Component based data-driven prognostics for complex systems: Methodology and applications

A. Mosallam, K. Medjaher, N. Zerhouni
FEMTO-ST Institute, AS2M department
Univ. Bourgogne Franche-Comté, Univ. de Franche-Comté/CNRS/ENSMM
24 rue Alain Savary, F-25000 Besançon, France
Email: kamal.medjaher@ens2m.fr

***Abstract*—In recent years, considerable research efforts have been applied in the field of fault prognostics. However, to the authors knowledge, there are few published works that address complete and systematic methods describing the steps required to develop data-driven prognostics approaches for complex systems. This paper presents a generic component-based prognostics methodology that can be customized for different applications and which can be useful for new researchers and engineers. The paper is divided into two parts. The first part provides a description of the procedures required before constructing data-driven prognostics, such as identifying critical components, selecting physical parameters to monitor, choosing monitoring sensors and defining the data acquisition system. The second part presents a novel data-driven prognostic method for direct remaining useful life (RUL) prediction. This method relies on two phases: offline and online. In the offline phase, a method for constructing health indicators (HI) from sensor data is presented. Such HIs can be used as offline models to display the deterioration evolution of components over time. In the online phase, similar HIs are constructed from the sensor data for a new component. Then, a discrete Bayesian filter is applied to estimate the current health status. Finally, the offline database is searched to find the closest group to the online HIs. The selected offline HIs can be used for estimating the RUL of the new component under operation. The performance of the method is demonstrated using two real data sets taken from the NASA Ames prognostics data repository.**

***Index Terms*—Data-driven prognostics, remaining useful life, health indicators construction, discrete Bayes filter, Gaussian process regression.**

I. Introduction

Achieving high reliability and availability of complex systems is a crucial task. This can be done by adopting efficient maintenance activities to detect and correct problems before they become severe and shut down the system. Effective maintenance was shown to increase the reliability and availability by offering greater utilization of any facility of complex systems and reducing costs through managing work and downtime. Many types of maintenance strategies have been developed over last decades. Due to recent development of sensor and monitoring technology, Condition-Based Maintenance (CBM) has emerged as a promising strategy. It uses visual inspection and sensor data to assess the machinery condition. CBM replaces predefined maintenance tasks with only the necessary ones, based on the equipment current condition. In this way, CBM reduces maintenance costs while increasing efficiency by performing maintenance actions only when there is evidence of abnormal behavior. Recently, CBM+ strategy is proposed to deal with new requirements in the maintenance domain. Such requirements necessitate predicting the system health condition in the future and take decisions accordingly. CBM+ can be defined as an updated maintenance concept that emphasizes on prognostics or predictive capabilities, assessment of the material condition and estimation of the remaining useful life at any time during a system's life. Moreover, Prognostics and Health Management (PHM) is a set of advanced diagnostic, prognostic, and health management research activities that enables and supports CBM+. PHM activity attracts significant research interest due to the need for prediction and decision models, which are important concepts for performing efficient CBM+ strategy.

Performing PHM for a whole complex system, however, is challenging in practice. Instead, component-based PHM approaches are more feasible. Such approaches are based on two main parts. In the first part, system experts 1) identify the critical components, 2) select the physical parameters to monitor, 3) select the monitoring sensors, 4) perform the data acquisition and 5) pre-process the sensor signals. In the second part, the generated sensor data are used by PHM researchers and data analytics engineers to perform one or more component-based PHM tasks, such as: 1) data analysis, 2) fault detection, 3) diagnostic, 4) prognostics, 5) decision making and 6) human machine interface. Prognostics, in particular, has recently attracted a lot of research interest due to the need of predictive models. In [1], prognostics is defined as the estimation of the remaining useful life (RUL) of a component (or a system) based on its current health state and knowing its future operating conditions (Figure 1).

Generally, prognostics can be realized using three main approaches: 1) model-based (physics of failure) approach, 2) data-driven approach and 3) hybrid approach. Data-driven prognostic approaches are becoming popular due to their intuitive nature, fast developmental cycle and the advances of modern sensor systems as well as data storage and processing technologies. These approaches can be used when the first principles of the system operation are complex such that developing of accurate physics of failure model is not feasible

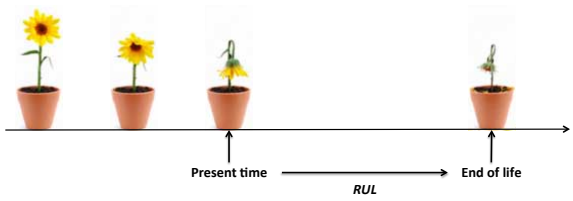

Fig. 1. Illustration of prognostics.

[2].

Data-driven approaches use empirical models to learn the degradation mechanisms from monitoring data. Empirical models map the relation between the system state variables, namely input, internal and output variables without explicit knowledge of the physical behavior of the monitored component. Such models can be divided into two main overlapping groups, namely Computational Intelligence (CI) and Machine Learning (ML) models. CI group includes bio-inspired models, such as neural networks and fuzzy systems. ML based approaches learn from experience and can enhance its performance over time, such as similarity based approaches and Bayesian based approaches. Bayesian approaches have a natural way of representing the uncertainty in a probabilistic form. This property is paramount for performing data-driven prognostics. The RUL appears to be a random variable and can be modeled as a stochastic process [3]. Therefore, uncertainty bounds or confidence intervals should be applied and accompany RUL estimation [4]. In addition, building Bayesian models does not require understanding the system behavior and it can be used to model multidimensional dynamic systems.

There are two main approaches to build data driven models, namely cumulative degradation and direct RUL mapping prognostics approaches [5]. In cumulative degradation prognostics approach, empirical models are used to map the degradation evolution of the desired system. These models are later used to estimate the new system health status. After knowing the new system's current health status, the RUL can be predicted based on the expected future behavior (Figure 2).

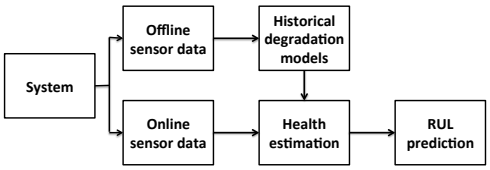

Fig. 2. Cumulative degradation based prognostics.

In direct RUL mapping prognostics approach, empirical models are also employed to build RUL models. However, these approaches directly map the relation between sensor data and the corresponding EOL value without the need to estimate the health status and from that estimate the RUL of the monitored component (Figure 3). To do this, health indicators are extracted from the raw monitoring signals, which may have originated from single sensor or from a number of sensors aggregated to represent the degradation evolution over time. This approach is relatively easy to implement and there are few published examples in the literature [2].

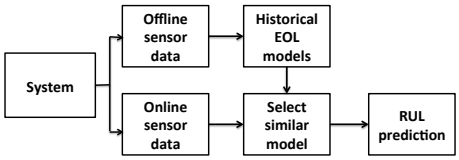

Fig. 3. Direct RUL mapping approach.

The main contribution of this paper is to present a complete method for developing a data-driven prognostics models for complex systems. This paper is structured as follows. Section II presents the proposed method. The applications and results are depicted in Section III. Section IV concludes the paper.

II. Component Based Data-Driven Prognostics for Complex System

The presented method is divided into two main parts, namely, towards data-driven prognostics and remaining useful life estimation of critical components based on Bayesian approaches. The first part provides a description of the procedures required before constructing data-driven prognostics, such as identifying of critical components, physical parameters to monitor, the monitoring sensors, data acquisition and signal pre-processing. The second part presents a novel method for direct RUL mapping prognostics based on Bayesian approaches (Figure 4).

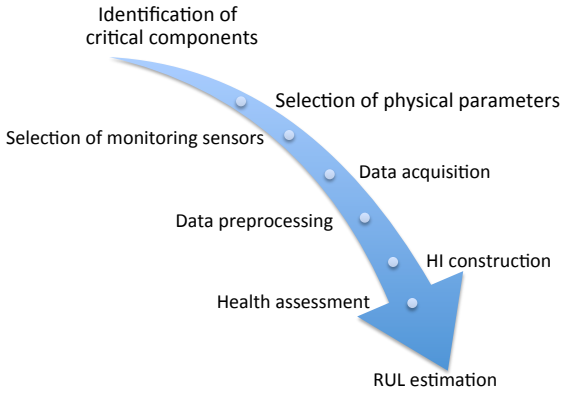

Fig. 4. Steps of the proposed method.

A. Towards data-driven prognostics

The efficiency of data-driven prognostics models depends on the quality of the available historical data. Extracting such data from the industrial system is a challenging step due to the increased complexity of modern industrial systems and due to the noise that might affect the acquired signals. Therefore, system experts have to study the system to decide the monitoring level. Monitoring complex systems can be done

on two different levels. 1) System level: it is used with large-scale systems consisting of multiple components or/and subsystems and the fault propagates through such components and 2) component level: components that show high failure rate are considered critical and should be monitored. For example, building prognostics models for a whole airplane can be challenging and still quite difficult in practice. Instead, component-oriented prognostics approaches build on identifying critical subsystems or components in the systems to be monitored and maintained individually.

1) Identification of critical components: one way to identify critical components in a complex system is by using hazard analysis [6]. Hazard analysis is a methodology to estimate the likelihood that a condition or event might happen, which could lead to an undesirable circumstance [7]. A successful hazard analysis requires sufficient technical knowledge about the desired system and appropriate hazard analysis methodology. There are many hazard evaluation techniques which complement rather than supplant the others. Generally, hazard evaluation can be divided into two main techniques, namely qualitative and quantitative (Figure 5). Each technique

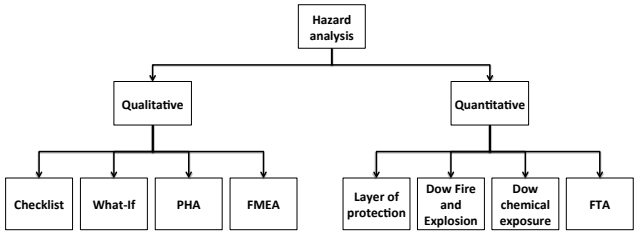

Fig. 5. Summary of hazard analysis discussed in this section.

approaches the system in a different way. Therefore, there is no one technique that is suitable for all situations. However, Failure Mode and Effects Analysis (FMEA) and Fault tree analysis (FTA) have been proposed in the PHM context [8]. FMEA is a systematic technique for analyzing component failure and documenting the resulting effect on system performance whereas FTA represents the factors and events using standard logic symbols. The result of a hazard analysis for a desired system is a list of all possible hazards that could result from a failed component or subsystem and their likelihood. Components with high failure rate are considered critical and should be monitored (Figure 6).

2) Selection of physical parameters: after locating the critical components, system expert chooses the appropriate physical parameters to monitor. These parameters are chosen on the basis of experience gathered from dealing with such systems. Quantities such as position, speed, acceleration, torque, vibration, temperature and strain are studied for long time and chosen to monitor mechanical systems. For example, the cause vibration in different machines can be linked to fault progression. Accurate monitoring of the vibration using appropriate sensors is therefore required to monitor health status of such machines. Table I depicts an example of

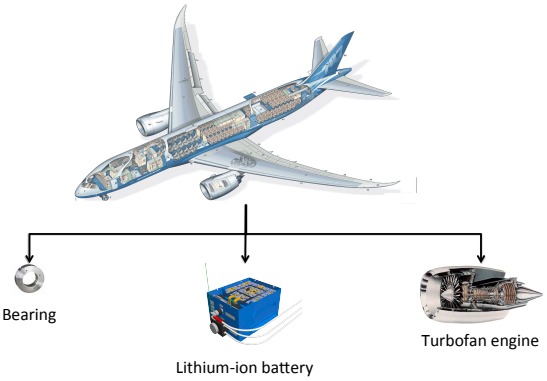

Fig. 6. Example of critical components in a commercial airplane.

possible parameters which can be used to characterize failure mechanism for the critical components shown in Figure 6.

TABLE I
EXAMPLES OF PHYSICAL PARAMETERS FOR CRITICAL COMPONENTS.

Component	Physical parameters
Bearing	Temperature, vibration and acoustics
Lithium-ion batteries	Charge and discharge voltage, charge and discharge current, temperature, voltage and battery impedance
Turbofan engine	Temperature at fan inlet, pressure at fan inlet, physical fan speed, physical core speed and demanded fan speed

3) Selection of monitoring sensors: after choosing the parameters that represent failure propagation, system expert chooses the appropriate sensors to record data from such components. Various sensors, such as micro-sensors, ultrasonic sensors, acoustic emission sensors, etc., have been designed to collect different types of data (Figure 7). The criteria

Fig. 7. Example of different commercial sensors.

of selecting sensors for monitoring a system should take in consideration six aspects, namely: 1) parameters to measure, 2) reliability, 3) accuracy, 4) measurement range, 5) resolution, 6) characteristics and 7) cost . Once the sensors are fixed and the system is operating, the system expert starts collecting data from such system for processing tasks.

4) Data acquisition: it is the process of gathering signals from measurement sources, such as sensors attached to critical components and digitizing the signals for storage on Personal Computers (PC). Generally, data collected from the critical component can be categorized into two main types. 1) Event data: include qualitative information about

the monitored component such as description of installation, breakdown, overhaul, causes etc., and the description of what was done to fix the failure and the severity of the repair and 2) condition monitoring data: measurements related to the health condition/state of the physical asset. They can be vibration data, acoustic data, oil analysis data, temperature, pressure, moisture, humidity, weather or environment data, etc. Event data and condition monitoring data are equally important in PHM. However, in this work we consider only condition monitoring data.

5) Data pre-processing: data acquisition step introduces some errors to the signals due to different kinds of noise, which can be reduced by data pre-processing. Data pre-processing is defined as the process of manipulating raw signals to be suitable for the next stage. It is not used to extract features or reduce dimensions of the raw signals. It is used as a preparation step to enhance the input signal quality and to remove the outliers. In this way, pre-processing raw signals reduces the computational complexity and prepares the signal for better analysis in the later steps. Data pre-processing approaches can be divided in four main groups, such as 1) handling missing data, 2) noise reduction, 3) normalization and 4) smoothing.

B. Remaining useful life estimation of critical components based on Bayesian approaches

Measurements collected from critical components are usually multidimensional time series signals that contain immense number of data. Thus, it is important to first extract information that represent the degradation evolution over time. The relation between the extracted information and EOL should be modeled to predict the RUL. To do this, the proposed method learns the model from the offline data set. It estimates the current health status from new online data and predicts the RUL by measuring the similarity to the offline data. The method is summarized in Figure 8 and will be explained hereafter.

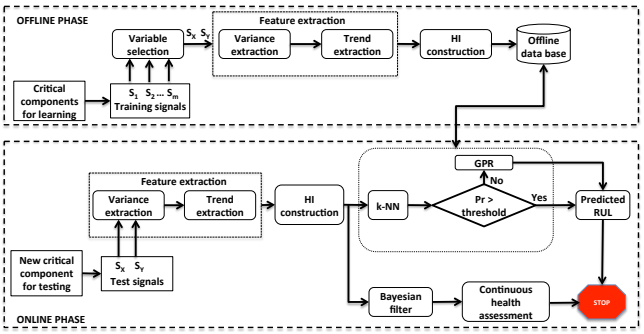

Fig. 8. Overall scheme of the proposed method.

1) Health indicators construction: in this step, different data analysis techniques are used to discover useful information about the degradation process. Such information can be used to construct health indicators (HI). A HI can be defined as a set of features extracted from monitored component which represent the component's degradation evolution as a function of time. In [9] a method for HI construction has been proposed. This method contains three steps which are summarized hereafter.

- Variable selection: this can be done using unsupervised variable selection as proposed in [10]. The method selects the sensor signals that depict information about the degradation evolution over time using symmetrical uncertainty (SU), for all the input signals, defined by:

$$SU(X,Y) = 2 \times \frac{I(X,Y)}{H(X) + H(Y)} \tag{1}$$

  where, $I(X,Y)$ is the mutual information between two random variables X and Y; $H(X)$ and $H(Y)$ are information entropy values of the random variables X and Y respectively.
- Dimensionality reduction: this can be done using projection algorithms such as principal component analysis (PCA) to represent the selected variables in a compact form.
- Trend extraction: this can be done by using empirical mode decomposition algorithm (EMD) to extract the residual signal $r_n(t)$, which should be constant or monotonic signal from the projected variables:

$$r_n(t) = X(t) - \sum_{i=1}^{n} imf_i(t) \tag{2}$$

  where, $X(t)$ is the input signal, imf_i is the intrinsic mode functions (IMF) and n is the maximum number of IMFs [11].

A feature vector $F = [a, b, \bar{x}, s^2]$ is then extracted from each trend at each cycle/time, where a and b are the slope and y-intercept of a linear curve fit of the input trend respectively, $\bar{x}$ and s^2 are the mean and the variance of the input trend respectively [5]. The resulting features are then used to represent each trend according to its EOL time. HIs are constructed for the offline data sets and saved on the data base as reference models. In the online step, the method uses the same previously selected variables and constructs HIs up to the current time. The online HIs are then used to assess the health status of the monitored component.

2) Health assessment: one way to do that is by applying recursive estimating algorithms. Such algorithms estimate the HIs from the online data until it reaches stopping criteria. This can be done first by estimating the health indicators values recursively from the monitored component sensor data using discrete Bayesian filter, see Algorithm 1.

where the input to the algorithm is a discrete probability distribution $\{p_{k,t}\}$ along with the recent measurement z_t, $\bar{p}_{k,t}$ is prediction probability, $p_{k,t}$ is the posterior probability, $p(X_t = x_k | X_{t-1} = x_i)$ is the state transition model and $p(z_t | X_t = x_k)$ is the measurement transition model. Discrete Bayesian filter can be used to represent the uncertainty about the health status in a probabilistic form, which is useful for decision making in later steps [5].

Algorithm 1 Discrete Bayesian filter.

Input : $\{p_{k,t-1}\}$, z_t
Output: $\{p_{k,t}\}$

forall the k **do**
  $\bar{p}_{k,t} = \sum_i p(X_t = x_k | X_{t-1} = x_i) p_{i,t-1}$
  $p_{k,t} = \eta p(z_t | X_t = x_k) \bar{p}_{k,t}$
end

3) Remaining useful life estimation: to estimate the RUL, the method looks for the most similar offline model by using a k-NN classifier:

$$p(C_k|\alpha) = \frac{p(\alpha|C_k) \times p(C_k)}{p(\alpha)} \tag{3}$$

where, α is the new online feature vector, C_k is the class or the group of trends that has similar EOL value, $p(\alpha|C_k)$ is the probability of observing α given C_k (also known as the likelihood), $p(C_k)$ is class priors and $p(\alpha)$ is the marginal likelihood. If the posterior probability is less than a certain threshold, the Gaussian process regression (GPR) model is then used [12]. GPR is defined as follows:

$$f(x) = \mathcal{GP}(m(x), k(x, x') \tag{4}$$

where, $\mathcal{GP}$ is the Gaussian process function defined by a mean function $m(x)$ and a covariance function $k(x, x')$ collected for all possible pairs of the input vector x. The estimated value of the RUL can be used as a stopping criteria for the recursive Bayesian filter. Figure 9 shows the final result of the prognostics method.

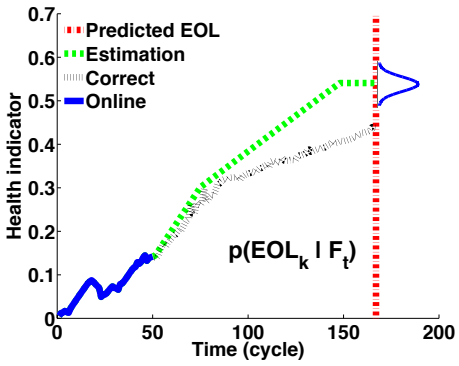

Fig. 9. Final result of the online process for one HI.

III. Applications and Results

Two real life data sets, available and described on the NASA prognostic center of excellence web site (ti.arc.nasa.gov), are used in the experiments: turbofan engine and lithium-ion battery aging data sets.

A. Turbofan engine data

The turbofan engine data sets are generated using commercial modular aero-propulsion system simulation (C-MAPSS). In this work, the data file "train_FD001.txt" is used for offline training and "test_FD001.txt" is used for online testing. Each file contains data for 100 engines and the objective is to predict the number of remaining operational cycles before failure in the test set. The true RUL values for the test data are presented in the data file "RUL_FD001.txt".

1) Variable selection: one of the results of the selection algorithm is the pair of sensors number {8,13}, i.e. physical fan speed and corrected fan speed respectively. The selected group is interesting as the two variables are correlated and both are related to the fan speed. Then, the algorithm starts constructing the monotonic trends iteratively from each pair at each cycle/time.

2) Health indicator construction: as mentioned before, four features are extracted from each trend at each cycle/time and labeled with EOL time to be saved in the offline database. The features represent the relation between the extracted trends. Each trend is then saved in offline database and labeled with the EOL time and will be used for predicting the RUL of new trends.

3) RUL estimation results: to assess the performance of the proposed method, mean absolute percentage error (MAPE) is calculated for all 100 online predictions:

$$E = \frac{100\%}{n} \times \sum_{i=1}^{n} \left| \frac{RUL_i - RUL_i^*}{RUL_i} \right| \tag{5}$$

where, RUL and RUL^* are the actual and predicted RUL values respectively and n is the number of total predictions. The error is calculated only for the last cycles of all 100 test signals. The MAPE over the 100 test data is 11.41%. Figure 10 shows the result of RUL estimation for engine #81 at all cycles.

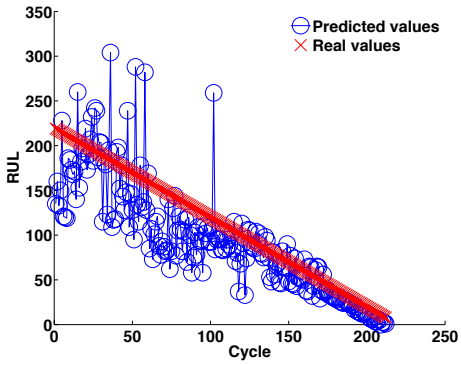

Fig. 10. RUL estimation results of engine #81.

B. Lithium-ion battery data

These data are collected on 34 lithium-ion batteries run through different operational profiles (e.g. charge, discharge and impedance) at different temperatures. In this work only charge and discharge cycles are used. Each cycle is presented by the mean value to reduce the processing time. In order

TABLE II
MEAN ABSOLUTE PERCENTAGE ERROR FOR BATTERY DATA SETS

Fold #1	Fold #2	Fold #3	Average
33.0966%	32.2086%	35.2726%	33.5259%

to validate the proposed method a 3-fold cross-validation is performed, i.e. the available data sets are partitioned into three groups of equal size. Each group is then divided into training and testing data set.

1) Variable selection: one of the results of the selection algorithm is the pair {6, 11}, i.e. the voltage measured at discharge and the capacity of the battery. The selected group is interesting as the two variables are correlated. Also, the capacity is related to the battery health as the decrease in the capacity indicates health degradation.

2) Health indicator construction: four features are extracted from each trend at each cycle/time and labeled with EOL time to be saved in the offline database.

3) RUL estimation results: to assess the performance of the proposed method, MAPE is calculated for all cycles of each battery. The total MAPE per fold is calculated as follows:

$$MAPE_f = \frac{1}{n} \times \sum_{i=1}^{n} MAPE_{i,f} \tag{6}$$

where $MAPE_f$ is the average MAPE for a complete fold, $MAPE_{i,f}$ is the MAPE for test battery i in fold f. The final results are calculated and summarized in Table II. Figure 11 shows a plot of the RUL predicted for the battery B0025. Only 10 cycles were considered as late predictions. Furthermore, the error was decreasing at the later cycles.

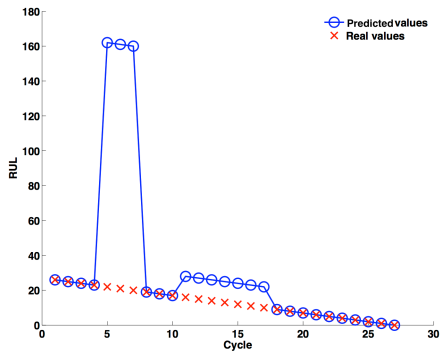

Fig. 11. RUL estimation results of battery #B0025.

IV. CONCLUSION

In this paper, a complete approach for developing a prognostics models for a complex system is presented. The first part of the approach presented different methods that can be used to select the critical component(s) in a complex system, choose the physical parameters from the critical component, choose the monitoring sensors, gather and pre-process sensor data. In the second part, a data driven method for RUL prediction based on Bayesian approaches is presented. The method builds on unsupervised selection of interesting variables from the input offline signals. It construct representative features that can be used as health indicators. The method represents the current status of the online signals as well as the uncertainty about the predictions in a probabilistic form. The performance of the predictions is enhanced by integrating two models, namely k-NN and GPR. The performance of the algorithm is demonstrated using two real data sets taken from the NASA Ames prognostics data repository. The selected variables from the two applications are shown to be interesting. Moreover, the prediction results show low MAPE values for both applications. This paper provided a generic prognostics approach that can be customized and integrated for different applications. It can also be useful for new PHM and data analytics researches to understand the whole process required to develop data-driven prognostics methods for complex systems.

ACKNOWLEDGMENT

This work is supported by the MainPreSI project realized within the framework of the European Territorial Cooperation program INTERREG IV. A France - Switzerland, financed by the European Regional Development Fund.

REFERENCES

[1] K. Goebel, A. Saxena, M. Daigle, J. Celaya, and I. Roychoudhury, "Introduction to prognostics," in *European PHM conference*, 2012, online tutorial, last visited October 2014.

[2] J. Sikorska, M. Hodkiewicz, and L. Ma, "Prognostic modelling options for remaining useful life estimation by industry," *Mechanical Systems and Signal Processing*, vol. 25, no. 5, pp. 1803 – 1836, 2011.

[3] G. Wilcock, "Enhanced fault management for future ima systems," in *IEEE Digital Avionics Systems Conference*, 1997.

[4] P. Kalgren, C. Byington, and M. Roemer, "Defining phm, a lexical evolution of maintenance and logistics," in *IEEE Autotestcon conference*, 2006, pp. 353 – 358, DOI: 10.1109/AUTEST.2006.283685.

[5] A. Mosallam, K. Medjaher, and N. Zerhouni, "Data-driven prognostic method based on bayesian approaches for direct remaining useful life prediction," *Journal of Intelligent Manufacturing*, 2014, published online 13 June 2014, DOI: 10.1007/s10845-014-0933-4.

[6] N. Crutchfield and J. Roughton, *Developing the Job Hazard Analysis, In Safety Culture*. Butterworth-Heinemann, Oxford, 2014, ch. 12, pp. 235–248.

[7] H. R. Booher, *Handbook of Human Systems Integration*, 2003, ISBN: 978-0-471-02053-0.

[8] A. Saxena, I. Roychoudhury, J. Celaya, B. Saha, S. Saha, and K. Goebel, "Requirement flowdown for prognostics health management," in *Proceedings of the AIAA Infotech*, 2012.

[9] A. Mosallam, K. Medjaher, and N. Zerhouni, "Nonparametric time series modelling for industrial prognostics and health management," *Int. J. Adv. Manuf. Technol*, vol. 69, no. 5-8, pp. 1685 – 1699, 2013.

[10] A. Mosallam, S. Byttner, and M. Svensson, "Nonlinear relation mining for maintenance prediction," in *IEEE Aerospace Conference*, 2011, pp. 1 – 9, DOI: 10.1109/AERO.2011.5747581.

[11] N. Huang, Z. Shen, S. Long, M. Wu, H. Shih, Q. Zheng, N. Yen, C. Tung, and H. Liu, "The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis," in *Proceedings of the Royal Society of London Series A. Mathematical, Physical and Engineering Sciences*, 1998, pp. 903 – 995.

[12] A. Mosallam, K. Medjaher, and N. Zerhouni, "Integrated bayesian framework for remaining useful life prediction," in *IEEE International Conference on Prognostics and Health Management, PHM'2014*, 2014, pp. 1 – 6.