Received: March 25, 2021. Revised: May 12, 2021.

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

A Singular Spectrum Analysis-based Synthetic Dataset Generation Method for Remaining Useful Life Estimation of Turbo Fan Engines

Peerapol Yuvapoositanon[1]* **Prakit Intachai[2]**

[1]*Mahanakorn Institute of Innovation, Faculty of Engineering and Technology, Mahanakorn University of Technology, Bangkok, Thailand*
[2]*Department of Information and Communication Engineering, Faculty of Engineering and Industrial Technology, Phetchaburi Rajabhat University, Thailand*
* Corresponding author's Email: peerapol@mutacth.com

Abstract: In this paper, we propose a novel method of generating synthetic datasets by means of singular spectrum analysis (SSA) with the optimal window length for substituting the actual datasets that are needed for remaining useful life (RUL) estimation of turbofan engines. The validity of proposed method is confirmed by testing with 200 actual datasets from turbofan engine datasets and 200 synthetic datasets generated by the proposed method in comparison to those generated by three algorithms: the Fourier Decomposition Method (FDM), the Fast Fourier Transform (FFT) and the Empirical Mode Decomposition (EMD). The performance of the SSA-based synthetic datasets for RUL estimation was compared with those of the FFT, EMD and FDM algorithms by means of the regression performed by the Long Short Term Memory (LSTM) neural networks. All the results were measured in terms of the mean absolute error (MAE) and the root mean squared error (RMSE) of their RUL estimates averaged over 200 datasets. The results were also compared with those of the actual feature dataset which provided the MAE of 23.828 and RMSE of 35.284. For the synthetic datasets, the results showed the MAE of 27.126 and RMSE of 38.472 for the FFT, the MAE of 28.362 and RMSE of 39.402 for the EMD and the MAE of 30.410 and RMSE of 41.705 for the FDM. It was revealed that the synthetic datasets generated by the proposed SSA-based method performed the best with the MAE of 25.123 and RMSE of 36.825 confirming the applicability of the proposed SSA-based synthetic datasets in substitution of the actual datasets for RUL estimation.

Keywords: Intelligent system, Synthetic system, Artificial system, Similarity-based, Singular spectrum analysis, Basis functions, RUL estimation, Long short term memory neural network.

1. Introduction

For the past decade, predictive maintenance (PdM) approach for the proactive maintenance scheme has tremendously gained interest as opposed to the traditional preventive or condition-based maintenance approaches [1]. In PdM, the condition of degradations can be tracked via consistently data monitoring and, as the degradation progresses, the end-of-life (EOL) of the machine can be predicted or estimated more effectively [2].

The remaining useful life (RUL), i.e., the length of time from the current time of operation until the EOL is met, needs to be accurately estimated so that planning for maintenance, logistics and safety management can be carried out efficiently [3-5]. However, it is a challenging task to achieve an accurate RUL prediction since RUL is considered as a random variable with possibly unknown distribution [6]. The quest for improvement of accuracy in RUL estimation has been conducted by a number of studies in the prognostics literature [7, 8].

In the data-driven approaches for PdM, datasets acquired from physical sensors are used to indirectly identify the patterns of degradation process needed for RUL estimation [1, 3, 9, 10]. These observed datasets are often called *features* and they can be temperature, pressure, magnetic field, flow and vibration, to name a few [1]. In order to make a

reliable RUL estimation, however, an adequate number of features associated with the run-to-failure data of the system is required [3]. Unfortunately, the run-to-failure data may not be readily available for most of practical systems. Indeed, in most cases, achieving run-to-failure datasets along with the associated features from a healthy system is an almost impossible task due to prohibitively expensive costs and a long period of time required in collecting the run-to-failure data.

One possible solution to this problem is to use *synthetic datasets* which are more readily for analyses and simulations than the actual datasets. [11]. In the literature, synthetic datasets find applications in diversified areas. In [12], an integrated ensemble based imbalance learning model for industrial prognostics was introduced and tested in a real wind turbine failure forecast challenge. The method is based on Synthetic Minority Oversampling Technique (SMOTE) [13], a synthetic technique which can add new minority class examples. In [14], real-valued failure data is generated for prognostics of air purge valve under the conditions of limited failure data availability. In [15], an analysis of characteristic features observable in phasor measurement unit (PMU) measurements obtained from a publically-sourced, industry dataset is carried out for input considerations in the production of realistic, synthetic power systems PMU data. Last but not least, in [16], synthetic daily stream flow time series at multiple sites are generated using Cholesky decomposition.

We may consider the well-known algorithms to be used as a synthetic dataset generation method, i.e., the Empirical Mode Decomposition (EMD) [17], the Fast Fourier Transform (FFT) [18] and the Fourier Decomposition Method (FDM) [19]. However, there are also considerations that need to be addressed before selecting an algorithm for generating synthetic datasets. One challenge when considering the EMD for synthesizing the datasets is the selection of intrinsic mode functions (IMFs); It is unclear in making physical sense of the individual components of the IMFs [20]. The FFT and the FDM are both based on the Fourier analysis. Therefore, their analysis-synthesis processes of a time series rely upon the series of the harmonic functions of the non-adaptive sinusoidal and cosinusoidal functions even in the case of a non-stationary time series [20]. Also, the spectral leakage problem inherent to any Fourier analysis algorithm can result in harmonics estimation errors and some additional measures are needed to eliminate the problem [21].

In this paper, we present a novel singular spectrum analysis (SSA)-based approach for generating the synthetic feature datasets for RUL estimation for Turbo fan engines. The method of constructing the synthetic datasets by using basis functions from the prototypes, i.e., the select time series chosen from available actual features datasets, is described. The validation of the proposed method was performed with the Turbofan engines datasets publicly provided by the Prognostics Center of Excellence (PCoE) at Ames Research Center [22]. Of all the analysis schemes currently being investigated, SSA is one of the most studied techniques for analysis and synthesis of general time series [23]. As a non-parametric method, SSA does not assume any data specification of the datasets, e.g., the levels of noise, observation duration and stationarity properties; hence its applications for practical time series is enhanced [24]. Extended from the principle of the well-established Fourier analysis, the concept of the proposed approach is based on the hypothesis that *any* time series can be synthesized from the linear combination of weighted basis functions derived by means of the SSA analysis-synthesis operations. Unlike the basis functions derived from the Fourier analysis method, however, those of the proposed SSA-based approach are a collection of time series synthesized from the prototypes. RUL estimation from the synthetic datasets can be performed by a subsequent regression method. In this regard, the long short term memory (LSTM) neural network is chosen as a regression method due to its capability in working with a trendy time series [25].

The paper is organized as follows. The methodologies for generating synthetic datasets are presented in Section 2. The relationship between the concepts of the basis functions of the time series in the continuous-time Fourier series analysis and SSA is described and the rationale of using SSA in determining the basis functions of non-stationary time series is also explained. In Section 3, we first describe the mechanism of perfect reconstruction of a time series from the basis functions derived another time series using SSA decomposition-reconstruction processes. Then the core concept of the synthetic dataset generation with scaled basis functions is formulated along with those of the three existing and well-recognized algorithms, i.e., the Empirical Mode Decomposition (EMD) [17], the Fast Fourier Transform (FFT) [18] and the Fourier Decomposition Method (FDM) [19]. In Section 4, the numerical experiments on the RUL estimation using the synthetic datasets generated by the proposed method in comparison with EMD, FFT and FDM are also provided. Finally, conclusions are discussed in Section 5.

2. Methodologies for generating synthetic datasets

In this section, we first formulate the concept of representing a time series by means of the basis functions. The similarities and differences between the concepts of basis functions of the Fourier analysis [19], the Empirical Mode Decomposition (EMD) [17], the Karhunen-Loève transform (KLT) (or the Hotelling transform) [26] and the singular spectrum analysis (SSA) [23, 24] are also discussed.

From the continuous-time Fourier series analysis viewpoint, any periodic continuous-time sequence $x(t)$ with period T can be *synthesized* from the combination of the infinite number of harmonically-related weighted sinusoidal functions [19]

$$x(t) = \frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \cos\left(\frac{2\pi kt}{T}\right) + b_k \sin\left(\frac{2\pi kt}{T}\right) \quad (1)$$

Where a_0 is the constant component of $x(t)$ and a_k, b_k are the k^{th} harmonic coefficients associated with the sinusoidal functions of $x(t)$ and can be determined by

$$a_k = \frac{2}{T}\int_0^T x(t) \cos\left(\frac{2\pi kt}{T}\right) dt, \quad (2)$$

$$b_k = \frac{2}{T}\int_0^T x(t) \sin\left(\frac{2\pi kt}{T}\right) dt. \quad (3)$$

In particular, the second and the third terms in Eq. (1) suggest that any periodic sequence can be formed via the combination of weighted sinusoidal *basis* functions. The benefit of using the sinusoidal basis functions is due largely to their simple behavior under a change of time scale without affecting its amplitudes [27]. Collectively, we use the term Fourier Decomposition Method (FDM) for the method of representing a nonlinear and non-stationary time series using the sinusoidal functions in Eq. (1) as the basis functions. Also, there are analysis schemes available to make FT more meaningful for time-frequency applications, e.g., the short-time Fourier transform, or spectrogram, [28] and the adaptive Fourier decomposition method (AFDM) [29].

The Fourier series analysis is extended to its non-periodic counterpart known as the Fourier transform (FT) where the period T is allowed to be infinity. For efficient analysis of the discrete-time version of FT, the Fast Fourier Transform (FFT) is one of the most commonly used methods and is considered as the most important numerical algorithm of our lifetime [30]. Unlike the Fourier Transform, the Empirical Mode Decomposition (EMD) does not use any fixed type of basis functions such as sine and cosine functions in both decomposition-reconstruction [23, 24]. The decomposition operation in EMD is based on the direct extraction of the energy associated with various intrinsic time scales leading to the collection of the intrinsic mode functions (IMFs) [31]. EMD decomposes a time series $x(t)$ into k numbers of IMFs by the decomposed component $c_k(t)$ and decomposed residue $r_k(t)$. Therefore for EMD, the original time series $x(t)$ is represented as:

$$x(t) = \sum_{k=1}^{N} c_k(t) + r_N(t), \quad (4)$$

where $r_N(t)$ is the total residue from the decomposition.

Similar to FDM and EMD, SSA is regarded as an alternative technique of time series analysis and forecasting [23]. Nevertheless, SSA differs in its principle of operation from the both methods. Unlike Fourier series analysis, the concept of SSA uses the data-driven trajectory matrix to create *adaptive basis* functions. Also, unlike EMD, the concept of oscillatory IMFs and instantaneous frequency are discarded in SSA and the adaptive basis functions are derived by the SVD-based method instead.

It is important to note that the concept of SSA-based basis functions is akin to that of the Karhunen-Loève transform (KLT) which states that a centered (zero-mean) stochastic time series can be thought of as a collection of orthogonal basis functions associated by uncorrelated coefficients [32]. However, SSA and KLT differs in two main considerations. First, KLT needs to compute the theoretical covariance matrix of the centered (zero-mean) input time series. Since there is no fast KLT transform and for a covariance matrix of size N, the calculations must be of the order of N^2 [33]. In contrast, SSA does not need to estimate the covariance matrix of the time series. The multivariate statistical analysis in SSA can be performed with a single time series by means of the collection of lagged copies of a single series in the name of the trajectory matrix [34]. Second, the definition of basis functions of KLT are the eigenvectors derived from the decomposition of the covariance matrix of zero-mean input vectors. But for the SSA-based approach, the basis functions can be the reconstructed elements of the time series itself where each of the basis functions has its own structurally interpretable meanings, i.e., trendy, oscillatory and noise [35].

Such basis functions are beneficial in generating synthetic datasets that resemble to the actual datasets being synthesized.

In the next subsection, we first introduce the system model and the basic SSA process. Then the concept of the SSA-based basis functions is formulated.

2.1 Analysis and synthesis of basis functions in SSA

Consider a real-valued time series $x(t)$ of length T, the trajectory matrix of the time series is arranged as the $L \times K$ Hankel matrix

$$\mathbf{X} = \begin{bmatrix} x(1) & x(2) & x(3) & \cdots & x(K) \\ x(2) & x(3) & x(4) & \cdots & x(K+1) \\ x(3) & x(4) & x(5) & \cdots & x(K+2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x(L) & x(L+1) & x(L+2) & \cdots & x(T) \end{bmatrix}, \quad (5)$$

where L is the window length typically set to $L < T/2$ [34, 36] and $K = T - L + 1$ is the embedding dimension. The optimally estimated window length (L_{est}) used in this paper is explained in [25]. By performing the singular value decomposition (SVD) of $\mathbf{X}$, we arrive at

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T, \quad (6)$$

where $\boldsymbol{U} = (\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_L)$ and $\boldsymbol{V} = (\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_K)$ are the eigenvector matrices. The diagonal matrix $\boldsymbol{D} = diag(\sigma_1, \sigma_2, \ldots, \sigma_L)$ is a collection of the singular values σ_i [24]. The value of σ_i is sorted in a decreasing order, i.e., $\sigma_1 > \sigma_2 > , \ldots, > \sigma_L$ [23]. Collectively, the group of the i^{th} singular value and associated eigenvectors $\{\sigma_i, \boldsymbol{u}_i, \boldsymbol{v}_i\}$ is called the i^{th} eigentriple. From the decomposition process, the trajectory matrix $\boldsymbol{X}$ can be represented by the summation of

$$\boldsymbol{X} = \boldsymbol{X}_1 + \boldsymbol{X}_2 + \cdots + \boldsymbol{X}_d, \quad (7)$$

where $\boldsymbol{X}_i = \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T$ is the i^{th} elementary matrix and $d = rank(\boldsymbol{X})$ is the number of nonzero eigenvalues σ_i. For practical time series, however, $\boldsymbol{X}$ is usually full rank hence $d = \min\{L, K\}$. To reconstruct the original components constituting the trajectory matrix $\boldsymbol{X}$, eigentriple grouping is first performed to arrange multiple elementary matrices $\boldsymbol{X}_i$ from the set of indices $i = 1, \ldots, d$ into m disjoint subsets, i.e., $I_1, \ldots, I_m$. The j^{th} subset has p components, i.e., $I_j = \{j_1, \ldots, j_p\}$. If $m = d$, the grouping is called the elementary grouping and each group has only one member, i.e., $I_j = \{j\}$, $j = 1, \ldots, d$. So the trajectory matrix can be decomposed as $\boldsymbol{X} = \boldsymbol{X}_{I_1} + \cdots + \boldsymbol{X}_{I_m}$ where $\boldsymbol{X}_{I_j} = \boldsymbol{X}_{j_1} + \cdots + \boldsymbol{X}_{j_p}$. The element $x_{l,o}^{(k)}$, $1 \le l \le L$ and $1 \le o \le K$, of $\boldsymbol{X}_{I_k}$ for $I_k = I_1, \ldots, I_m$ is arranged such that the diagonal averaging can be performed. The diagonal averaging operation is defined as

$$\theta_k(t) = \begin{cases} \frac{1}{t}\sum_{q=1}^{t} x_{q,t-q+1}^{(k)} & ; 1 \le t < L, \\ \frac{1}{L}\sum_{q=1}^{L} x_{q,t-q+1}^{(k)} & ; L \le t < K, \\ \frac{1}{T-t+1}\sum_{q=t-K+1}^{L} x_{q,t-q+1}^{(k)} & ; L \le t < K, \end{cases} \quad (8)$$

where $\theta_k(t)$, $k = 1, \ldots, m$ and $t = 1, \ldots, T$, is denoted as the k^{th} *basis function* of the SSA decomposition. The reconstructed version $\hat{x}(t)$ of the original time series $x(t)$ can then be derived by the summation of $\theta_k(t)$, $k = 1, \ldots, m$:

$$\hat{x}(t) = \sum_{k=1}^{m} \theta_k(t). \quad (9)$$

3. The concept of generating synthetic datasets from prototypes

3.1 System model for prototypes

In this section, we describe the mechanism in reconstruction of *any* non-stationary time series by the concept of basis functions developed in the last section. The concept of how to synthesize new time series from another time series is derived from those presented in [25].

We choose to explain this concept via the reconstruction of a pair of uncorrelated time series a, $x^{(a)}(t)$, and time series b, $x^{(b)}(t)$. We designate the time series $\hat{x}^{(a)}(t)$ as prototype a and the time series $\hat{x}^{(b)}(t)$ as prototype b. The estimate $\hat{x}^{(a)}(t)$ can be directly reconstructed by summation of its basis functions $\theta_k^{(a)}(t)$ as in (8). Interestingly, it is possible to also reconstruct $\hat{x}^{(a)}(t)$ by scaling the basis functions of the prototype b, $\theta_k^{(b)}(t)$, with the weights of the prototype a, $w_k^{(a)}(t)$ [25]:

$$\hat{x}^{(a)}(t) = \sum_{k=1}^{m} w_k^{(a)}(t)\theta_k^{(b)}(t). \quad (10)$$

The weight, $w_k^{(a)}(t)$ in Eq. (10) is determined by the knowledge of both the basis functions $\theta_k^{(a)}(t)$ and $\theta_k^{(b)}(t)$ where $\theta_k^{(b)}(t) \neq 0, \forall k$,

$$w_k^{(a)}(t) = diag\left(\theta_k^{(b)}(t)\right)^{-1} diag\left(\theta_k^{(a)}(t)\right), \quad (11)$$

In the same fashion, the estimate $\hat{x}^{(b)}(t)$ can be indirectly represented by means of the basis functions of prototype a, $\theta_k^{(a)}(t)$, scaled by the weights of the prototype b, $w_k^{(b)}(t)$:

$$\hat{x}^{(b)}(t) = \sum_{k=1}^{m} w_k^{(b)}(t)\theta_k^{(a)}(t). \quad (12)$$

Also, the weights $w_k^{(b)}(t)$ in (11) is determined by the knowledge of both of the basis functions $\theta_k^{(a)}(t)$ and $\theta_k^{(b)}(t)$ where $\theta_k^{(a)}(t) \neq 0, \forall k$,

$$w_k^{(b)}(t) = diag\left(\theta_k^{(a)}(t)\right)^{-1} diag\left(\theta_k^{(b)}(t)\right). \quad (13)$$

where $diag(\cdot)$ represents the matrix diagonalization operation and $(\cdot)^{-1}$ the inverse matrix operation. Eqs. (10)-(13) constitute the core idea of synthesizing a time series using the basis functions and the coefficients obtained from another time series [25]. The model of the analysis and synthesis operations is illustrated in Fig. 1 and also described in Algorithm 1.

3.2 System models for synthetic datasets

In this section, we describe how a dataset whose characteristics descended from a prototype can be

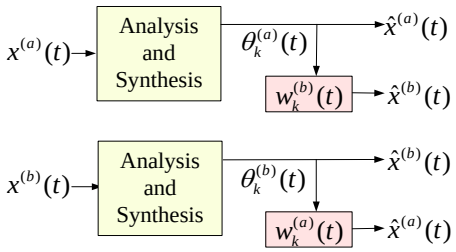

Figure.1 The illustration of the synthesizing process of $\hat{x}^{(a)}(t)$ and $\hat{x}^{(b)}(t)$ by means of the basis functions $\theta_k^{(a)}(t)$, $\theta_k^{(b)}(t)$ and the weights $w_k^{(a)}(t)$, $w_k^{(b)}(t)$ from the prototypes $x^{(a)}(t)$, $x^{(b)}(t)$ which derived from the SSA analysis and synthesis. Notice that $\hat{x}^{(a)}(t)$ *can* be reconstructed by means of $\theta_k^{(b)}(t)$ and $w_k^{(a)}(t)$ and $\hat{x}^{(b)}(t)$ *can* be reconstructed by means of $\theta_k^{(a)}(t)$ and $w_k^{(b)}(t)$

Algorithm 1: Synthesizing of time series $\hat{x}^{(a)}(t)$ and $\hat{x}^{(b)}(t)$ by means of the basis functions $\theta^{(a)}(t)$ and $\theta^{(b)}(t)$.

Input : Prototype time series $x^{(a)}(t)$ and $x^{(b)}(t)$.
Output : Estimated time series $\hat{x}^{(a)}(t)$ and $\hat{x}^{(b)}(t)$.
Let L and calculate $K = T - L + 1$.
For $l = 1$ to L
For $k = 1$ to K
Create trajectory matrix $\mathbf{X}^{(a)}$ and $\mathbf{X}^{(b)}$ by (5).
End
End
Perform SVD as $svd(\mathbf{X}^{(a)})$ and $svd(\mathbf{X}^{(b)})$.
While $i <= d$ do
$\boldsymbol{X}_i^{(a)} = \sigma_i^{(a)}\boldsymbol{u}_i^{(a)}\boldsymbol{v}_i^{T(a)}$.
$\boldsymbol{X}_i^{(b)} = \sigma_i^{(b)}\boldsymbol{u}_i^{(b)}\boldsymbol{v}_i^{T(b)}$.
End
For $k = 1$ to m
Derive the basis functions $\theta_k^{(a)}(t)$ and $\theta_k^{(b)}(t)$ by (8).
End
For $k = 1$ to m
$w_k^{(a)}(t) = diag\left(\theta_k^{(b)}(t)\right)^{-1} diag\left(\theta_k^{(a)}(t)\right)$.
$w_k^{(b)}(t) = diag(\theta_k^{(a)}(t))^{-1} diag(\theta_k^{(b)}(t))$.
End
Estimate time series $\hat{x}^{(a)}(t)$ by (10) and $\hat{x}^{(b)}(t)$ by (12).

synthesized via the collection of m basis functions of the f^{th} feature $x^{(f)}(t)$, i.e., $\theta_k^{(f)}(t)$ $k = 1, \ldots, m$. In this regard, the s^{th} synthetic dataset for the f^{th} feature for at time t generated by the SSA algorithm is generated from the basis functions $\theta_k^{(f)}(t)$ as

$$\tilde{x}_{(\mathrm{SSA})}^{(s,f)}(t) = \sum_{k=1}^{m} w_k^{(s,f)}(t)\theta_k^{(f)}(t), \quad (14)$$

where $\tilde{x}_{(\mathrm{SSA})}^{(s,f)}(t)$ is the synthetic dataset generated by the SSA algorithm and the weight function for the s^{th} synthetic dataset for the f^{th} feature $w_k^{(s,f)}(t), k = 1, \ldots, m$ is a normal distribution function with a mean of one and a variance of δ^2,

$$w_k^{(s,f)}(t) \sim \mathcal{N}(1, \delta^2). \quad (15)$$

Note that $w_k^{(s,f)}(t)$ is a random variable introduced to provide diversity in the synthetic datasets being generated.

For general representations, Eq. (14) can be written in the matrix-vector representation. First, the $L_{est} \times T$ basis function matrix for the f^{th} feature is defined as

$$\boldsymbol{\theta}^{(f)} = \begin{bmatrix} \theta_1^{(f)}(1) & \cdots & \theta_1^{(f)}(T) \\ \vdots & \ddots & \vdots \\ \theta_m^{(f)}(1) & \cdots & \theta_m^{f}(T) \end{bmatrix}, \quad (16)$$

where $f = 1,2,\ldots,F$ and F is the number of recorded feature datasets. The $L_{est} \times T$ matrix of the weights is defined as

$$\mathbf{W}^{(s,f)} = \begin{bmatrix} w_1^{(s,f)}(1) & \cdots & w_1^{(s,f)}(T) \\ \vdots & \ddots & \vdots \\ w_m^{(s,f)}(1) & \cdots & w_m^{(s,f)}(T) \end{bmatrix}, \quad (17)$$

where $s = 1,2,\ldots,S$ and S is the number of synthetic datasets to be generated. As a result, the s^{th} synthetic dataset for the f^{th} feature generated by the SSA algorithm $\tilde{\mathbf{x}}_{(\mathrm{SSA})}^{(s,f)}$ is then determined by

$$\tilde{\mathbf{x}}_{(\mathrm{SSA})}^{(s,f)} = \left(\mathbf{W}^{(s,f)} \circ \boldsymbol{\theta}^{(f)}\right)^T \mathbf{1}, \quad (18)$$

where $\circ$ is the Hadamard (element-wise) product and $\tilde{\mathbf{x}}_{(\mathrm{SSA})}^{(s,f)} = [\,\tilde{x}_{(SSA)}^{(s,f)}(1) \quad \tilde{x}_{(SSA)}^{(s,f)}(2) \quad \cdots \quad \tilde{x}_{(SSA)}^{(s,f)}(T)]^T$ is a synthetic datasets vector and $\mathbf{1} = [\,1 \;\; 1 \;\; \cdots 1]^T$ is an $m \times 1$ vector of ones.

We can also generate the s^{th} synthetic time series for the prototype $x^{(f)}(t)$ by means of the decomposed component and the residue of the EMD algorithm as

$$\tilde{x}_{(\mathrm{EMD})}^{(s,f)}(t) = \sum_{k=1}^{m} w_k^{(s,f)}(t) c_{EMD,k}^{(f)}(t) + r_m^{(f)}(t), \quad (19)$$

where $\tilde{x}_{(\mathrm{EMD})}^{(s,f)}(t)$ is the synthetic datasets generated by the EMD algorithm, $c_{EMD,k}^{(f)}(t)$ is the k^{th} decomposed component from $x^{(f)}(t)$ and $r_m^{(f)}(t)$ is total of residue from decomposition of $x^{(f)}(t)$.

The next algorithm that can be used to generate synthetic datasets is the FDM algorithm. Since the reconstruction of $x^{(f)}(t)$ by FDM is essentially based on the concept of Fourier series, both the harmonic coefficients and the basis functions are required in the synthesis process. The s^{th} synthetic dataset for the f^{th} feature generated by FDM is therefore

$$\tilde{x}_{(\mathrm{FDM})}^{(s,f)}(t) = \frac{a_0}{2} + \sum_{k=1}^{m} w_k^{(s,f)}(t) \boldsymbol{\alpha}_{FDM}^T(k) \mathbf{c}_{FDM,k}^{(f)}(t), \quad (20)$$

where $\tilde{x}_{(\mathrm{FDM})}^{(s,f)}(t)$ is the synthetic datasets generated by the FDM algorithm, a_0 is the constant component and $\boldsymbol{\alpha}_{FDM}(k) = [a_k \quad b_k]^T$ is the vector of the k^{th} harmonic coefficients as mentioned in Eq. (2) and Eq. (3) and $\mathbf{c}_{FDM,k}^{(f)}(t) = \left[\cos(\frac{2\pi kt}{T}) \quad \sin(\frac{2\pi kt}{T})\right]^T$ is the vector of the k^{th} FDM basis functions.

The last algorithm to be examined is the Fast Fourier Transform (FFT) algorithm [18]. Unlike all the above-mentioned algorithms, the FFT coefficients are generally complex valued therefore they cannot be directly operated with the real-valued weight $w_k^{(s,f)}(t)$. One possible solution to this problem is to symmetrically rearrange the time series $x^{(f)}(t)$ into a $(2m-1) \times 1$ vector prior to the FFT operation as

$$\acute{\boldsymbol{x}}(t) = [\acute{x}(0) \quad \acute{x}(1) \quad \cdots \quad \acute{x}(2m-2)]^T = [x^{(f)}(0) \cdots x^{(f)}(m-1) : x^{(f)}(m-1) \cdots \; x^{(f)}(1)\,]^T. \quad (21)$$

The k^{th} FFT coefficient for $\acute{\boldsymbol{x}}(t)$ is then given by

$$X(k) = \acute{\boldsymbol{x}}(t)^T \mathbf{c}_{FFT,k}^{*(f)}(t), \quad (22)$$

where $\mathbf{c}_{FFT,k}^{(f)}(t)$ is a $(2m-1) \times 1$ vector of the k^{th} basis function for the FFT operation,

$$\mathbf{c}_{FFT,k}^{(f)}(t) = \left[1 \quad e^{\frac{2\pi k}{2m-1}} \quad \ldots \quad e^{\frac{2\pi k(2m-2)}{2m-1}}\right]^T, \quad (23)$$

and $(\cdot)^*$ is the conjugation operator. With this arrangement of $x^{(f)}(t)$, the FFT coefficients $X(k)$ for $k = 0,\ldots,2m-2$ are all now constrained to be real valued numbers. In order to generate synthetic datasets from $X(k)$, $w_k^{(s,f)}(t)$ must also be arranged into a $(2m-1) \times 1$ vector as

$$\acute{\boldsymbol{w}}_k^{(s,f)}(t) = \left[\acute{w}_k^{(s,f)}(0) \quad \acute{w}_k^{(s,f)}(1) \cdots \acute{w}_k^{(s,f)}(2m-2)\right]^T = \left[w_k^{(s,f)}(0) \cdots w_k^{(s,f)}(m-1) : w_k^{(s,f)}(m-1) \cdots w_k^{(s,f)}(1)\right]^T. \quad (24)$$

The synthetic time series generated by the FFT algorithm is therefore

$$\tilde{\acute{x}}(t) = \sum_{k=0}^{2m-2} \acute{w}_k^{(s,f)}(t) X(k)\, e^{\frac{2\pi kt}{2m-1}} = \left(\acute{\boldsymbol{w}}_k^{(s,f)}(t) \circ \boldsymbol{X}(k)\right)^T \mathbf{c}_{FFT,k}^{(f)}(t), \quad (25)$$

where $\boldsymbol{X}(k) = [X(0),\ldots,X(2m-2)]^T$ is a $(2m-1) \times 1$ vector of the FFT coefficients $X(k), k = 0,\ldots,2m-2$.

Finally, the s^{th} synthetic dataset for the f^{th} feature by FFT is then a truncated version of $\tilde{\acute{x}}(t)$ for $t = 0,\ldots,m-1$,

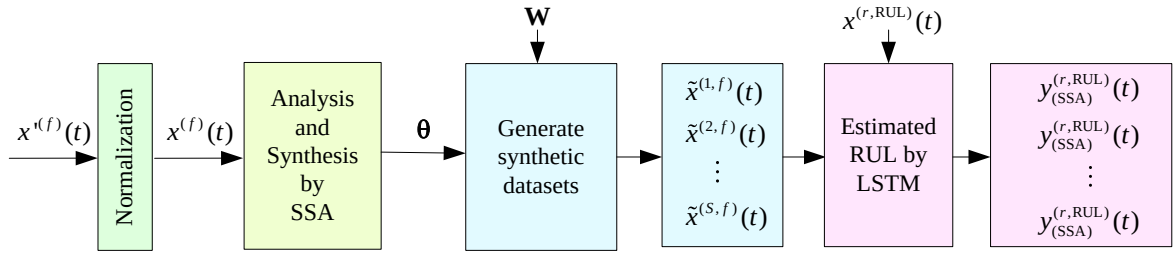

Figure.2 The system of SSA-LSTM RUL estimation with synthetic datasets is illustrated. The prototype features datasets $x^{(f)}(t)$ are applied to from analysis and synthesis parts of SSA for generation of synthetic datasets $\tilde{x}^{(s,f)}(t)$. The feature $\tilde{x}^{(s,f)}(t)$ for synthetic datasets and RUL $x^{(\mathrm{RUL})}(t)$ for prototype datasets are used by LSTM in order to generate the estimated RUL $y^{(\mathrm{RUL})}_{(\mathrm{SSA})}(t)$

$$\tilde{x}^{(s,f)}_{(\mathrm{FFT})}(t) = \{\tilde{\hat{x}}(t),\ t = 0, \dots, m-1\}. \quad (26)$$

where $\tilde{x}^{(s,f)}_{(\mathrm{FFT})}(t)$ is the synthetic datasets generated by the FFT algorithm.

Algorithm 2 summarizes the methods for generating the s^{th} synthetic dataset for the f^{th} feature for the four above-mentioned algorithms.

3.3 The singular spectrum analysis and long short term memory (SSA-LSTM) neural network RUL estimation for synthetic time series

In [25], the hybrid system of SSA-LSTM has been proposed and is shown to enhance performance of RUL estimation for the Turbofan datasets of [22] as compared to the systems utilizing either SSA or LSTM alone. In this paper, the SSA-LSTM RUL estimation architecture of [25] is revisited for performance comparison testing of the synthetic dataset generation algorithms developed in Section 3.2. The complete system is shown in Fig.2.

Algorithm 2: Generation of synthetic datasets from SSA, FFT, EMD and FDM algorithms

Input: Time series $x^{(f)}(t)$.

Output: Estimated features of synthetic time series $\tilde{x}^{(s,f)}(t)$ for SSA, $\tilde{x}^{(s,f)}_{(\mathrm{EMD})}(t)$ for EMD, $\tilde{x}^{(s,f)}_{(\mathrm{FDM})}(t)$ for FDM and $\tilde{x}^{(s,f)}_{(\mathrm{FFT})}(t)$ for FFT.

Analysis basis function of SSA by $\theta^{(f)}_i(t)$ derived by (8).
Analysis component of EMD derived by (4) and complex exponentials of CTFT derived by (1).
Let S and $s = 1,2, \dots, S$.
For $k = 1$ to m
  $w^{(s,f)}_k(t) \sim \mathcal{N}(0, \delta^2)$.
End
For $s = 1$ to S
  For $f = 1$ to F
    $\tilde{x}^{(s,f)}_{(\mathrm{SSA})}(t) = \left(\mathbf{W}^{(s,f)} \circ \boldsymbol{\theta}^{(f)}\right)^T \mathbf{1}$.
    $\tilde{x}^{(s,f)}_{(\mathrm{EMD})}(t) = \sum_{k=1}^{m} \left[w^{(s,f)}_k(t) c^{(f)}_k(t) + r^{(f)}_m(t)\right]$.
    $\tilde{x}^{(s,f)}_{(\mathrm{FDM})}(t) = \sum_{k=1}^{m} \left[w^{(s,f)}_k(t) X(k) \cdot e^{j\omega_0 kt}\right]$.
    $\tilde{x}^{(s,f)}_{(\mathrm{FFT})}(t) = \tilde{\hat{x}}(t),\ t = 0, \dots, m-1$, where
      $\tilde{\hat{x}}(t) = \left(\dot{\boldsymbol{w}}^{(s,f)}_k(t) \circ \boldsymbol{X}(k)\right) \boldsymbol{c}^{(f)}_{FFT,k}(t)$.
  End
End

In this paper, the commonly used min-max normalization method is adopted to obtain a normalized version of a feature. The f^{th} feature prototype $x^{(f)}(t)$ results from the min-max normalization of $x'^{(f)}(t)$ as

$$x^{(f)}(t) = \frac{x'^{(f)}(t) - x'^{(f)}_{\min}}{x'^{(f)}_{\max} - x'^{(f)}_{\min}}, \quad (27)$$

where $x'^{(f)}(t)$ is the f^{th} raw feature data with the maximum and the minimum values being $x'^{(f)}_{\max}$ and $x'^{(f)}_{\min}$ respectively. The normalized prototype feature $x^{(f)}(t)$ is used as the system input of analysis and synthesis by SSA as in Fig. 2.

The f^{th} feature prototype $x^{(f)}(t)$ is then analyzed and synthesized to create the basis function of the f^{th} feature $\theta^{(f)}(t)$. The f^{th} feature of the s^{th} synthetic dataset $\tilde{x}^{(s,f)}(t)$ is generated from the j^{th} basis function $\theta^{(j)}(t)$ with the f^{th} feature of the s^{th} synthetic dataset weight parameter $w^{(s,f)}_i(t)$ as described in Eq. (15). The features for synthetic datasets $\tilde{x}^{(s,f)}(t)$ and true RUL datasets $x^{(r,\mathrm{RUL})}(t)$ for $r = 1,2, \dots, 200$ are assigned in the training networks of LSTM, and the model of RUL estimation is then used in LSTM testing networks to achieve the estimated RUL datasets $y^{(r,\mathrm{RUL})}_{(\mathrm{SSA})}(t), r = 1,2, \dots, 200$.

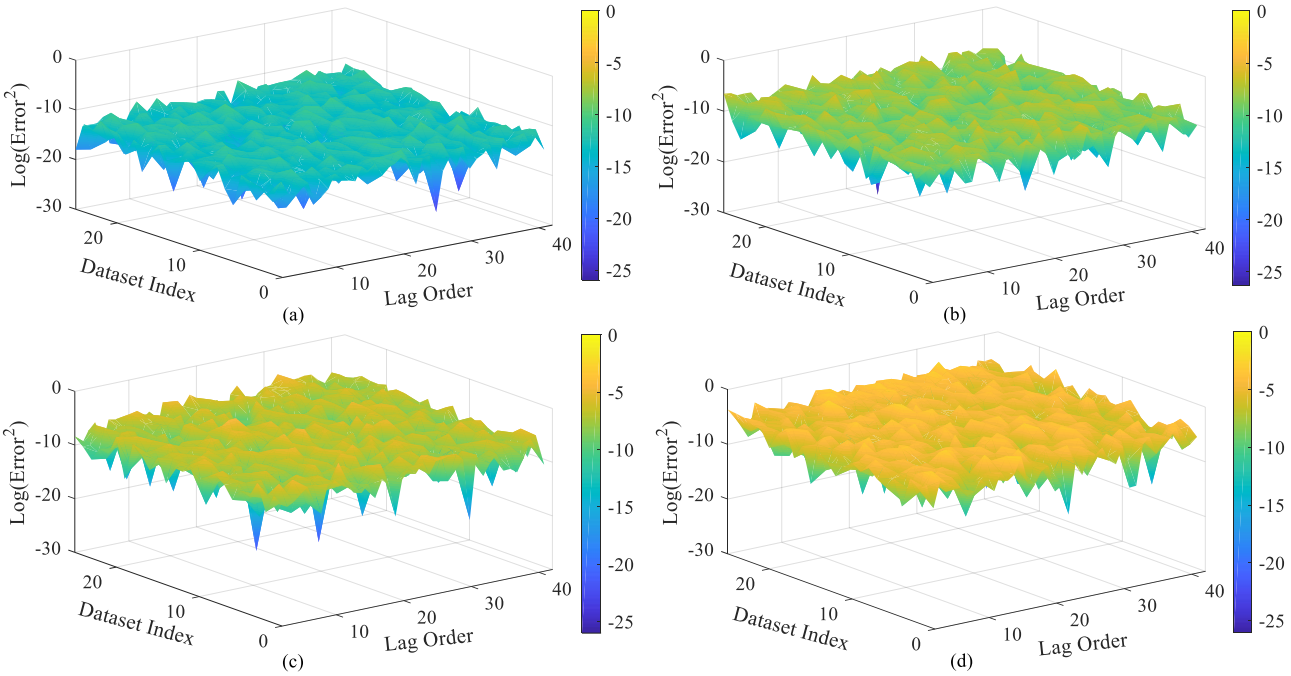

Figure. 3 The log squared cross-correlation errors across 40 lag orders and 25 actual dataset indexes of the synthetic datasets generated by (a) SSA (b) FFT (c) EMD and (d) FDM

This also applies to $y_{(\mathrm{EMD})}^{(r,\mathrm{RUL})}(t)$, $y_{(\mathrm{FDM})}^{(r,\mathrm{RUL})}(t)$ and $y_{(\mathrm{FFT})}^{(r,\mathrm{RUL})}(t)$ by the synthetic models for EMD, FDM and FFT from the synthetic datasets $\tilde{x}_{(\mathrm{EMD})}^{(s,f)}(t)$, $\tilde{x}_{(\mathrm{FDM})}^{(s,f)}(t)$ and $\tilde{x}_{(\mathrm{FFT})}^{(s,f)}(t)$ respectively.

4. Experimental results

4.1 The cross-correlation error as a performance measure for a synthetic dataset

Initially in this section, we need to establish a performance measure to determine how close the statistics of the synthetic datasets generated by SSA, FFT, EMD and FDM algorithms are to those of the actual datasets. In [38], the cross-site correlation of real-space hydrologic variables between the synthetic and historical datasets was used as a performance measure of the data generation of synthetic streamflows. In a similar yet different fashion, we tested the cross-correlation of the synthetic datasets generated by each of the four algorithms and the actual datasets with the cross correlation of the prototype dataset and the actual datasets.

The difference between the two cross-correlations across the lag orders and the dataset indexes or *the cross-correlation error* is therefore the deviation in the statistics of the synthetic dataset from the prototype, i.e., a select feature, referenced to the actual datasets. The m-lagged cross-correlation of the prototype derived from the f^{th} feature, $x^{(f)}(t)$, and the j^{th} actual dataset, $x^{(j)}(t)$, is given by

$$R(o,j) = E\{x^{(f)}(t+o)x^{(j)}(t)\}, \qquad (28)$$

where o denotes the lag of the two time series. In the same fashion, the o-lagged cross correlation of the prototype-based f^{th} feature of the s^{th} synthetic dataset of algorithm A, $\tilde{x}_{(A)}^{(s,f)}(t)$, and the j^{th} actual dataset, $x^{(j)}(t)$, is given by

$$\tilde{R}_{(A)}(o,j) = E\left\{\tilde{x}_{(A)}^{(s,f)}(t+o)x^{(j)}(t)\right\}, \qquad (29)$$

where $A \in \{SSA, FFT, EMD, FDM\}$ represents the synthetic dataset generation algorithm. The difference between $R(o,j)$ and $\tilde{R}_{(A)}(o,j)$ at the m^{th} lag and the j^{th} actual dataset is calculated by the squared error of both cross-correlations:

$$\varepsilon_{(A)}^2(o,j) = \left(R(o,j) - \tilde{R}_{(A)}(o,j)\right)^2. \qquad (30)$$

In Fig. 3 (a)-(d), the surface plots of the logarithm of $\varepsilon_{(A)}^2(o,j)$ of the four algorithms, i.e., SSA, FFT, EMD and FDM, across 40 lag orders and 25 actual dataset indexes are shown respectively. Notice that all the plots were negatively-oriented meaning that lower values were preferable. Among the four algorithms, the average level of the log of errors

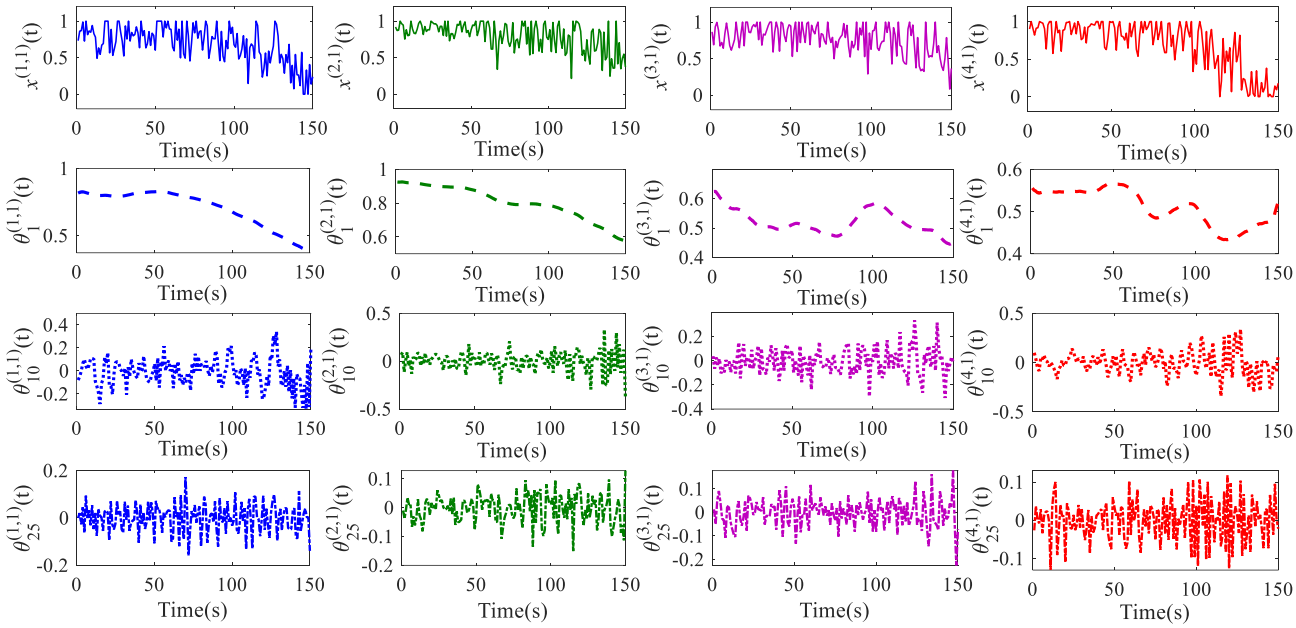

Figure. 4 The plots of the basis functions $\theta_k^{(r,1)}(t)$ for $k = 1,10$ and 25 using $L_{est} = 24$ derived from the SSA algorithm for the 1^{st} normalized feature $x^{(r,1)}(t)$ are shown in four columns for four prototypes, i.e., $r = 1, \ldots, 4$. Starting from the 1^{st} (leftmost) column to the 4^{th} (rightmost) column, $x^{(r,1)}(t)$ for $r = 1, \ldots, 4$ are plotted consecutively at the 1^{st} (topmost) row of each column. In each column, the basis functions $\theta_k^{(r,1)}(t)$ for $k = 1,10$ and 25 are plotted consecutively in panels from the 2^{nd} to the 4^{th} (bottommost) rows

$\log(\varepsilon_{(SSA)}^2(o,j))$ for the synthetic datasets generated by SSA as shown in Fig. 3 (a) was the lowest at -13.60. This indicated that the SSA algorithm was able to generate synthetic datasets most similar to the prototype. For both the synthetic datasets generated by FFT and EMD, the similarities to the prototype were less pronounced as the average levels of the log of errors $\log(\varepsilon_{(FFT)}^2(o,j))$ and $\log(\varepsilon_{(EMD)}^2(o,j))$ shown in Fig. 3 (b)-(c) were at -9.61 and -8.38 respectively. Finally, for the synthetic datasets generated by FDM, the average level of the log of errors $\log(\varepsilon_{(FDM)}^2(o,j))$ in Fig. 3 (d) was at -6.34 suggesting that the synthetic datasets generated by FDM were the most dissimilar to the prototype.

4.2 The remaining useful life (RUL) estimation performance testing of synthetic datasets

We selected four datasets from the 200 turbofan datasets publicly provided by the Prognostics Center of Excellence (PCoE) at Ames Research Center [22] to be the prototypes for performance testing of the synthetic datasets. Each of the prototypes $x^{(r,\mathrm{RUL})}(t)$, $r = 1,2,3,4$ was normalized before being processed.

The f^{th} reconstructed features of prototype $\hat{x}^{(r,f)}(t)$ and RUL of prototype datasets $x^{(r,\mathrm{RUL})}(t)$ for $r = 1,2,3,4$ were applied to the LSTM network for RUL estimation. The true RUL $x^{(r,\mathrm{RUL})}(t)$ of the four prototypes contained $T = 116$ time cycles. $\hat{x}^{(r,f)}(t)$ was then arranged to be the trajectory matrix as in Eq. (5) with the window length L as the design parameter for generating trajectory matrix [23, 24].

In [25], the value of L that associates with the minimum root mean squared error (RMSE) for the particular datasets is $L_{est} \approx 24$, and the range of minimum errors from four prototype datasets is 20 to 30 or $20 \leq L_{est} \leq 30$. The value of $L_{est} \approx 24$ and the range of $20 \leq L_{est} \leq 30$ coincide nicely with the suggested values of $L_{est} < T/2$ in [23], $L_{est} \approx T/4$ in [34], $L_{est} > 2\sqrt{T}$ in [36,39] and $L_{est} = \log(T)^c$ in [38,40].

Three basis functions of the four prototype datasets functions $\theta_k^{(r,1)}(t)$ for $k = 1,10$ and 25 using $L_{est} = 24$ of the 1^{st} normalized feature $x^{(r,1)}(t)$ for four prototypes, i.e., $r = 1, \ldots, 4$, with $L_{est} = 24$ are shown in Fig.4. Notice that at the same k the basis functions for each prototype behave in a similar fashion; for $k = 1$, all the basis functions represent the trends for the prototypes and as k is

Table 1. MAE and RMSE of the estimated RUL $y_{(\mathrm{SSA})}^{(r,\mathrm{RUL})}(t)$ for four prototype datasets at $L_{est} = 24$

Algorithm	$r = 1$	$r = 2$	$r = 3$	$r = 4$
MAE	6.65	7.59	5.23	9.26
RMSE	7.81	8.95	6.34	10.76

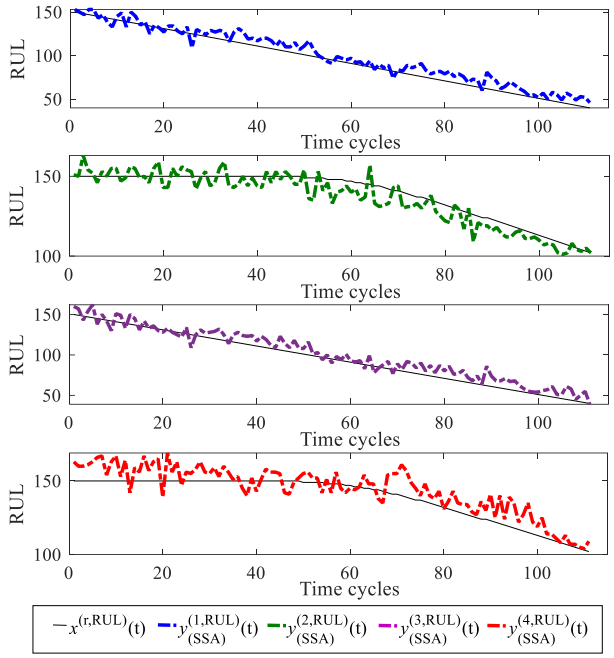

Figure. 5 Four estimated RULs $y^{(r,\mathrm{RUL})}_{(\mathrm{SSA})}(t)$ for $r = 1, \ldots, 4$ are plotted respectively in comparison with their true RUL counterparts $x^{(r,\mathrm{RUL})}(t)$ using the estimated window length $L_{est} = 24$. In all of the four plots, $y^{(r,\mathrm{RUL})}_{(\mathrm{SSA})}(t)$ are plotted as dashed-dotted lines and the true RULs $x^{(r,\mathrm{RUL})}(t)$ as thin solid lines

higher, i.e. $k = 10$ and 25, the basis functions become more oscillatory representing high frequency components buried in the prototypes. The estimated RULs derived by the LSTM, $y^{(r,\mathrm{RUL})}_{(\mathrm{SSA})}(t)$ for $L_{est} = 24$ are plotted in comparison with the true RULs $x^{(r,\mathrm{RUL})}(t)$ for four prototypes in Fig. 5 where the dash-dotted line and the solid line represent $y^{(r,\mathrm{RUL})}_{(\mathrm{SSA})}(t)$ and $x^{(r,\mathrm{RUL})}(t)$ respectively.

The mean absolute errors (MAEs) and RMSEs from four prototype datasets for SSA $y^{(r,\mathrm{RUL})}_{(\mathrm{SSA})}(t)$ are expressed in Table 1. It is shown that the estimated RUL $y^{(3,\mathrm{RUL})}_{(\mathrm{SSA})}(t)$ for $x^{(3,\mathrm{RUL})}(t)$ in Fig.5 has lowest errors for this testing with MAE = 5.23 and RMSE = 6.34.

For performance testing in RUL estimation, synthetic datasets for SSA as in Eq. (14) or (18), FFT [18] in Eq. (26), EMD [17] in Eq. (19) and FDM [19] in Eq. (20) were generated in order to compare with 200 features from the actual datasets. All synthetic datasets and the actual datasets were then regressed by the LSTM network for RUL estimation with the batch of 116 records of the actual RUL datasets. The testing of RUL estimation by using the features of the actual datasets has been conducted in [25] and the

Table 2. Averaged MAE and averaged RMSE from 200 synthetic datasets by SSA, FFT [18], EMD [17] and FDM [19] compared with the actual feature datasets [25] for RUL estimation with 116 time cycles

Algorithm	MAE	RMSE
Actual feature [25]	23.828	35.284
SSA	25.123	36.825
FFT [18]	27.126	38.472
EMD [17]	28.362	39.402
FDM [19]	30.41	41.705

performed and the averaged MAEs and the averaged RMSEs of all the synthetic datasets are given in Table. 2.

From Table 2, it is shown that averaged MAE and the averaged RMSE associated with the synthetic datasets generated by SSA are the lowest as compared to those associated by FFT [18], EMD [17] and FDM [19].

In Fig. 6 the synthetic datasets generated by the four algorithms are shown in comparison with the three select prototypes $x^{(f)}(t)$ which are plotted in order from left to right across the columns as black solid lines in all panels. It is shown in the 1^{st} (topmost) row of Fig. 6 that those generated by the SSA algorithm, the blue dashed-dotted lines, performed the best in replicating all the three prototypes. For the following three consecutive rows, the synthetic datasets generated by the EMD $y^{(f)}_{(EMD)}(t)$, by the FFT $y^{(f)}_{(FFT)}(t)$ and by the FDM $y^{(f)}_{(FDM)}(t)$ are shown respectively. Of all the four results, the synthetic datasets generated by the FDM $y^{(f)}_{(FDM)}(t)$ algorithm performed the worst in all of the comparison.

Finally, for visualization of the performance comparison, two true RULs from the actual datasets are chosen and compared with the estimated RULs with the 116 batch records of the synthetic datasets as in Fig. 7 and Fig. 8. In Fig. 7 (a)-(d), the true RUL of the actual dataset number 1 is compared with the estimated RUL with the synthetic dataset generated by SSA, FFT, EMD and FDM respectively. In Fig. 8 (a)-(d), the comparison between the four algorithms and the true RUL of the actual dataset number 2 is performed in the same manner as in Fig. 7 (a)-(d). In both Fig. 7 and Fig. 8, the estimated RULs derived by the synthetic datasets generated by SSA are the closest to the true RULs. These results as well as the averaged MAE and the averaged RMSE results in Table 2 suggest that the synthetic datasets generated by SSA as in Eq. (14) or (18) perform the best in RUL estimation as compared to those generated by FFT

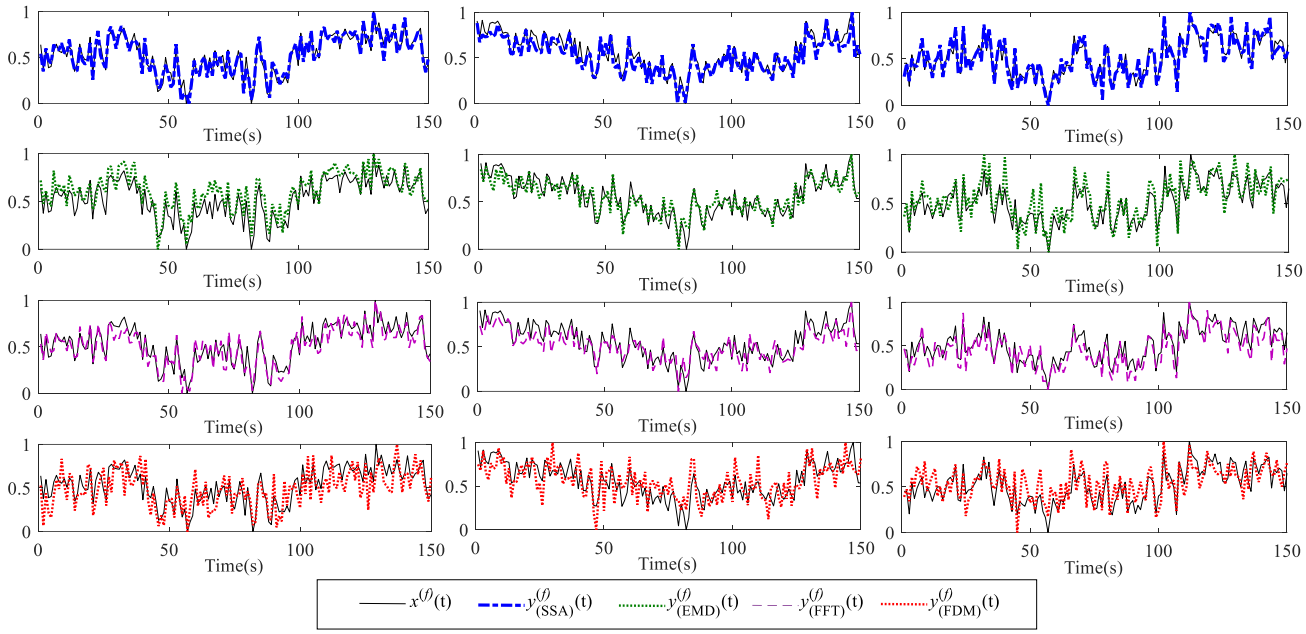

Figure. 6 The synthetic datasets generated by the SSA, EMD, FFT and EMD algorithms are shown in comparison with the three select features $x^{(f)}(t)$ which are plotted in black solid lines in all panels. The three panels in the 1^{st} (topmost) row represent the plots of the synthetic datasets generated by the SSA algorithm $y^{(f)}_{(SSA)}(t)$ for $f = 1,2,3$ respectively in blue dashed-dotted lines starting from the 1^{st} (leftmost) column. For the following three consecutive rows, the three panels of the synthetic datasets generated by the EMD $y^{(f)}_{(EMD)}(t)$ for $f = 1,2,3$ are plotted in green dotted lines, by the FFT $y^{(f)}_{(FFT)}(t)$ in purple dashed lines and by the FDM $y^{(f)}_{(FDM)}(t)$ in red dotted lines respectively

[18] in Eq. (26), EMD [17] in Eq. (19) and FDM [19] in Eq. (20).

5. Conclusions

In this paper, we propose a novel synthetic dataset generation method for turbofan engines datasets [22] based on the singular spectrum analysis (SSA) algorithm. The proposed system generates synthetic datasets by means of the basis functions derived from the SSA analysis-synthesis operations of the select prototypes. Diversity in multiple synthetic datasets can be achieved by applying the random weight functions to the basis functions. The other three algorithms, i.e., EMD [17] as described in Eq. (19), FDM [19] in Eq. (20) and FFT [18] in Eq. (26), were also considered for the performance comparison of synthetic dataset generations.

In order to test the similarity between the synthetic and the prototype datasets, the cross-correlation of all synthetic datasets from the four algorithms and the actual datasets were compared with the cross correlation of the prototype datasets and the actual datasets. It was shown that the SSA-based synthetic datasets provided the lowest average cross-correlation errors across 40 lag orders and 25 actual dataset indexes. This implied that the SSA algorithm provided synthetic datasets whose characteristics were the closest to that of the prototype. The performance of the SSA-based synthetic datasets for remaining useful life (RUL) estimation and those of the existing algorithms, i.e., FFT, EMD and FDM, were compared by means of the regression performed by the Long Short Term Memory (LSTM) neural networks. The performance metrics of the synthetic datasets generated by the proposed SSA-based and the three existing algorithms were measured in terms of the mean absolute error (MAE) and the root mean squared error (RMSE) of their RUL estimates averaged over 200 datasets. The results were compared with those of the actual feature dataset [25] which provided the MAE of 23.828 and RMSE of 35.284. For the synthetic datasets, the results showed the MAE of 27.126 and RMSE of 38.472 for the FFT [18], the MAE of 28.362 and RMSE of 39.402 for the EMD [17] and the MAE of 30.410 and RMSE of 41.705 for the FDM [19]. It was revealed that the synthetic datasets generated by the proposed SSA-based method performed the best with the MAE of 25.123 and RMSE of 36.825 confirming the applicability of the proposed SSA-based synthetic datasets in substitution of the actual datasets for RUL estimation.

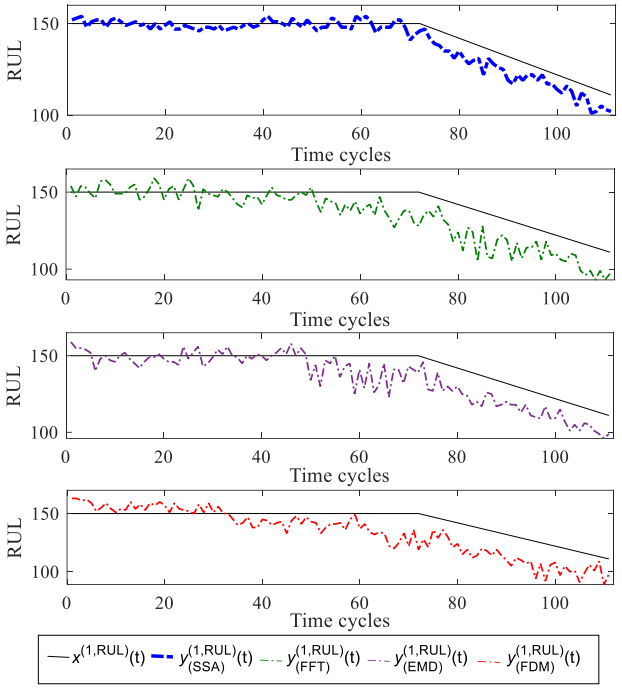

Figure. 7 The true RUL of the actual dataset number 1 $x^{(1,RUL)}(t)$ plotted in thin black lines in each panel is compared with the estimated RULs derived from the synthetic datasets generated by SSA $y^{(1,RUL)}_{(SSA)}(t)$, by FFT $y^{(1,RUL)}_{(FFT)}(t)$, by EMD $y^{(1,RUL)}_{(EMD)}(t)$ and by FDM $y^{(1,RUL)}_{(FDM)}(t)$ which are plotted in four consecutive panels

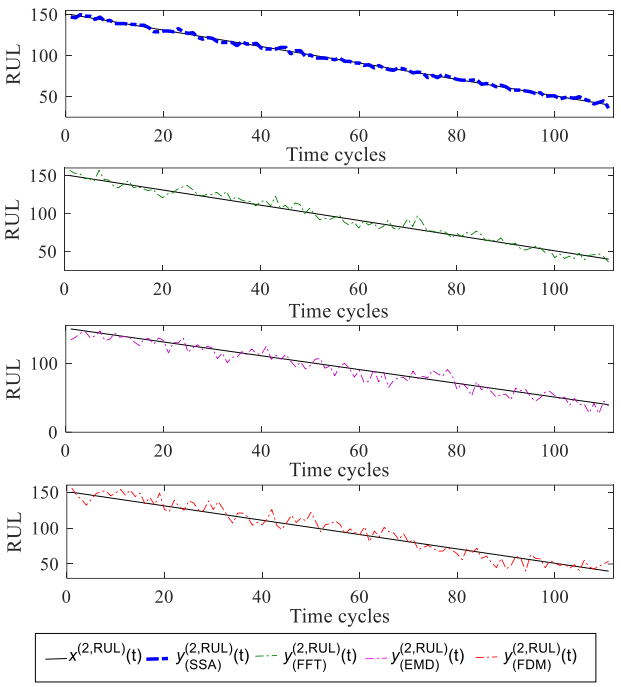

Figure. 8 The true RUL of the actual dataset number 2 $x^{(2,RUL)}(t)$ plotted in thin black lines in each panel is compared with the estimated RULs derived from the synthetic datasets generated by SSA $y^{(2,RUL)}_{(SSA)}(t)$, by FFT $y^{(2,RUL)}_{(FFT)}(t)$, by EMD $y^{(2,RUL)}_{(EMD)}(t)$ and by FDM $y^{(2,RUL)}_{(FDM)}(t)$ which are plotted in four consecutive panels

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

The paper conceptualization, methodology, writing—review, editing, validation content, supervision and project administration have been done by 1st author. The software, formal analysis and synthetic, testing, investigation, evaluation system, resources management, data curation, writing—original draft preparation and editing visualization have been done by 2nd author.

References

[1] N. Sakib and T. Wuest, "Challenges and opportunities of condition-based predictive maintenance: a review", In: *Proc. of Global Web Conf. Envisaging the future manufacturing, design, technologies and systems in innovation era*, Vol. 78, pp. 267-272, 2018.

[2] G. A. Susto, A. Schirru, S. Pampuri, S. McLoone, and A. Beghi, "Machine learning for predictive maintenance: A multiple classifier approach", *IEEE Transactions on Industrial Informatics*, Vol. 11, No. 3, pp. 812-820, 2014.

[3] N. H. Kim, D. An, and J. H. Choi, "Introduction", in *Prognostics and health management of engineering systems*, Springer International Publishing, Switzerland, ch. 1, pp. 1-24, 2017.

[4] R. Gouriveau, K. Medjaher, and N. Zerhouni, "PHM and Predictive Maintenance", in *From prognostics and health systems management to predictive maintenance 1: Monitoring and prognostics*, John Wiley and Sons, United State of America, Vol. 4, ch. 1, pp. 1-13, 2016.

[5] C. M. Brigitte, J. M. Nicod, and C. Varnier, "Traceability of Information and Knowledge Management", in *From Prognostics and Health Systems Management to Predictive Maintenance 2: Knowledge, Reliability and Decision*, John Wiley and Sons, United State of America, Vol. 7, ch. 1, pp. 1-21, 2017.

[6] X. S. Si, W. Wang, C. H. Hu, D. H. Zhou, and M. G. Pecht, "Remaining useful life estimation based on a nonlinear diffusion degradation process", *IEEE Transactions on Reliability*, Vol. 61, No. 1, pp. 50–67, 2012.

[7] Z. Zhao, B. Liang, X. Wang, and W. Lu, "Remaining useful life prediction of aircraft

engine based on degradation pattern learning", *Reliability Engineering and System Safety*, Vol. 164, pp. 74–83, 2017.

[8] Y. Liu, X. Hu, and W. Zhang, "Remaining useful life prediction based on health index similarity", *Reliability Engineering and System Safety*, Vol. 185, pp. 502–510, 2019.

[9] M. Schwabacher, "A survey of data-driven prognostics", In: *Proc. of Intelligent Prognostics, Diagnostics, and Health Management*, Arlington, Virginia, United States of America, p. 7002, 2005.

[10] X. S. Si, W. Wang, C. H. Hu and D. H, Zhou. "Remaining useful life estimation–a review on the statistical data driven approaches", *European journal of operational research*, Vol. 213, No. 1, pp. 1-14, 2011.

[11] P. Klein and R. Bergmann, "Generation of Complex Data for AI-based Predictive Maintenance Research with a Physical Factory Model", In: *Proc. of International Conf. on Informatics in Control, Automation and Robotics, Prague, Czech*, pp. 40-50, 2019.

[12] Z. Wu, W. Lin, and Y. Ji, "An integrated ensemble learning model for imbalanced fault diagnostics and prognostics", *IEEE Access*, Vol. 6, pp. 8394–8402, 2018.

[13] L. Demidova and I. Klyueva, "SVM classification: Optimization with the SMOTE algorithm for the class imbalance problem", In: *Proc. of Mediterranean Conf. on Embedded Computing*, Budva, Montenegro, pp. 1-4, 2017.

[14] G. D. Ranasinghe and A. K. Parlikad, "Generating real-valued failure data for prognostics under the conditions of limited data availability", In: *Proc. of IEEE International Conf. on Prognostics and Health Management*, San Francisco, United States of America, pp. 1–8, 2019.

[15] I. Idehen, W. Jang, and T. Overbye, "Pmu data feature considerations for realistic, synthetic data generation*",* In: *Proc. of North American Power Symposium*, Wichita, Kansas, United State of America, 2019.

[16] B. Kirsch, G. Characklis, and H. Zeff, "Evaluating the Impact of Alternative Hydro-Climate Scenarios on Transfer Agreements: Practical Improvement for Generating Synthetic Streamflows", *Journal of Water Resources Planning and Management*, Vol. 139, No. 4, pp. 396-406, 2013.

[17] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu, "The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis", In: *Proc. of the Royal Society of London. Series A: mathematical, physical and engineering sciences*, Vol. 454, No.1971, pp. 903–995, 1998.

[18] H. J. Nussbaumer, "The fast Fourier transform", *Fast Fourier Transform and Convolution Algorithms*, Springer Series in Information Sciences, Berlin, Germany, Vol. 2, pp. 80-111, 1981.

[19] A. V. Oppenheim, A. S. Willsky, and S. H. Nawab, "Fourier Series Representation of Periodic Signals", in *Signal and System*, 2nd ed., Prentice Hall Signal Processing Series, Upper Saddle River, New Jersey, ch. 3, pp. 177-283, 1998.

[20] P. Bonizzi, J. M. H. Karel, O. Meste, and R. L. M. Peeters, "Singular spectrum decomposition: A new method for time series decomposition", *Advances in Adaptive Data Analysis*, Vol. 6, No. 04, p. 1450011, 2014.

[21] R. Kazala, "Algorithm of DFT leakage minimization by using optimization methods", In: *Proc. Signal Processing: Algorithms, Architectures, Arrangements, and Applications*, Poznan, Poland, pp. 280-285, 2017.

[22] NASA, "Prognostics Center of Excellence Data Repository", in Turbofan engine degradation simulation data set, Accessed on: Apr. 20, 2021. [Online]. Available: *https://ti.arc.nasa.gov/tech/dash/groups/pcoe/prognostic-data-repository/*.

[23] N. Golyandina and A. Zhigljavsky, "Basic SSA", in *Singular Spectrum Analysis for time series*, Springer Science and Business Media, Berlin, Germany, ch. 2, pp. 11-90, 2013.

[24] H. Hassani and R. Mahmoudvand, "Applications of Singular Spectrum Analysis", in *Singular Spectrum Analysis: Using R*, Springer Business and Economics, New York, United State of America, ch. 3, pp. 87-102, 2018.

[25] P. Intachai and P. Yuvapoositanon, "A Prototype Similarity-based System for Remaining Useful Life Estimation for Future Industry by Singular Spectrum Analysis-Long Short Term Memory Neural Networks Algorithm", *Journal of Mobile Multimedia*, Vol. 16, pp. 181-202, 2020.

[26] N. Ahmed and K. R. Rao, "Optimal Diagonal Filters", in *Orthogonal transforms for digital signal processing*, Springer-Verlag, Berlin, Germany, ch. 8.5, pp. 189-191, 1975.

[27] C. Leon, "Time-frequency distributions-a review", In: *Proc. of the IEEE*, Vol. 77, No. 7, pp. 941-981, 1989.

[28] A. V. Oppenheim, R. W. Schafer, *Discrete-time signal processing*, 3rd ed., Prentice Hall Signal

Processing Series, Upper Saddle River, New Jersey, 2010.
[29] P. Singh, S. D. Joshi, R. K. Patney, and K. Saha, "The Fourier decomposition method for nonlinear and non-stationary time series analysis", In: *Proc. of the Royal Society A: Mathematical, Physical and Engineering Sciences*, Vol. 473, No. 2199, p. 20160871, 2017.
[30] G. Strang, "Wavelets", *American Scientist*, Vol. 82, No. 3, pp. 250–255, 1994.
[31] P. Venkatappareddy and B. Lall, "Characterizing empirical mode decomposition algorithm using signal processing techniques", *Circuits Systems and Signal Processing*, Vol. 37, No. 7, pp. 2969-2996, 2018.
[32] S.-J. Orfanidis, "SVD, PCA, KLT, CCA, and All That", *Optimum Signal Processing*, pp. 332-525, 2007.
[33] C. Maccone, "The KLT (Karhunen–Loève Transform) to extend SETI searches to broad-band and extremely feeble signals", *Acta Astronautica*, Vol. 67, No. 11-12, pp. 1427-1439, 2010.
[34] J. B. Elsner and A. A. Tsonis, "Foundations of SSA", in *Singular spectrum analysis: a new tool in time series analysis*, Springer Science and Business Media, Berlin, Germany, ch. 4, pp. 39-50, 2013.
[35] A. Mosallam, K. Medjaher, and N. Zerhouni, "Bayesian approach for remaining useful life prediction", *Chemical Engineering Transactions*, Vol. 33, pp. 139-144, 2013.
[36] M. A. R. Khan and D. S. Poskitt, "A note on window length selection in singular spectrum analysis", *Australian and New Zealand Journal of Statistics*, Vol. 55, No. 2, pp. 87–108, 2013.
[37] The Mathworks Inc., "Sequence-to-sequence regression using deep learning", in Mathworks Help Center, Accessed on: Apr. 20, 2021. [Online]. Available: *https://www.mathworks.com/help/deeplearning/ug/sequence-to-sequence-regression-using-deep-learning.html*.
[38] M. Giuliani, J. Herman, and J. Quinn, "Kirsch-Nowak Stream flow Generator", in *GitHub*, Accessed on: Apr. 20, 2021. [Online]. Available: *https://github.com/julianneq/Kirsch-Nowak_Streamflow_Generator*/.
[39] M. A. R. Khan and D. Poskitt, "Window length selection and signal-noise separation and reconstruction in singular spectrum analysis", in *Monash Econometrics and Business Statistics Working Papers 23/11*, Monash University, Australia, 2011.
[40] M. A. R. Khan and D. Poskitt, "Moment tests for window length selection in singular spectrum analysis of short-and long-memory processes", *Journal of Time Series Analysis*, Vol. 34, No. 2, pp. 141–155, 2013.