Shapelet-based remaining useful life estimation

Simon Malinowski, Brigitte Chebel-Morello and Noureddine Zerhouni
FEMTO-ST Institute, Université de Franche-Comté / ENSMM
[simon.malinowski/brigitte.morello/noureddine.zerhouni]@femto-st.fr

***Abstract*— In the Prognostics and Health Management domain, estimating the remaining useful life (RUL) of critical machinery is a challenging task. Various research topics as data acquisition and processing, fusion, diagnostics, prognostivs and decision are involved in this domain. This paper presents an approach for estimating the Remaining Useful Life (RUL) of equipments based on shapelet extraction and characterization. This approach makes use in a first step of an history of run-to-failure data to extract discriminative rul-shapelets, i.e. shapelets that are correlated with the RUL of the considered equipment. A library of rul-shapelets is extracted from this step. Then, in an online step, these rul-shapelets are compared to different test units and the ones that match these units are used to estimate their RULs. This approach is hence different from classical similarity-based approaches that matches the test units with training ones. Here, discriminative patterns from the training set are first extracted and then matched to test units. The performance of our approach is assessed on a data set coming from a previous PHM Challenge. We show that this approach is efficient to estimate the RUL compared to other approaches.**

I. Introduction

Remaining Useful Life estimation is one of the main task in the Prognostics and Health Management (PHM) domain. The aim of any RUL estimation technique is to provide accurate prediction of the time after which an equipment will not be able to meet its operating requirements. RUL estimation is hence very important for industrial purposes as it can help in adjusting maintenance strategies, maximizing the useful operational life of equipments, reducing maintenance costs and avoiding breakdowns that might have critical impacts.

RUL estimation techniques in the literature are separated into two families : model-based and data-driven approaches. Techniques combining these two approaches are called hybrid methods. Model-based approaches rely on building a physical model describing the behavior of the equipment. These approaches are very accurate but the knowledge of the physical degradation of the system needs to be available, which is not always the case. In addition, model-based approaches are specific to an application and cannot be generalized. Data-driven approaches make use of available run-to-failure data to build models or extract information based on a learning process. These approaches are usually easier to obtain and implement, but are often less accurate than model-based ones. They hence offer a trade-off between accuracy and complexity. Most of the data-driven approaches in the literature are based on machine learning and statistical tools. Neural Networks have been widely considered to model the system and estimate the RUL. These approaches mainly rely on time series prediction ([1], [2], [3], [4] for instance). Hidden Markov models [5], ARMA models [6] have also been considered in the literature. A good survey of machine learning and statistical techniques for prognostics can be found in [7].

Recently, similarity-based approaches have been introduced for the RUL estimation problem. In this kind of approaches, test units (whose remaining lives are to be predicted) are matched to the library of training units (available from run-to-failure data) and the most similar instances are used to estimate the RUL. The authors of [8] have won the 2008 PHM challenge with a similarity based approach that relied on a modified euclidean distance between training and test units. In this method, the whole test trajectory is used to be matched to the library of training units.

Shapelets have been introduced in [9] and [10] for classification and early classification of time series. We extend here the notion of shapelet and define a new kind of shapelets called rul-shapelets, that correspond to patterns carrying information about the remaining useful life. In this paper, we first describe how to extract discriminative rul-shapelets from a training set of time series representing run-to-failure data of an equipment. This extraction step produces a library of rul-shapelets that will be used in the RUL estimation of test time series. The RUL estimation relies here on finding similar behaviors between some parts of test time series and the shapelets that have been extracted because of their correlation with the remaining useful life. Hence, a major difference with other traditional approaches is that the estimation is based on some parts of the test unit (and not the whole test unit or only last instants), these parts being chosen because of their high correlation with the RUL.

This approach is compatible with any applications satisfying the following assumptions:

- Run-to-failure data is available
- Test components are assumed to go through the same degradation process as train component
- Sensory data captures the health status evolution.

The rest of this paper is organized as follows. Section II describes how discriminative shapelets are extracted from a set of time series representing run-to-failure data. Section III explains how these shapelets are used to perform

The authors would like to acknowledge the European Regional Development Fund (FEDER) who funded this work as a part of a project named ALTIDE: Aid to Lifecycle Traceability for Intelligently Developed Equipment..

RUL estimation on test units, and Section IV evaluates the performance of the proposed approach on a data set available online.

II. Shapelet extraction and selection

Let $\mathscr{T}$ be a training set composed of $|\mathscr{T}|$ time series, $T_1,\ldots,T_{|\mathscr{T}|}$. In this set, the time series may have different lengths. The length of the time series T_i is denoted $l(T_i)$. With this notation, a time series T_i in the set $\mathscr{T}$ can be written $T_i = t_1^i, t_2^i, \ldots, t_{l(T_i)}^i$. For sake of clarity, we assume here that the time series T_i are univariate, i.e. $t_j^i \in \mathbb{R}, \forall\, 1 \leq j \leq l(T_i)$. The feature extraction process described in this section can be extended to multivariate time series by applying it to every dimension of the time series.

As we focus in this paper on prognostics (estimation of the remaining useful life of equipments before failure), time series in the set $\mathscr{T}$ represent the monitoring of an equipement's behavior from the beginning up to its failure. In this section, we aim at extracting, from the set $\mathscr{T}$ of time series, features that we will be able to correlate with the remaining length of a time series (time period between the instant when the feature is met and the failure of the equipment).

In the following, we consider features under the form of time series of small length (relatively to the average length of the time series in $\mathscr{T}$), that will be denoted *rul-shapelets*.

Definition 1: A rul-shapelet is defined by a tuple $f = (S, \delta, \mu, \sigma)$, where $S = s_1,\ldots,s_{l(S)}$ is a time series and δ is a distance threshold. μ and σ represent respectively the mean and variance of a Normal distribution that is associated to the shapelet f. This distribution models the remaining length of a time series that is matched by f (cf. Definition 2).

Definition 2: A rul-shapelet $f = (S, \delta, \mu, \sigma)$ is said to match a time series T if there exists a subsequence T' of T (whose length is $l(S)$) such that the euclidean distance between S and T' is less or equal than δ. In other words, f matches T if

$$\exists k \in [1, l(T) - l(S) + 1], \text{s. t.} \sqrt{\sum_{j=1}^{l(S)} (s_j - t_{k+j-1})^2} \leq \delta. \quad (1)$$

The match is hence defined at a time instant k. If more than one k satisfies (1), the instant of the match is defined as the one that leads to the minimal distance.

According to Definition 1 and 2, when a time series T_i is matched by a rul-shapelet $f = (S, \delta, \mu, \sigma)$ at a time instant k, we can estimate the probability density function of the RUL of T_i (from time k) as a Normal distribution $\mathscr{N}(\mu, \sigma)$, i.e. we estimate the length of the time series T_i to follow a Normal distribution $\mathscr{N}(k + \mu, \sigma)$.

In this section, we describe how to extract rul-shapelets from a set of time-series $\mathscr{T}$ and select the ones that convey sufficient and accurate information about the RUL when it matches a time series.

A. Shapelet extraction

To extract a rul-shapelet f, we first need to extract a time series of small length (which represents the feature S of the rul-shapelet) from the set $\mathscr{T}$, and we then need to estimate the other three features associated with f : δ, μ and σ. All the subsequences of lengths $l_1,\ldots,l_N$ from the set $\mathscr{T}$ are first extracted. The lengths l_i are parameters chosen by the user. Depending of the number of time series in $\mathscr{T}$ and their lengths, the number of subsequences extracted here can be very important. In order to keep a reasonable amount of shapelets, subsequences of length $l_i, \forall\, 1 \leq i \leq N$ can be quantized using the K-means algorithm into a smaller number of subsequences. After this step, a set $\mathscr{F} = \{f_1,\ldots,f_{|\mathscr{F}|}\}$ of shapelets is obtained. For all these shapelets, only the first feature S is known for the moment. In this case, f can also be written $f = (S, ?, ?, ?)$. We explain in the following how to obtain the other three features.

B. Shapelet selection

Definition 3: Let $f = (S, ?, ?, ?)$ be a rul-shapelet and T a time series from $\mathscr{T}$. The best-match features (BMF) between f and T is the pair (d, rul), where :

$$\begin{cases} d = & \min\limits_{1 \leq j \leq l(T)-l(S)+1} \sqrt{\sum\limits_{i=1}^{l(S)} (s_i - T_{i+j-1})^2} \\ rul = & l(T) - \arg \min\limits_{1 \leq j \leq l(T)-l(S)+1} \sqrt{\sum\limits_{i=1}^{l(S)} (s_i - T_{i+j-1})^2}. \end{cases} \quad (2)$$

In other words, d is the minimum euclidean distance between S and a subsequence of T of the same length and *rul* is the remaining time before the end of T when this minimum distance d is met. In the following, d will also be denoted best-match distance and *rul* best-match RUL.

We can, according to Definition 3, compute the BMF between a rul-shapelet f and every time series of the set $\mathscr{T}$.

Definition 4: For a rul-shapelet $f = (S, ?, ?, ?)$ and a set $\mathscr{T}$ of time series, we define the best-match features list between f and $\mathscr{T}$ as the list :

$$L_f = \langle bmf_1 = (d_1, rul_1), \ldots, bmf_{|\mathscr{T}|} = (d_{|\mathscr{T}|}, rul_{|\mathscr{T}|}) \rangle, \quad (3)$$

where $bmf_i, 1 \leq i \leq \mathscr{T}$ is the BMF between f and the i^{th} time series of $\mathscr{T}$. This list of pairs is then ordered so that the d_i are in an increasing order ($d_1 \leq d_2 \leq \cdots \leq d_{\mathscr{T}}$).

We are now interested, for a rul-shapelet $f = (S, ?, ?, ?)$, in finding the parameters δ, μ and σ such that when f matches a time series T (i.e. the best-match distance between f and T is lower than δ), we have a high confidence in estimating that the remaining time (from the instant of the match) before the end of the series T follows a Normal distribution $\mathscr{N}(\mu, \sigma)$. For that purpose, we use the best-match features list L_f between f and the set of time series $\mathscr{T}$. From this list, we first extract the list of best-match RULs: $R = \langle rul_1, \ldots, rul_{|\mathscr{T}|} \rangle$. This list R is normalized so

that its average value equals 0 and its variance equals 1. We compute the index i ($2 \leq i \leq |\mathscr{T}|$) defined by:

$$i = \arg \min_{2 \leq j \leq |\mathscr{T}|} var(rul_1, \ldots, rul_j), \quad (4)$$

where *var* denotes the statistical variance. This equation also means that we are searching for the i first elements of R of minimum variance. In order to select only the most discriminative rul-shapelets, the shapelets such that the minimal partial variance (computed at Equation 4) is above a threshold τ can be discarded. When the values of R are normalized, the variance of the whole list is 1 (whatever the considered rul-shapelet). As τ gets close to 0, the more discriminative are the shapelets.
Then, the parameters of the selected rul-shapelet f are computed as :

$$\begin{cases} \delta = d_i \\ \mu = (rul_1 + \cdots + rul_i)/i \\ \sigma = var(rul_1, \ldots, rul_i)^{1/2}, \end{cases} \quad (5)$$

where the values rul_j correspond to the ones before normalization of R.

Example 1:

Figure 1 shows the steps described above to estimate the parameters of a rul-shapelet f. On this exemple, a training set of 100 time series is used. Figure 1-(a) shows the values of the best-match RULs between f and the 100 time series. Note that these values are ordered according to the best match distances as explained above. The right part of Figure 1 shows the partial variances $var(rul_1, \ldots, rul_j), 2 \leq j \leq 100$. From these values, the index $i = 11$ is chosen according to Equation (4). The 11^{st} best-match distance (d_{11} in the best-match features list) is selected as the paramter δ and the 11 first values of the left image are selected (shown by red triangles) to estimate μ and σ according to Equation (5).

After these steps, the set $\mathscr{F} = \{f_1, \ldots, f_{|\mathscr{F}|}\}$ of rul-shapelets is filled, i.e. every f_i is defined with its four parameters.

III. Shapelet-based RUL estimation

In this section, we explain how we perform RUL estimation using a set of rul-shapelets $\mathscr{F}$ obtained as described in the previous section. The context is the following: we are monitoring the behavior of a component and our aim is to predict when this component is likely to face a failure, based on previous experiences. The monitoring of the component is modeled by a time series $U = u_1, \ldots, u_{l(U)}$. No information is given about the last monitoring instant $l(U)$: it can be at an early stage of the life of the component, a late stage or any stage in between.

A. Extracting the rul-shapelets that match U

The first step of the RUL estimation described in this section is to find the rul-shapelets in $\mathscr{F}$ that match the test time series U, and the time instant of the match when applicable. Let $f = (S, \delta, \mu, \sigma)$ be a rul-shapelet of $\mathscr{F}$. The best-match distance between f and U is computed (i.e. the mimimum euclidean distance between S and a subsequence of U of same length as S). If this best-match distance is lower than δ then, according to Definition 2 f matches U. The time instant idx_f of the match (i.e. the time index of the beginning of the subsequence of U that leads to the minimal distance) is stored together with f. If the best-match distance between f and U is greater than δ, then f is discarded. The same operation is repeated for all the rul-shapelets of $\mathscr{F}$, leading to a set $Match(U) = \{(f_1, idx_1), \ldots, (f_k, idx_k)\}$, where $f_i, 1 \leq i \leq k$ is a rul-shapelet and idx_i the time instant when f_i matches U.

B. RUL estimation

Every rul-shapelet in $Match(U)$ conveys an information about the RUL of U given by the probability density function of the Normal distribution. Given a rul-shapelet $f_i = (s_i, \delta_i, \mu_i, \sigma_i)$ in $Match(U)$, we can define the likelihood that the RUL of U is equal to $r, r \in \mathbb{N}$ according to the information brought by f_i as

$$\mathscr{L}(U, r, i) = \frac{1}{\sigma_i \sqrt{2\pi}} e^{-\frac{1}{2}(\frac{r-\mu_i}{\sigma_i})^2} \quad (6)$$

Taking into account all the information brought by the rul-shapelets of $Match(U)$, the likehood that the RUL of U is equal to r is given by

$$\mathscr{L}(U, r) = \sum_{i=i}^{k} \mathscr{L}(U, r, i). \quad (7)$$

Note that weights can also be inserted in Equation (7) to favor for instance rul-shapelets that match U at late time instants (as late time instants are closer to failures than early ones). We use in this paper a weight that correspond to the ratio between the instant when the shapelet is matched and the length of the time series where it is matched. Ratios around 1 mean that the shapelet is matched close to the last instants.

Fianlly, the RUL of U can be estimated by :

$$RUL(U) = \arg \max_{r \in \mathbb{N}} \mathscr{L}(U, r). \quad (8)$$

Figure 2 gives an example of the probability density function obtained following the method described in this section. The estimated RUL here is equal to 156.

IV. Experimental results

A. Turbofan data set

The data used to assess the performance of our approach is the Turbofan engine degradation simulation data set [11] available on the NASA prognostic data repository[1]. We use here the first experiment of this data set. It is composed of 100 training time series that represent 100 comlete run-to-failure monitoring of the same engine model and 100 test time series that represent only a partial monitoring (up to a certain time) of 100 similar engine models. The lengths of the time series in the training set vary between 128 and

[1] http://ti.arc.nasa.gov/tech/dash/pcoe/prognostic-data-repository/

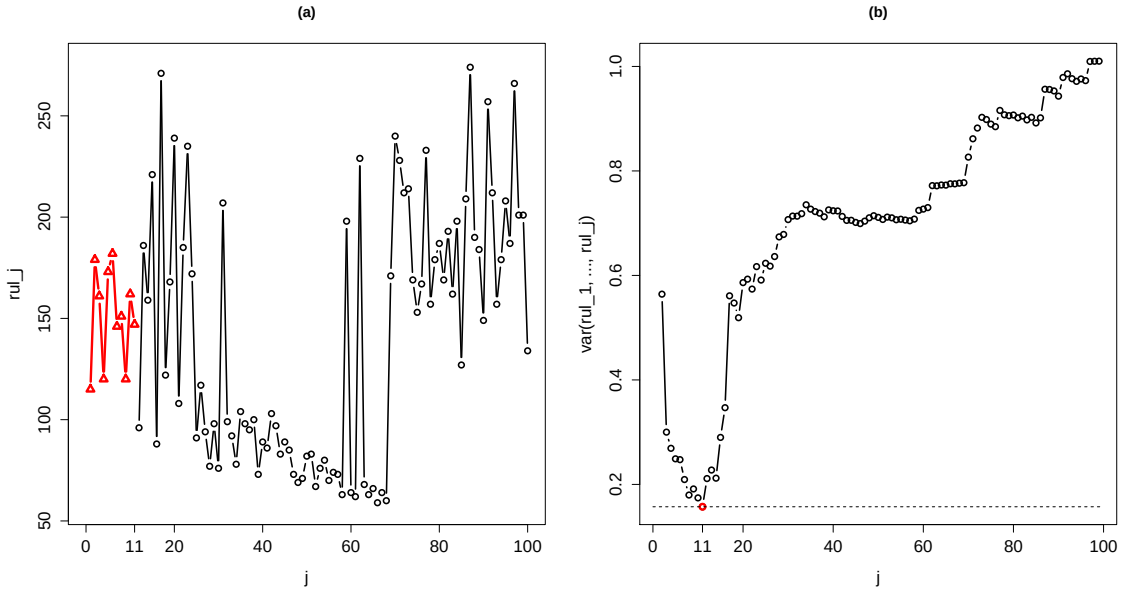

Fig. 1. (a) Best-match RULs obtained between a rul-shapelet and a set of 100 time series, ordered according to the best-match distances (Equation 2) - (b) Partial variances of the best-match RULs of (a). The index of the minimum partial variance is equal to 11 here.

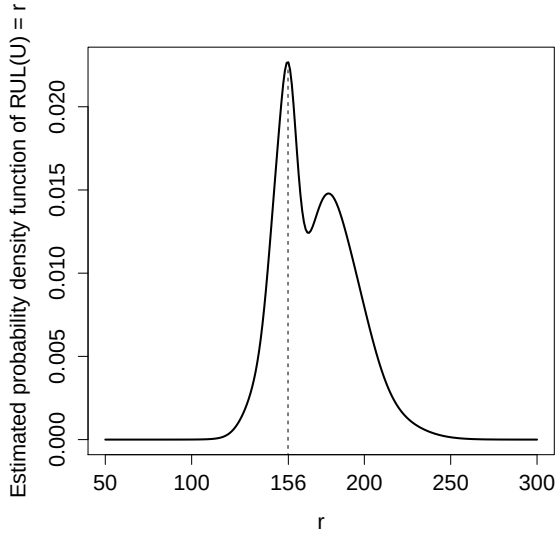

Fig. 2. Example of the probability density function obtained by the proposed method of RUL estimation.

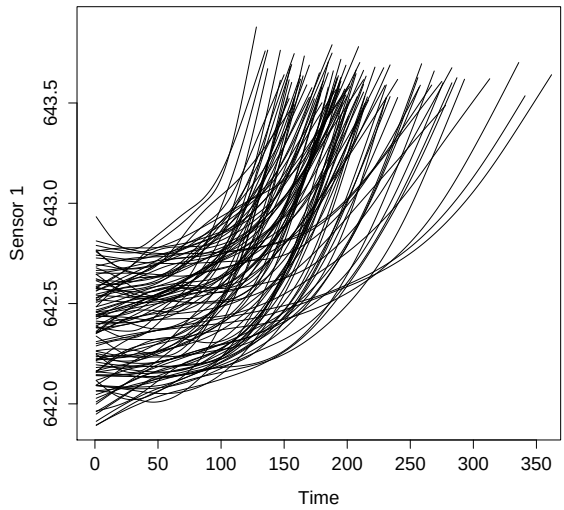

Fig. 3. Evolution of the measures of sensor 1 for the 100 training time series. These curves are obtained after smoothing with a polynomial curve.

362. Our goal is to predict the RUL of the 100 testing time series. Each time series is multivariate : 21 measures (including temperature, pression and speed at various points) are available at every time instant. We have selected only seven sensors according to the study of [8] that pointed out these sensors as the most significative ones. The values given by these sensors are corrupted by noise. We used a third-order polynomial curve to smooth these values and keep only the trends of the time series. The 100 time series given by sensor 1 after smoothing are shown in Fig. 3.

B. Results

Two different approaches are considered here : the first one consists in working directly with the 7-dimensional time series and the second one consists in computing a 1-dimensional time series (health indicator) by using a linear regression on the original time series. This second method was proposed in [8]. The seven values available at each time instant are converted by linear regression into a unique value

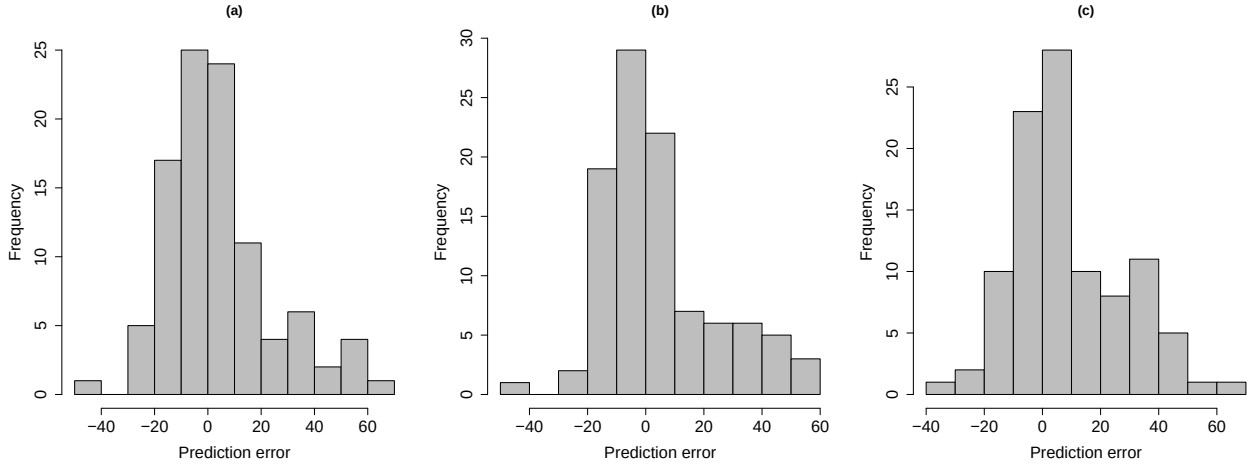

Fig. 5. Histograms of the prediction errors (actual RUL − predicted RUL) obtained by applying (a) an estimation approach based on [8], the proposed approach on (b) the 7-dimensional time series and on (c) the health indicator obtained after linear regression on the original time series.

TABLE I

PERFORMANCE EVALUATION (PERCENTAGE OF CORRECT, EARLY AND LATE PREDICTIONS AND SCORE) OF THE PROPOSED APPROACH AND COMPARISON WITH [8].

Method	correct pred. (%)	early pred. (%)	late pred. (%)	Score of Eqn (11)
Proposed approach (7 sensors)	53	25	22	807.47
Proposed approach (health indicator)	54	33	13	**651.01**
Estimation based on Wang et al. [8]	55	22	23	791.02

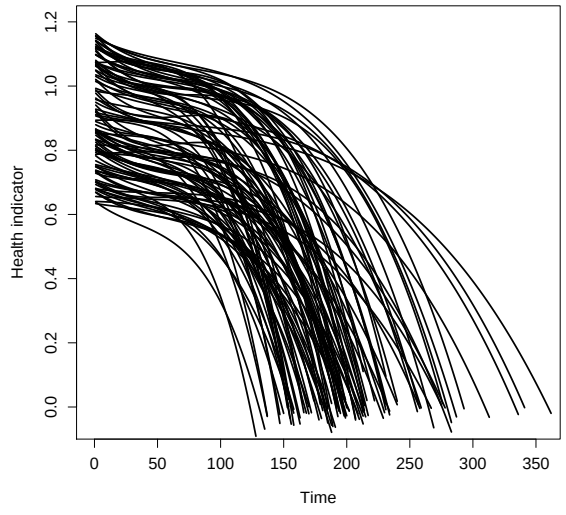

Fig. 4. Health indicators obtained by linear regression for the 100 training time series. These curves are obtained after smoothing with a third-order polynomial curve.

between 0 and 1 that represent the health of the system (value 1 means good health). The health indicators obtained by linear regression are shown in Figure 4. We will compare the effectiveness of both methods in this section.

For each of these approaches, we have built a library of rul-shapelets as explained in Section II that we used to perform the estimation of the RUL of the 100 time series of the test set, using the method explained in Section III. We used 5 differents lengths for the rul-shapelets : $10, 20, 30, 40$ and 50. The threshold τ (Section II-B) is set to 0.35.

As late predictions can have more harmful impact than early ones, a prediction is considered correct if

$$diff = \text{Actual RUL} - \text{Predicted RUL} \in [-10, 13]. \quad (9)$$

This interval was defined by the PHM competition.

Figure 5 represents the histograms of the prediction error when the proposed RUL estimation technique is applied for three different approaches : (a) an estimation approach based on [8], (b) the proposed approach using the 7 dimensional time series and extracting rul-shapelets on each dimension and (c) the proposed approach using the health indicator time series obtained by linear regression. We can observe that the prediction errors are more concentrated around zero for the most right histogram.

A score can also be computed to evaluate the performance of the predictions. The following score S was defined by the

PHM competition :

$$S = \sum_{k=1}^{100} S_k, \tag{10}$$

where

$$S_k = \begin{cases} e^{-diff_k/10} - 1 \text{ if } diff_k \leq 0 \\ e^{diff_k/13} - 1 \text{ if } diff_k > 0, \end{cases} \tag{11}$$

where $diff_k$ is the difference between the actual RUL and the estimated RUL for the k^{th} time series of the testing set.

Table I sums up the performance of the proposed approach and compares it to the method of [8]. For all the different methods the percentage of correct, early and late predictions are given, together with the score given in Equation (11). In the method proposed by Wang et al. [8], the RUL estimation depends on a parameter that determines the number of nearest neighbours kept. A train trajectory is kept if the euclidean distance between this training trajectory and a test one is less than α times the minimum distance between the test trajectory and all the training ones. We fixed this α here to 3 as it gives the best results.

From this table, we can see that using the health indicator approach yields better results than working directly with the sensors. In addition, a better score is obtained using the shapelet-based approach than the one based on [8].

Another advantage of the rul-shapelet based RUL estimation is that the number of parameters in this method is low and they are quite intuitive to fix. Indeed, only the lengths of the shapelets and the threshold τ (to discard non discriminating shapelets) need to be fixed. The lengths of the shapelets depend on the application (size of the training time series) and do not have a huge impact on the performance when reasonnably chosen. The threshold τ is intuitive to fix : a very low $\tau(\leq 0.1)$ will lead to discard a lot of shapelets and might lead to select too few shapelets while a higher $\tau(\geq 0.5)$ may lead to select too many shapelets that are less discriminating.

V. Conclusion

We have proposed in this paper a RUL estimation technique based on shapelet extraction. The shapelet extraction process aims at selecting, from a training set of run-to-failure data, patterns (under the form of small time series) that can be correlated with the remaining time before failure. These extracted patterns convey each an information about the remaining life of the equipment from the instant they are met. These patterns are then used in an online step to estimate the RUL of test units (units for which the remaining useful life is not known). Hence, the RUL estimation is based here on matching discriminative patterns (in terms of RUL estimation) from the training units to test units. This approach was tested on a Turbofan data set and the prediction results showed efficient performance.

References

[1] N. Gebraeel, M. Lawley, R. Liu, and V. Parmeshwaran, "Residual life predictions from vibration-based degradation signals: a neural network approach," *Industrial Electronics, IEEE Transactions on*, vol. 51, no. 3, pp. 694–700, 2004.

[2] R. Huang, L. Xi, X. Li, C. R. Liu, H. Qiu, and J. Lee, "Residual life predictions for ball bearings based on self-organizing map and back propagation neural network methods," *Mechanical Systems and Signal Processing*, vol. 21, no. 1, pp. 193 – 207, 2007.

[3] A. P. Vassilopoulos, E. F. Georgopoulos, and V. Dionysopoulos, "Artificial neural networks in spectrum fatigue life prediction of composite materials," *International Journal of Fatigue*, vol. 29, no. 1, pp. 20 – 29, 2007.

[4] K. Javed, R. Gouriveau, and N. Zerhouni, "Novel failure prognostics approach with dynamic thresholds for machine degradation," in *Industrial Electronics Society, IECON 2013 - 39th Annual Conference of the IEEE*, pp. 4404–4409, 2013.

[5] C. Bunks, D. McCarthy, and T. Al-Ani, "Condition-based maintenance of machines using hidden markov models," *Mechanical Systems and Signal Processing*, vol. 14, no. 4, pp. 597 – 612, 2000.

[6] W. Wu, J. Hu, and J. Zhang, "Prognostics of machine health condition using an improved arima-based prediction method," in *Industrial Electronics and Applications, 2007. ICIEA 2007. 2nd IEEE Conference on*, pp. 1062–1067, 2007.

[7] 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.

[8] T. Wang, J. Yu, D. Siegel, and J. Lee, "A similarity-based prognostics approach for remaining useful life estimation of engineered systems," in *International Conference on Prognostics and Health Management, PHM 2008*, pp. 1–6, 2008.

[9] L. Ye and E. Keogh, "Time series shapelets: a new primitive for data mining," in *Proc. of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 947 – 956, 2009.

[10] Z. Xing, J. Pei, P. S. Yu, and K. Wang, "Extracting interpretable features for early classification on time series," in *Proc. of the Eleventh SIAM International Conference on Data Mining, SDM 2011, April 28-30, 2011, Mesa, Arizona, USA*, pp. 247–258, 2011.

[11] A. Saxena and K. Goebel, "C-MAPSS Data Set," in *NASA Ames Prognostics Data Repository*, 2008.