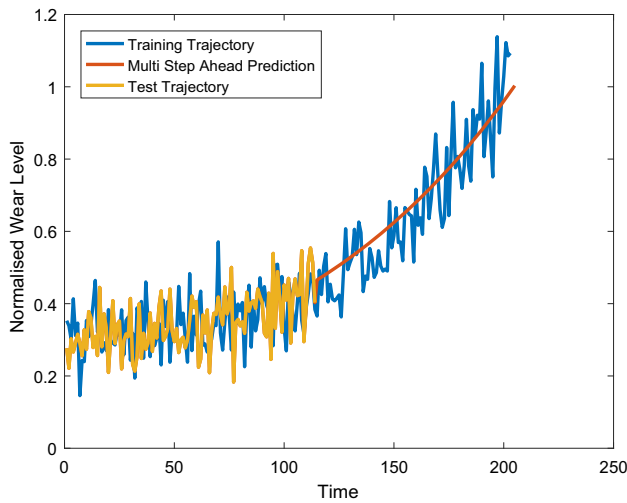


Fig. 6 Remaining useful life

practical relevance in prognostic designs and have found their way into multi-step predictions. The metrics used in this research are based on the works of [14, 19].

Mean Absolute Error

MAE calculates an average of the absolute error terms.

$$MAE = \frac{1}{n} \sum_{i=1}^n |e_i| = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \tag{Eq 16}$$

Mean Absolute Percentage Error

MAPE averages the absolute percentage errors in the predictions of multiple RUL calculations at the same prediction horizon.

$$MAPE = \frac{100}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \tag{Eq 17}$$

Mean Square Error

MSE is a risk function that calculates the average of the squares of the errors.

$$MSE = \frac{1}{n} \sum_{i=1}^n e_i^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \tag{Eq 18}$$

False Positive Rate (FP) and False Negative Rate (FN)

FP is the ratio where a fault is predicted in spite of the asset performing within desired conditions. Conversely, a negative rate is the ratio of unanticipated predictions when the system would fail.

$$FP(i) = \begin{cases} 1 & \text{error} > t_{FP} \\ 0 & \text{otherwise} \end{cases} \tag{Eq 19}$$

$$FN(i) = \begin{cases} 1 & 1 - \text{error} > t_{FN} \\ 0 & \text{otherwise} \end{cases}$$

Table 2 Prognostic performance metrics

Dataset	MAE	MAPE	MSE	FP (%)	FN (%)
FD001	13.6	18.2	332.6	60	40
FD002	16.2	26.1	555.5	52	48
FD003	15.7	20.9	498.5	63	37
FD004	18.3	24.3	630.3	56	44

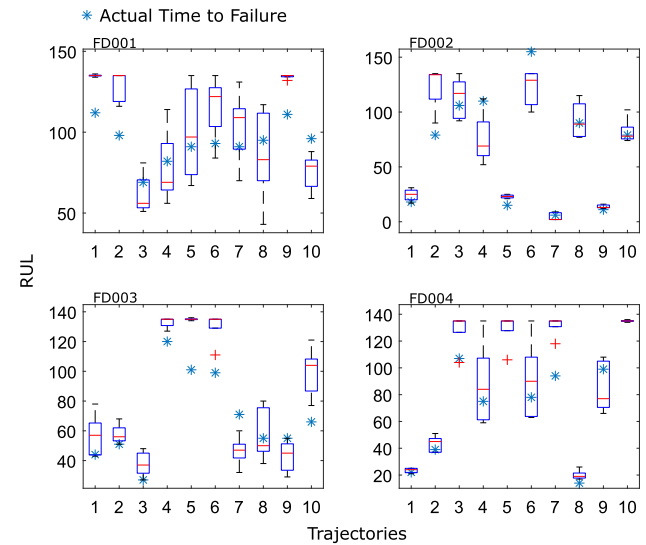


Fig. 7 Comparison of true and predicted RUL

where t_{FP} , and t_{FN} are the user-defined acceptable early or late prediction limits.

In Table 2, the prognostic metric results are shown. The multi-step forecast performance over the long-term cycles is calculated in a close-range to true remaining useful life. The performance evaluation prognostic metrics have been prepared to determine whether or not the designed algorithm or multi-step prediction results can show their practical results. The developed model seems to exhibit promising results at multi-step long-term time series predictions for exponential wear growths. The training of network could accomplish learning as desired, while training performance is substantially increased by multiple predictor function use and the recurrence relation calculation.

Figure 7 provides a comparison of true and predicted RULs for the first ten trajectories of all datasets. The box plots demonstrate that most of the true RULs are within the range of upper and lower whiskers, whereas a considerable number are actually between the upper and lower quartiles. However, some cases are particularly dangerous for the performance evaluation made by prognostic metrics because their high error rates are detrimental to the performance level of the entire dataset.

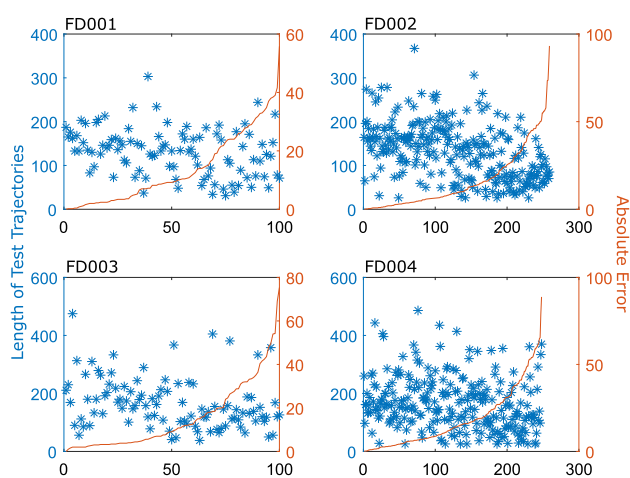


Fig. 8 Comparison of absolute error and unit test length

A comparison of absolute error rates and test trajectory unit lengths is provided in Fig. 8. In each dataset case, the error rates show a clear increase as the unit lengths decrease. Considering that the longer test trajectories are mature enough to adequately represent system behavior, the RUL predictions in these samples are generally consistent and do not result in high error rates. It is observed that the consistency in the mature trajectories is a result of the grown patterns, which are not affected by undesired fluctuations in the data. On the other hand, in the case of short trajectories, the variance in data fluctuations is a major concern as they might result in undesired failures in the overall accuracy.

Conclusion

In this paper, a multiple linear regression-based dimensionality reduction model is proposed for multi-step ahead remaining useful life estimation. The prediction method builds on a state-space model using frequency-domain data.

The performance of the proposed prognostic method is evaluated by four different subsets of turbofan engine degradation simulation dataset which were simulated under different combinations of operational conditions and fault modes. The results have shown that the combination and filtering of models can yield a low error rate in the remaining useful life prediction.

Analysis of the multi-step ahead estimation suggests that the model can determine the remaining useful life of an average operating system, and can adjust the estimation over time-based usage data. It is also observed that the dimensionality reduction model can detect the initial wear levels of different trajectories.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

1. A.K.S. Jardine, D. Lin, D. Banjevic, A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mech. Syst. Signal Process.* **20**, 1483–1510 (2006)
2. J. Lee, F. Wu, W. Zhao, M. Ghaffari, L. Liao, D. Siegel, Prognostics and health management design for rotary machinery systems—reviews, methodology and applications. *Mech. Syst. Signal Process.* **42**, 314–334 (2014)
3. M. Pecht, *Prognostics and Health Management of Electronics* (Wiley, Hoboken, 2008)
4. N.M. Vichare, M.G. Pecht, Prognostics and health management of electronics. *IEEE Trans. Compon. Packag. Technol.* **29**, 222–229 (2006)
5. B. Saha, K. Goebel, S. Poll, J. Christophersen, Prognostics methods for battery health monitoring using a Bayesian framework. *IEEE Trans. Instrum. Meas.* **58**, 291–296 (2009)
6. K. Goebel, B. Saha, A. Saxena, J.R. Celaya, J.P. Christophersen, Prognostics in battery health management. *IEEE Instrum. Meas. Mag.* **11**(4), 33–40 (2008)
7. C.S. Byington, M. Watson, D. Edwards, P. Stoelting, A model-based approach to prognostics and health management for flight control actuators, in *Proceedings of Aerospace Conference*, vol. 6, pp. 3551–3562 (2004)
8. T. Wang, J. Yu, D. Siegel, J. Lee, A similarity-based prognostics approach for remaining useful life estimation of engineered systems, in *Proceedings of International Conference on Prognostics and Health Management, PHM*, vol. 2008, pp. 1–6 (2008)
9. E. Ramasso, Investigating computational geometry for failure prognostics. *Int. J. Progn. Health Manag.* **5**, 005 (2014)
10. V.V. Osipov, D.G. Luchinsky, V.N. Smelyanskiy, C. Kiris, D.A. Timucin, S.H. Lee, In-flight failure decision and prognostics for the solid rocket booster, in *Proceedings of AIAA 43rd AIAA/ASME/SAE/ASEE Joint Propulsion Conference and Exhibit*, Cincinnati, OH, (2007)
11. S. Sankararaman, K. Goebel, Uncertainty in prognostics and systems health management. *Int. J. Progn. Health Manag.* **6**, 010 (2015)
12. S. Uckun, K. Goebel, P.J.F. Lucas, Standardizing research methods for prognostics, in *Proceedings of International Conference on Prognostics and Health Management, PHM 2008*, (2008)
13. A. Saxena, K. Goebel, “Turbofan engine degradation simulation data set”, NASA Ames prognostics data repository (NASA Ames Research Center, Moffett Field, CA, 2008), <https://ti.arc.nasa.gov/tech/dash/pcoe/prognostic-datarepository/>
14. A. Saxena, K. Goebel, D. Simon, N. Eklund, Damage propagation modeling for aircraft engine run-to-failure simulation, in *Proceedings of International Conference on Prognostics and Health Management, PHM 2008*, (2008)
15. S. Chatterjee, S.H. Ali, Influential observations, high leverage points, and outliers in linear regression. *Stat. Sci.* **1**, 1–6 (2008)
16. D.A. Freedman, *Statistical Models: Theory and Practice* (Cambridge University Press, New York, 2009), pp. 41–60
17. L. Ljung, *System Identification: Theory for the User*, PTR Prentice Hall Information and System Sciences Series (Prentice Hall, New Jersey, 1999), pp. 81–90

18. O. Bektas, J.A. Jones, NARX time series model for remaining useful life estimation of gas turbine engines, in Proceedings of Third European Conference of the Prognostics and Health Management Society, (2016)
19. K. Goebel, A. Saxena, S. Saha, B. Saha, J. Celaya, *Machine Learning and Knowledge Discovery for Engineering Systems Health Management—Prognostic Performance Metrics* (CRC Press, New York, 2011), p. 147