J. Mod. Power Syst. Clean Energy (2018) 6(2):317–329
https://doi.org/10.1007/s40565-018-0398-0

Short-term local prediction of wind speed and wind power based on singular spectrum analysis and locality-sensitive hashing

Ling LIU[1], Tianyao JI[1], Mengshi LI[1], Ziming CHEN[1], Qinghua WU[1]

Abstract With the growing penetration of wind power in power systems, more accurate prediction of wind speed and wind power is required for real-time scheduling and operation. In this paper, a novel forecast model for short-term prediction of wind speed and wind power is proposed, which is based on singular spectrum analysis (SSA) and locality-sensitive hashing (LSH). To deal with the impact of high volatility of the original time series, SSA is applied to decompose it into two components: the mean trend, which represents the mean tendency of the original time series, and the fluctuation component, which reveals the stochastic characteristics. Both components are reconstructed in a phase space to obtain mean trend segments and fluctuation component segments. After that, LSH is utilized to select similar segments of the mean trend segments, which are then employed in local forecasting, so that the accuracy and efficiency of prediction can be enhanced. Finally, support vector regression is adopted for prediction, where the training input is the synthesis of the similar mean trend segments and the corresponding fluctuation component segments. Simulation studies are conducted on wind speed and wind power time series from four databases, and the final results demonstrate that the proposed model is more accurate and stable in comparison with other models.

Keywords Wind power, Wind speed, Locality-sensitive hashing (LSH), Singular spectrum analysis (SSA), Local forecast, Support vector regression (SVR)

CrossCheck date: 16 January 2018

Received: 20 June 2017 / Accepted: 16 January 2018 / Published online: 13 March 2018

✉ Tianyao JI
tyji@scut.edu.cn

Ling LIU
201620110692@mail.scut.edu.cn

Mengshi LI
mengshili@scut.edu.cn

Ziming CHEN
c.zm03@mail.scut.edu.cn

Qinghua WU
wuqh@scut.edu.cn

[1] School of Electric Power Engineering, South China University of Technology, Guangzhou 510641, China

1 Introduction

As fossil fuels are gradually being depleted, wind energy, as a non-polluting type of renewable energy, has been rapidly developed [1]. In recent years, the proportion of wind energy keeps increasing and this trend will continue for a long time [2]. In China, wind power has been vigorously developed. By the end of 2016, the installed capacity of wind power has increased to 149 GW, with an increment of 3% over 2015 [3]. With more and more wind power feeding into power systems, the randomness and intermittent nature of wind speed and wind power jeopardizes the stability and reliability of power system operation and raises the operating cost [4]. Therefore, in order to alleviate the adverse effects of wind power integration, more accurate and stable forecasting is critical and urgently needed [5].

Many researchers have devoted themselves to improving the accuracy of wind speed and wind power forecasting, and have put forward a number of forecast models. The mainstream models can be roughly classified into two categories: physical models and statistical models [6]. A

physical model usually establishes a rigorous mathematical forecast model based on the principles of geophysical fluid dynamics and thermodynamics [7]. The numerical weather prediction (NWP) model, which is a typical physical model [8], simulates the physics of the atmosphere based on certain initial values and boundary conditions. However, the NWP model requires huge computational resources, and thus, it is operated on super computers, which limits its application in short-term forecasting [9]. Statistical models use historical data only, and outperform physical models in short-term forecasting [10]. Various models for wind speed and wind power forecasting are mushrooming in recent years, among which auto-regressive and moving average (ARMA) [11], artificial neural network (ANN) [12] and support vector regression (SVR) [13] are most widely used. The main advantages are that they compute the forecast results quickly, and can work on personal computers. In electricity markets, a great number of players need the real-time results of wind speed and wind power forecasting, such as system operators, wind power generators, and suppliers. In order to help them predict wind speed and wind power more accurately with less time, this paper proposes a short-term forecast model that is a statistical model.

In general, it is difficult to forecast from the original time series directly, as wind speed and wind power have the nature of rapid fluctuation and high randomness. Thus, it has been proposed to filter the original time series or to decompose it into multiple series by empirical mode decomposition (EMD) [14], wavelet transform (WT) [15], k-opening-closing and closing-opening (k-OCCO) filter [9], etc. EMD and WT decompose the original time series into several components and forecast each component respectively [14, 15]. However, such a strategy consumes more time, and the error arising in each component will be integrated, so that it leads to an unsatisfactory forecast result [16, 17]. The k-OCCO filter extracts the tendency of the original time series and treats the remainder as noise [18], which may induce a large fixed error.

Thus, in order to overcome the shortcomings mentioned above, this paper proposes a novel forecasting model which can avoid both error accumulation and fixed errors. The model is based on singular spectrum analysis (SSA) [19, 20] and locality-sensitive hashing (LSH) [21–23] for short-term forecasting. SSA is used for decomposing the original data into two components: the mean trend and the fluctuation component. According to previous research, local prediction performs better than global prediction [24]. Therefore, LSH is used to classify the sample segments based on hash functions and search for the similar segments of the forecast mean trend segment, which refers to the segment of the mean trend at the period of time when forecasting is to be done. To the best of our knowledge, it is the first time that LSH is used in wind speed/power forecasting to select similar segments, so that the size of the training data set can be reduced and the forecast result can be more accurate. Instead of forecasting the two components individually or the mean trend solely, this paper proposes to synthesize the similar mean trend segments and the corresponding fluctuation component segments into the training input of SVR for more accurate forecast results.

2 Methodologies

2.1 Singular spectrum analysis

SSA is a powerful method to study the meaningful features of nonlinear time series, which has gained more attention in recent years. It extracts and identifies the mean trend and the fluctuation component from an original time series, expecting to obtain a superior forecast result with a synthesis of these two components. Standard SSA is performed in the following steps.

Step 1: Embedding. For an original time series $\boldsymbol{y} = [y_1\ y_2\ \cdots\ y_N]$ of length N, L is the selected embedding dimension or the window length in the SSA. The original time series $\boldsymbol{y}$ is converted to L-lagged vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_K$, $K = N - L + 1$, and each L-lagged vector $\boldsymbol{x}_i$ is defined as

$$\boldsymbol{x}_i = [y_i\ y_{i+1}\ \cdots\ y_{i+L-1}]^{\mathrm{T}} \quad i = 1, 2, \cdots, K \tag{1}$$

The trajectory matrix $\boldsymbol{X} \in \mathbb{R}^{L\times K}$ is defined as

$$\boldsymbol{X} = [x_1\ x_2 \ldots x_K] \begin{bmatrix} y_1 & y_2 & \cdots & y_k \\ y_2 & y_3 & \cdots & y_{K+1} \\ \vdots & \vdots & \cdots & \vdots \\ y_L & y_{L+1} & \cdots & y_N \end{bmatrix} \tag{2}$$

Step 2: Singular value decomposition (SVD). The d eigenvalues of the covariance matrix $\boldsymbol{S} = \boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$ are calculated, which meet the requirement $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$. Correspondingly, $\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_d$ are the orthogonal eigenvectors. Then the trajectory matrix $\boldsymbol{X}$ can be decomposed by SVD into

$$\boldsymbol{X} = \boldsymbol{e}_1 + \boldsymbol{e}_2 + \cdots + \boldsymbol{e}_d \tag{3}$$

$$\boldsymbol{e}_i = \sqrt{\lambda_i}\boldsymbol{u}_i\boldsymbol{v}_i^{\mathrm{T}} \tag{4}$$

$$\boldsymbol{v}_i = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{u}_i\sqrt{\lambda_i} \tag{5}$$

$$d = \min\{L, K\} \quad i = 1, 2, \cdots, d \tag{6}$$

where the vectors $\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_d$ are considered to be principle components of trajectory matrix $\boldsymbol{X}$. Consequently, the collection $(\sqrt{\lambda_i}, \boldsymbol{u}_i, \boldsymbol{v}_i)$ is defined as the ith triple feature vector of matrix $\boldsymbol{X}$.

Step 3: Grouping. After decomposition, the indices $\{1,2,\cdots,d\}$ are divided into m independent groups $I_1, I_2, \cdots, I_m$ and $I_j = \{i_{j_1}, \cdots, i_{j_p}\}$, $j = 1,2,\cdots,m$ are denoted as a group of indices. The trajectory matrix $\boldsymbol{X}$ is expressed as

$$\boldsymbol{X} = \boldsymbol{E}_{I_1} + \boldsymbol{E}_{I_2} + \cdots + \boldsymbol{E}_{I_m} \tag{7}$$

$$\boldsymbol{E}_{I_j} = \boldsymbol{e}_{i_{j_1}} + \boldsymbol{e}_{i_{j_2}} + \cdots + \boldsymbol{e}_{i_{j_p}} \tag{8}$$

where the contribution rate of $\boldsymbol{E}_{I_j}$ is $\sum_{i\in I_j}\lambda_i / \sum_{i=1}^{d}\lambda_i$. In this paper, the original time series is decomposed into two components. Hence, we have $\boldsymbol{X} = \boldsymbol{E}_{I_1} + \boldsymbol{E}_{I_2}$.

Step 4: Diagonal averaging. Each matrix $\boldsymbol{E}_{I_j} \in \mathbb{R}^{L\times K}$ is converted to a time series by the following method. Let z_{il}, $i = 1,2,\cdots,L$, $l = 1,2,\cdots,K$ be the elements of the matrix with $\boldsymbol{E}_{I_j}$. Set $L^* = \min\{L,K\}$, $K^* = \max\{L,K\}$, and $N = K + L - 1$. $z^*_{il} = z_{il}$ when $L<K$; otherwise, $z^*_{il} = z_{li}$. Diagonal averaging converts matrix $\boldsymbol{E}_{I_j}$ into a time series $\left[y_1^{(j)}\ y_2^{(j)}\ \cdots\ y_N^{(j)}\right]$ via the following equation

$$y_k^i = \begin{cases} \dfrac{1}{k}\displaystyle\sum_{q=1}^{k+1} z^*_{q,k-q+2} & 1\le k\le L^* \\ \dfrac{1}{L^*}\displaystyle\sum_{q=1}^{L^*} z^*_{q,k-q+2} & L^*\le k\le K^* \\ \dfrac{1}{N-K+2}\displaystyle\sum_{q=k-K^*+2}^{N-K^*+1} z^*_{q,k-q+2} & K^*\le k\le N^* \end{cases} \tag{9}$$

In this paper, the diagonal averaging is applied to $\boldsymbol{E}_{I_1}$ and $\boldsymbol{E}_{I_2}$, respectively, and the two corresponding time series are $\boldsymbol{y}^{(1)} = \left[y_1^{(1)}\ y_2^{(1)}\ \cdots\ y_N^{(1)}\right]$ and $\boldsymbol{y}^{(2)} = \left[y_1^{(2)}\ y_2^{(2)}\ \cdots\ y_N^{(2)}\right]$ according to (9). Thus, the initial time series $\boldsymbol{y}$ is expressed by

$$\boldsymbol{y} = \boldsymbol{y}^{(1)} + \boldsymbol{y}^{(2)} \tag{10}$$

where $\boldsymbol{y}^{(1)}$ is the mean trend and $\boldsymbol{y}^{(2)}$ is the fluctuation component.

An example is given in Fig. 1, where a time series collected from the wind power database of Elia, the Belgian transmission system operator (TSO), is decomposed into the mean trend and the fluctuation component.

2.2 Locality sensitive hashing

Considering the continuity characteristics of meteorological data, wind speed and wind power at a certain time instant are strongly correlated with data collected in a short period of time before. Thus, instead of putting all the sample segments into training the model, it is more effective to use only the segments that follow the same tendency as the period of time when forecasting is to be done, called the forecast segment. In this paper, the LSH, which is a common and fast similarity search technique, is applied to select the similar segments. It proceeds in two steps: index building and similarity search.

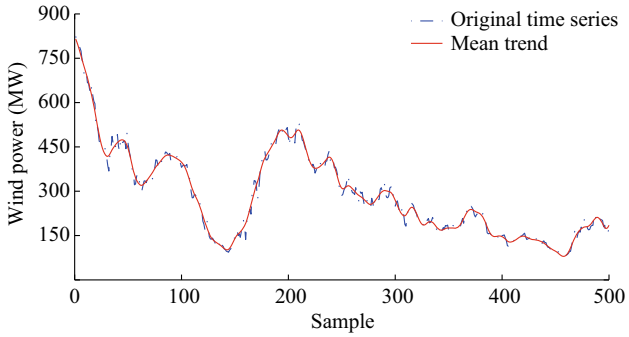

Fig. 1 A time series of wind speed from the Elia database and the mean trend extracted by SSA

Step 1: Index building. The mean trend $\boldsymbol{y}^{(1)}$ is reconstructed into a higher-dimensional phase space with the embedding dimension s and time constant τ. It can be reconstructed into $N-(s-1)\tau$ segments and each segment is expressed as

$$\boldsymbol{y}_i^{(1)} = \left[y_i^{(1)}\ y_{i+\tau}^{(1)}\ \cdots\ y_{i+(s-1)\tau}^{(1)}\right] \tag{11}$$

where $i = 1,2,\cdots,N-(s-1)\tau$.

Afterwards, segments $\boldsymbol{y}_i^{(1)}$ are passed through the locality-sensitive hash function (LSHF) of $h(\boldsymbol{y}_i^{(1)})$, and similar segments are hashed into a certain bucket with high probability, while dissimilar segments have low probability to be hashed into this bucket.

A family H is called locality sensitive, or more specifically (r_1, r_2, p_1, p_2)-sensitive for D, if for any $\boldsymbol{y}_i^{(1)}$ and $\boldsymbol{y}_j^{(1)}$, it holds

$$\begin{cases} \Pr_H\left\{h(\boldsymbol{y}_i^{(1)}) = h(\boldsymbol{y}_j^{(1)})\right\} \ge p_1 & D(\boldsymbol{y}_i^{(1)}, \boldsymbol{y}_j^{(1)}) \le r_1 \\ \Pr_H\left\{h(\boldsymbol{y}_i^{(1)}) = h(\boldsymbol{y}_j^{(1)})\right\} \le p_2 & D(\boldsymbol{y}_i^{(1)}, \boldsymbol{y}_j^{(1)}) \ge r_2 \end{cases}$$

where D is the distance measure; $\Pr_H$ is the probability with respect to a random choice of $h \in H$, and in order to ensure the family is practical, the conditions $p_1 > p_2$ and $r_1 < r_2$ should be guaranteed.

In this paper, the distance measure D based on normal Euclidean distance is selected. Correspondingly, the LSHF $h(\boldsymbol{y}_i^{(1)})$ is given by

$$h_{\boldsymbol{a},b}(\boldsymbol{y}_i^{(1)}) = \left\lfloor \frac{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{y}_i^{(1)} + b}{r} \right\rfloor \tag{12}$$

where $\boldsymbol{a}$ is a d-dimensional vector which obeys the normal distribution; r is the width of the interval of a random

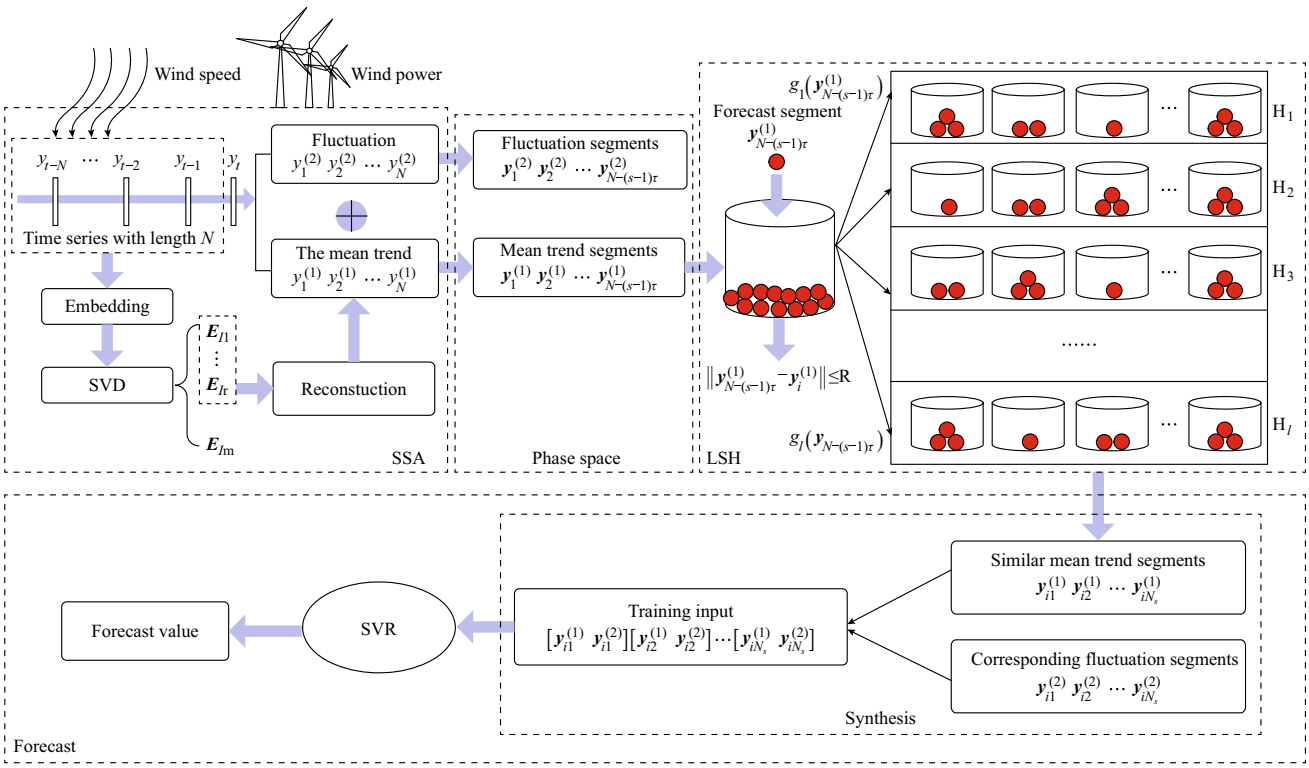

Fig. 2 Framework of proposed forecast model

Table 1 Contribution rates of eigenvectors

Order	Contribution rate (%)
1	98.438
2	1.153
3	0.194
4	0.074
5-20	0.141

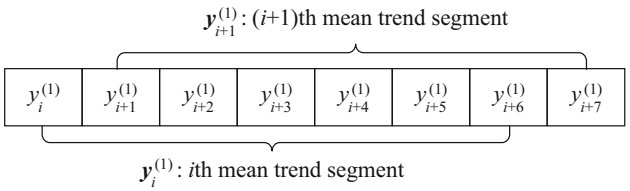

Fig. 3 Reconstruction of the mean trend

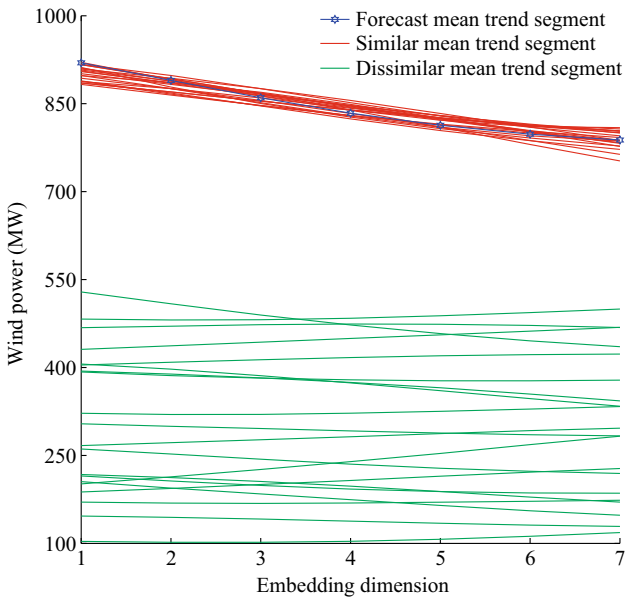

Fig. 4 Searching for similar segments to forecast mean trend segment

bucket; b is a real number selected randomly from the range of [0, r].

Similarly, a hash table function $g(\boldsymbol{y}_i^{(1)})$ is a concatenation of k LSHFs, $h_j(\boldsymbol{y}_i^{(1)}) \in H$, $j = 1, 2, \cdots, k$:

$$g(\boldsymbol{y}_i^{(1)}) = \left[h_1(\boldsymbol{y}_i^{(1)})\ h_2(\boldsymbol{y}_i^{(1)})\ \cdots\ h_k(\boldsymbol{y}_i^{(1)})\right]^{\mathrm{T}} \tag{13}$$

$$g_{\boldsymbol{A},b}(\boldsymbol{y}_i^{(1)}) = \left\lfloor \frac{\boldsymbol{A}^{\mathrm{T}}\boldsymbol{y}_i^{(1)} + \boldsymbol{b}}{r} \right\rfloor \tag{14}$$

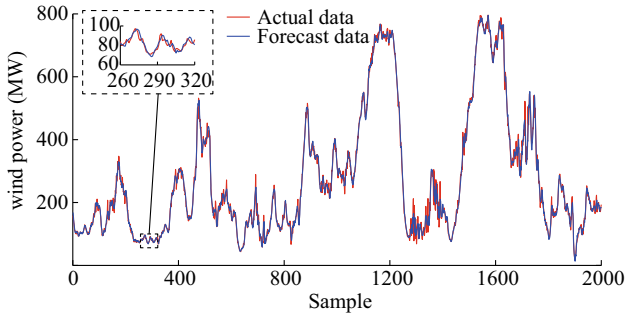

Fig. 5 Results of 4-step ahead forecast of Elia data set

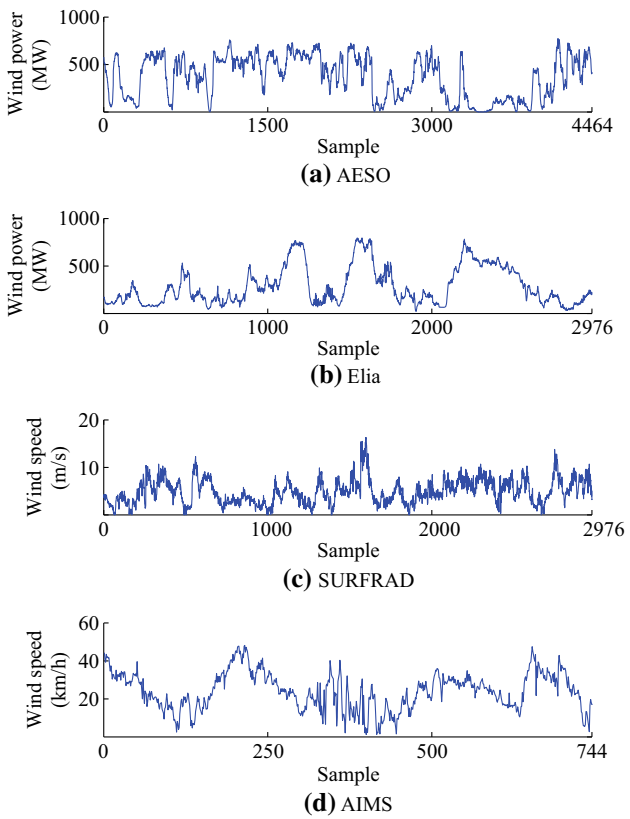

Fig. 6 Data from four data sets in March

where the projection matrix $\boldsymbol{A} = [\boldsymbol{a}_1\ \boldsymbol{a}_2\ \cdots\ \boldsymbol{a}_k] \in \mathbb{R}^{d\times k}$ and each column $\boldsymbol{a}_i$ follows a normal distribution; $\boldsymbol{b}$ is a k-dimensional vector with each element chosen uniformly from the range of $[0, r]$. $g_{\boldsymbol{A},\boldsymbol{b}}(\boldsymbol{y}_i^{(1)})$ transforms the d-dimensional segment $\boldsymbol{y}_i^{(1)}$ into buckets, such that segments in the same bucket are similar. A bucket is chosen at random and partitioned into uniform widths of r. To increase the accuracy of index building, l hash table functions $g_1(\boldsymbol{x}), g_2(\boldsymbol{x}), \cdots, g_l(\boldsymbol{x})$ are applied, thus forming l tables $\mathrm{H}_1 \sim \mathrm{H}_l$.

Step 2: Similarity search. The forecast mean trend segment $\boldsymbol{y}_{N-(s-1)\tau}^{(1)}$ is projected to l buckets by the l hash table functions. In each table, the segments hashed into the same bucket as the forecast segment are returned, and the N_s more similar ones, which have a smaller distance from the forecast segment, are selected to be used in local forecasting. Instead of computing the distance of all segments, LSH only calculates the distance of segments in the same bucket, which saves time remarkably.

3 Proposed forecast model

The framework of the proposed forecast model is shown in Fig. 2. In the first step, the decomposition of the original time series by SSA yields the mean trend and the fluctuation component, as shown in Fig. 1. The contribution rates of each eigenvector are illustrated in Table 1 and the first three eigenvectors are used for extracting the mean trend. Afterwards, as shown in Fig. 3, the mean trend will be reconstructed in the phase space and converted to the mean trend segments and fluctuation component segments, respectively. Due to the high randomness of the fluctuation component, similar segment searching is carried out on the mean trend segments, and a number of similar segments of the forecast mean trend segment are obtained by LSH. Figure 4 demonstrates an example of the result of similar segment searching. Finally, in order to avoid error accumulation and fixed errors, the training input is the synthesis of the similar mean trend segments and the corresponding fluctuation component segments. An example of the forecast result is given in Fig. 5, which intuitively is satisfactory.

For detailed instructions on how to implement the proposed model for forecasting, the pseudo code is listed in Algorithm 1.

Algorithm 1 The proposed model

Initialization:

Build the proposed model according to Section 2.1 and Section 2.2. Set the values of $L = 20$, $l = 10$, $k = 25$, $s = 7$, $\tau = 1$ and $N_\mathrm{s} = 500$.

Body:

1: Load original time series $\boldsymbol{y} = [y_1\ y_2\ \cdots\ y_N]$, where N represents the number of samples for each season of each database. Determine the look-ahead steps p, embedding dimension s and time constant τ.

2: Apply SSA to decompose the original time series $\boldsymbol{y}$ into the mean trend $\boldsymbol{y}^{(1)} = \left[y_1^{(1)}\ y_2^{(1)}\ \cdots\ y_N^{(1)}\right]$ and the fluctuation component $\boldsymbol{y}^{(2)} = \left[y_1^{(2)}\ y_2^{(2)}\ \cdots\ y_N^{(2)}\right]$.

3: Reconstruct the mean trend $\boldsymbol{y}^{(1)}$ and the fluctuation component $\boldsymbol{y}^{(2)}$ into mean trend segments $\boldsymbol{y}_i^{(1)}$ and corresponding fluctuation segments $\boldsymbol{y}_i^{(2)}, i = 1, 2, \cdots, N-(s-1)\tau$ in the phase space.
4: From the $(N-(s-1)\tau)$ mean trend segments, obtain N_s segments similar to the forecast mean trend segment $\boldsymbol{y}_{N-(s-1)\tau}^{(1)}$ by using LSH.
5: Synthesize the N_s mean trend segments $\boldsymbol{y}_i^{(1)}$ and fluctuation segments $\boldsymbol{y}_i^{(2)}$, to obtain the training input $\left[\boldsymbol{y}_i^{(1)}\ \boldsymbol{y}_i^{(2)}\right]$. Correspondingly, the training output is the actual value $y_{i+(s-1)\tau+p}$. Note that subscript i corresponds to N_s subscripts out of $1 \sim N-(s-1)\tau-1$.
6: Build an SVR model of the training input and training output.
7: Perform the SVR model with the forecast segment $\left[\boldsymbol{y}_{N-(s-1)\tau}^{(1)}\ \boldsymbol{y}_{N-(s-1)\tau}^{(2)}\right]$ as the input and obtain the forecast value $\hat{y}_{N+p}$.
8: **return** Forecast value $\hat{y}_{N+p}$

4 Simulation studies

4.1 Data sources

To evaluate the performance of the proposed model, wind speed and wind power data from four databases are used for forecasting. The data are collected from different wind farms with abundant wind resource. The first database provides inshore wind power data collected from AESO, located in Alberta, Canada [9]. The second database includes the total wind power generation from Elia. The third database is the wind speed data collected from a surface radiation network (SURFRAD) station [25]. The fourth database is offshore wind speed data collected from a testing station in Darwin supported by the Australian Institute of Marine Science (AIMS). The data from the four databases in March are shown in Fig. 6, and their statistical measures (mean, standard deviation (Std), minimum and maximum) are given in Table 2. To test the performance of the forecast models in different seasons, the prediction is carried out in March, June, September and December and compared with the actual data.

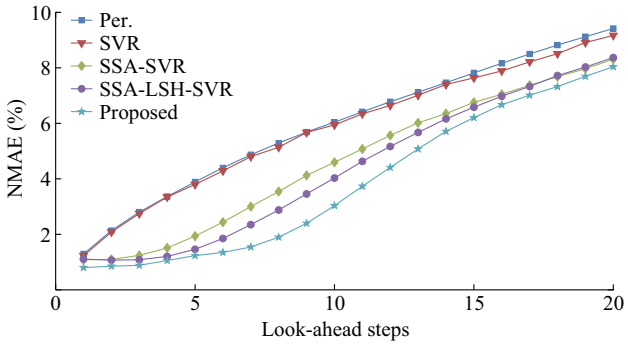

Fig. 7 NMAEs of five models on Elia data set in June

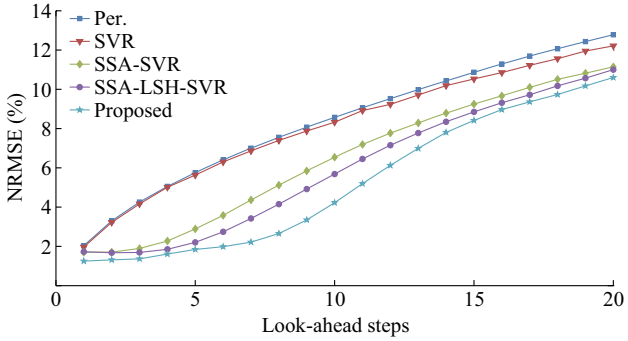

Fig. 8 NRMSEs of five models on Elia data set in June

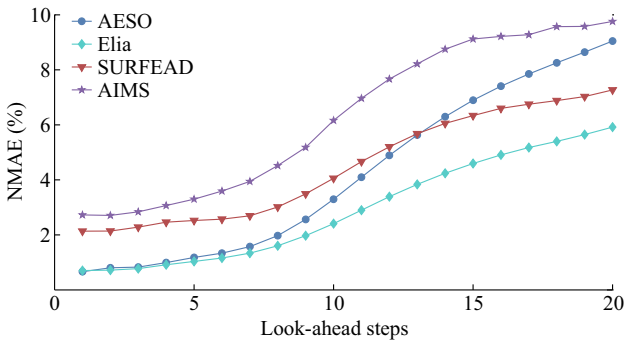

Fig. 9 NMAEs of proposed model on four data sets

Table 2 Statistical properties of the data of each data set

Database	Sample time	Sample interval	Statistical properties				Number of samples			
			Mean	Std	Max	Min	3	6	9	12
AESO(MW)	10-minute	01/01/2013-31/12/2013	293.0409	231.7803	904.1000	0	8495	21743	39311	48095
Elia (MW)	15-minute	01/01/2013-31/12/2013	350.1230	302.4225	1343.2	1.1400	5563	14495	23327	32063
SURFRAD (m/s)	15-minute	13/06/2014-09/06/2015	4.4439	2.9775	20.5000	0	25056	33888	7680	16416
AIMS (km/h)	1-hour	08/01/2014-07/01/2015	24.1593	10.7262	67.6440	0.8424	1416	3624	5830	8016

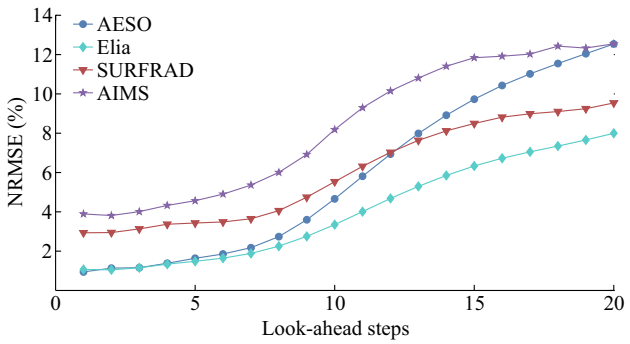

Fig. 10 NRMSEs of proposed model on four data sets

4.2 Performance qualification

To evaluate the forecasting accuracy of the proposed model, two commonly used criteria are employed to measure the errors: the normalized mean absolute error (NMAE) and normalized root mean squared error (NRMSE) [26]. NMAE represents how close are the forecast results to the final outcomes and the NRMSE assesses the standard deviation between the forecast value and actual value, which reflects the stability of the forecast model. The smaller the NMAE or NRMSE value is, the more accurate and stable the forecast results are. They are defined by

$$NMAE = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - \hat{y}_i|}{Y} \times 100\% \tag{15}$$

$$NRMSE = \frac{1}{Y}\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \times 100\% \tag{16}$$

where N is the size of the validation data set; $\hat{y}_i$ is the forecast value; y_i is the actual value. For wind power forecasting, Y is the installed capacity of the wind farm at the time of forecasting. For wind speed forecasting, Y is the maximum historical value of wind speed.

Table 3 Performance evaluation of five forecast models on AESO data set

Forecast model		Month	Performance evaluation (%)				
			Per.	SVR	SSA-SVR	SSA-LSH-SVR	Proposed
4-step ahead	NMAE	3	3.7487	3.5486	1.7593	1.1791	**0.9931**
		6	3.6017	3.4527	1.6832	1.1080	**0.9581**
		9	3.3865	3.2169	1.5420	0.9653	**0.8255**
		12	3.2587	3.1643	1.5171	0.9382	**0.7952**
	NRMSE	3	5.5308	5.1343	2.5264	1.6747	**1.3803**
		6	5.1487	4.8768	2.3751	1.5723	**1.3418**
		9	5.3357	4.9438	2.3681	1.4897	**1.2637**
		12	4.8820	4.5850	2.2177	1.4351	**1.1946**
12-step ahead	NMAE	3	7.8772	7.6632	6.6033	5.5858	**4.8926**
		6	7.5769	7.3882	6.2499	5.3132	**4.4633**
		9	7.4317	7.2392	6.1986	5.1497	**4.2807**
		12	6.8628	6.8602	6.0333	4.9038	**4.1070**
	NRMSE	3	11.3931	10.7018	9.3641	8.0207	**6.9354**
		6	10.4301	9.9668	8.5134	7.3046	**6.1824**
		9	11.3043	10.6386	9.1800	7.6844	**6.4945**
		12	10.1215	9.5907	8.2631	7.0079	**5.8755**
20-step ahead	NMAE	3	11.0825	10.8482	10.0106	9.3361	**9.0498**
		6	10.4255	10.1824	9.5325	8.8819	**8.5205**
		9	10.6095	10.3015	9.5018	8.7968	**8.4593**
		12	9.8205	10.2096	9.4999	8.3658	**7.9643**
	NRMSE	3	15.6906	14.8015	13.5531	12.8930	**12.5400**
		6	14.0942	13.2676	12.4526	11.7541	**11.4103**
		9	15.4803	14.5035	13.4658	12.5580	**12.2770**
		12	14.2865	13.6007	12.6949	11.5683	**11.0371**

Note: Bold values indicate the result corresponding to the highest accuracy in each case

4.3 Models employed for performance comparison

To validate the accuracy of the proposed model, simulation studies are conducted with the persistence (Per.), SVR, SSA-SVR and SSA-LSH-SVR models. The Per. model indicates that future wind data are consistent with the last measured value. This model performs well in short-term forecasting of wind speed and wind power owing to the slow scale of meteorological changes. Therefore, the Per. model is always utilized as a benchmark for reference.

As a typical machine learning model, the SVR model generally performs well in a majority of forecasting applications, and is also regarded as a benchmark. The SSA-SVR model extracts the mean trend of the wind speed or wind power and treats the remainder as noise. Afterwards, the mean trend of the wind speed or wind power is input into the SVR model. It is expected that the application of SSA can reduce the effect of randomness. As for the SSA-LSH-SVR model, the function of SSA in this model is the same as in the SSA-SVR model, and LSH is employed to find similar segments to the forecast mean trend segment, which are then used to form the training input to build an SVR model. In this manner, a local forecast is achieved, and it is expected to show better performance than a global forecast. By comparing the SVR, SSA-SVR and SSA-LSH-SVR models with the Per. model, the contribution of the SVR, SSA and LSH algorithms can be highlighted.

Unlike the models mentioned above, the proposed model regards the remainder, which is produced by SSA, as the fluctuation component, and LSH is also used to find similar segments to the forecast mean trend segment. The training input for SSA-SVR and SSA-LSH-SVR models are the mean trend segments only, but for the proposed model, it is the synthesis of similar mean trend segments and the corresponding fluctuation component segments. In this way, the two components are treated independently in the SVR model, and the problems of error accumulation and fixed error can be solved.

Table 4 Performance evaluation of the five forecast models on the Elia data set

Forecast model		Month	Performance evaluation (%)				
			Per.	SVR	SSA-SVR	SSA-LSH-SVR	Proposed
4-step ahead	NMAE	3	2.6018	2.5433	1.3322	1.0097	**0.9142**
		6	3.3608	3.2890	1.6240	1.2039	**1.0555**
		9	2.2986	2.2146	1.0423	0.7262	**0.6311**
		12	3.3782	3.3599	1.6283	1.0902	**0.9655**
	NRMSE	3	3.6154	3.5836	1.8750	1.4720	**1.3353**
		6	5.0544	4.9797	2.4484	1.8519	**1.6080**
		9	3.3571	3.2318	1.5565	1.1442	**1.0005**
		12	4.7742	4.6830	2.3031	1.6681	**1.5197**
12-step ahead	NMAE	3	5.1307	5.1524	4.2830	3.8509	**3.3905**
		6	6.7770	6.6244	5.6338	5.1300	**4.4332**
		9	4.8966	4.7201	3.9833	3.4160	**2.9205**
		12	7.3559	7.2613	6.2025	5.1939	**4.4494**
	NRMSE	3	6.9956	6.9042	5.8747	5.2901	**4.6795**
		6	9.5244	9.2134	7.8885	7.1261	**6.1448**
		9	6.8697	6.5273	5.5086	4.7600	**4.0875**
		12	10.2746	9.6897	8.2331	6.9645	**6.0193**
20-step ahead	NMAE	3	7.0272	6.9700	6.2935	6.1646	**5.9190**
		6	9.4153	9.1108	8.3938	8.3639	**8.0316**
		9	6.8036	6.7274	6.1369	5.7608	**5.5788**
		12	10.6998	10.8987	9.7743	9.0728	**8.7172**
	NRMSE	3	9.5660	9.1924	8.3764	8.3184	**7.9948**
		6	12.7807	12.2040	11.2540	10.9973	**10.5984**
		9	9.3852	8.9187	8.1482	7.7690	**5.5788**
		12	14.7701	14.3054	12.8965	12.0753	**11.5808**

Note: Bold values indicate the result corresponding to the highest accuracy in each case

4.4 Forecast results and discussion

Simulation studies were conducted by the Per., SVR, SSA-SVR, SSA-LSH-SVR models and the proposed model on the four data sets, respectively. Take the time series of Sect. 2, which is wind power data, as an example. The forecast results of 4 look-ahead steps in March based on the proposed model are shown in Fig. 5, which shows the forecast curve almost coincides with the actual one.

In order to assess the performance of the proposed model when forecasting for a range of look-ahead steps, simulation studies of $1\sim20$ look-ahead steps are conducted and the forecasting performance is evaluated using NMAE and NRMSE. The results are shown in Figs. 7 and 8. As shown in the two figures, the NMAE and NRMSE of the proposed model are the smallest among the five models in all cases. Furthermore, NMAE and NRMSE are improved by 69.32% and 69.02% at 6 look-ahead steps, which is the biggest improvement, and by 14.62% and 17.06% at 20 look-ahead steps, which is the least, compared to the Per. model. As for the SSA-LSH-SVR model, it is more accurate than the SVR model and the SSA-SVR model, which demonstrates the advantage of the similar segment searching strategy of LSH. Last but not the least, the proposed model, which also adopts LSH for similar segment searching but is based on the synthesis of the similar mean trend segments and the corresponding fluctuation component segments, outperforms the other four models.

To examine the performance of the proposed model on various data sets, the NMAEs and NRMSEs of the proposed model on the four data sets for $1\sim20$ look-ahead steps are plotted in Figs. 9 and 10. For the purpose of making a more comprehensive study, numerical results for 4, 12, and 20 look-ahead steps have been summarised in Tables 3, 4 5 and 6. As wind may have various natural characteristics in different seasons, the forecast is carried out in March, June, September and December. It can be seen that the proposed model gives the smallest NMAEs and NRMSEs for all of the four data sets, which indicates

Table 5 Performance evaluation of the five forecast models on the SURFRAD data set

Forecast model		Month	Performance evaluation (%)				
			Per.	SVR	SSA-SVR	SSA-LSH-SVR	Proposed
4-step ahead	NMAE	3	5.0653	4.8633	2.8113	2.5653	**2.4615**
		6	5.1903	4.7650	3.0632	2.8423	**2.7659**
		9	4.8368	4.4912	2.8301	2.6880	**2.5553**
		12	4.3706	4.0705	2.4597	2.2831	**2.1889**
	NRMSE	3	6.9570	6.6548	3.8784	3.5382	**3.3683**
		6	7.8259	7.1308	4.4766	4.1373	**4.0048**
		9	6.5681	6.0300	3.8657	3.6758	**3.5208**
		12	5.6982	5.3097	3.2413	3.0153	**2.8879**
12-step ahead	NMAE	3	7.5099	7.0259	6.1475	5.7557	**5.2010**
		6	6.9620	6.2809	5.7662	5.6304	**5.1279**
		9	6.4090	5.7873	5.2481	5.0724	**4.8155**
		12	5.8834	5.5116	4.8772	4.7785	**4.3367**
	NRMSE	3	10.0015	9.2935	8.2986	7.7483	**7.0265**
		6	9.6425	8.6669	8.0837	7.9991	**7.3079**
		9	8.5900	7.8104	7.0485	6.7051	**6.3830**
		12	7.5972	7.1519	6.3236	6.1938	**5.6389**
20-step ahead	NMAE	3	9.1928	8.3717	7.7597	7.4777	**7.2695**
		6	8.4090	7.3594	6.7626	6.8632	**6.6777**
		9	7.4027	6.7165	6.1970	6.1240	**6.0945**
		12	7.1019	6.4800	6.1095	6.1063	**5.8659**
	NRMSE	3	12.0228	10.9011	10.1763	9.7511	**9.5373**
		6	11.6459	10.1598	9.4526	9.2863	**8.9680**
		9	10.0471	9.0650	8.4565	8.1024	**8.0407**
		12	8.9335	8.2040	7.7717	7.7847	**7.5363**

Note: Bold values indicate the result corresponding to the highest accuracy in each case

that the proposed model outperforms the other ones in both accuracy and stability.

4.5 Comparison studies

In order to test the effect of LSH, simulation studies are conducted using the SSA-K-means-SVR, SSA-SOM-SVR and SSA-LSH-SVR models on the Elia data set, respectively, where the K-means and SOM (Self Organizing Maps) algorithms are employed to select the similar segments. The forecast results for 4, 12, and 20 look-ahead steps have been demonstrated in Table 7, and the SSA-LSH-SVR is superior to SSA-K-means-SVR in all cases.

Although when the look-ahead step becomes larger, LSH is not as accurate as SOM, it has great advantage in terms of computation load. All simulation studies are run on a Windows 7 PC with an Inter Core i7 CPU 3.40 GHz and 8GB memory, and the program is coded in MATLAB 2014a. Take the case of forecasting 4 look-ahead steps using the December data of the Elia data set for example. The computation time of SSA-K-means-SVR, SSA-SOM-SVR and SSA-LSH-SVR is 6.7798, 57.7116 and 1.0396 seconds, respectively. Therefore, LSH is a better choice as it sacrifices little accuracy (the largest difference is 0.46% for NMAE) with huge savings in time.

4.6 Discussion on parameters

The purpose of this paper is to make LSH effective and applicable in forecasting, and the proposed model further improves over the SSA-LSH-SVR model by replacing the training input by the synthesis of similar mean trend segments and the corresponding fluctuation component segments. However, it is necessary to analyze how to obtain the best parameters for the model.

A series of experiments are conducted on the Elia data set whose sample interval is 15 minutes, and the results show that as the number of look-ahead steps increases, a

Table 6 Performance evaluation of the five forecast models on the AIMS data set

Forecast model		Month	Performance evaluation (%)				
			Per.	SVR	SSA-SVR	SSA-LSH-SVR	Proposed
4-step ahead	NMAE	3	6.7554	6.7196	3.5731	3.4415	**3.1653**
		6	8.1777	7.6566	4.3599	4.2099	**3.9675**
		9	8.2750	7.5252	4.0651	3.8741	**3.6130**
		12	6.8461	6.5338	3.3525	3.1752	**2.7968**
	NRMSE	3	9.5042	9.3127	5.0002	4.8552	**4.4445**
		6	10.8030	9.7637	5.7422	5.5591	**5.2120**
		9	10.5954	9.5442	5.3315	5.1160	**4.6981**
		12	9.3149	8.6524	4.6519	4.4532	**3.8454**
12-step ahead	NMAE	3	10.2709	9.8030	7.9571	7.4296	**6.2146**
		6	11.0733	9.8055	8.3434	7.4303	**6.6665**
		9	11.9016	9.8551	8.1905	7.3367	**6.3884**
		12	9.7997	8.4262	6.7212	6.1360	**4.9810**
	NRMSE	3	12.9482	12.2571	10.1847	9.5633	**8.8435**
		6	13.9213	11.9185	10.3613	9.3666	**8.4221**
		9	14.6832	12.0225	10.3835	9.3934	**8.0788**
		12	12.7870	10.8204	8.7184	7.9646	**6.4888**
20-step ahead	NMAE	3	10.7971	10.4322	9.8362	9.5021	**9.4728**
		6	10.9828	9.8750	9.7214	9.7856	**9.7651**
		9	11.8759	9.8435	9.7624	9.7262	**9.6618**
		12	9.4529	8.0755	7.8977	7.7785	**7.7266**
	NRMSE	3	13.7177	12.9174	12.8506	12.2216	**12.2724**
		6	13.6620	11.9579	11.6980	11.9206	**11.9559**
		9	14.7073	12.1174	11.9378	11.9441	**11.9508**
		12	12.2981	10.1450	10.0368	9.9901	**9.8796**

Note: Bold values indicate the result corresponding to the highest accuracy in each case

Table 7 Performance evaluation of three forecast models on Elia data set

Forecast model		Month	Performance evaluation (%)		
			SSA-K-means-SVR	SSA-SOM-SVR	SSA-LSH-SVR
4-step ahead	NMAE	3	1.0276	1.0318	**1.0097**
		6	1.2217	1.2129	**1.2039**
		9	0.7333	0.7316	**0.7262**
		12	1.0930	1.1074	**1.0902**
	NRMSE	3	1.4938	1.4994	**1.4720**
		6	1.8757	1.8661	**1.8519**
		9	1.1571	1.1504	**1.1442**
		12	1.6697	1.6866	**1.6681**
12-step ahead	NMAE	3	3.9912	3.8897	**3.8509**
		6	5.2808	**5.1062**	5.1300
		9	3.4367	3.4471	**3.4160**
		12	5.1973	**5.1865**	5.1939
	NRMSE	3	5.4866	5.2941	**5.2901**
		6	7.3743	7.1293	**7.1261**
		9	4.7896	4.7814	**4.7600**
		12	7.0003	7.0041	**6.9645**
20-step ahead	NMAE	3	6.3282	6.1695	**6.1646**
		6	8.5485	**8.2494**	8.3639
		9	5.7213	**5.7239**	5.7608
		12	9.0435	**8.9914**	9.0728
	NRMSE	3	8.5143	**8.2779**	8.3184
		6	11.2943	11.0038	**10.9973**
		9	**7.7324**	7.7337	7.7690
		12	12.0813	**12.0250**	12.0753

Note: Bold values indicate the result corresponding to the highest accuracy in each case

more accurate forecast could be obtained by increasing the embedding dimension L; yet, when L reaches a critical point, the accuracy begins to decrease. On the other hand, the number of hash tables, l, and the number of LSHFs, k, do not significantly affect the forecast results. For the number of segments, N_s, the accuracy increases with the growth of N_s and finally converges to a certain value.

The best parameters for different kinds of data are different and should be obtained based on a large number of experiments. Some general advice may be given for carrying out the experiments more effectively. First, choose a suitable embedding dimension, L, for SSA ; considering the critical point described above, for each look-ahead step, there exists a value of L for which the NMAE is smallest. There is no strict restriction on the number of hash tables, l, and the number of LSHFs, k. Finally, to decide the best number of the similar segments, N_s, simulation studies should be carried out starting with $N_s = 150$ and increasing in steps of 50.

5 Conclusion

In this paper, a forecast model based on SSA, LSH and SVR is proposed for short-term wind speed and wind power prediction. The SSA is used to decompose the original time series into two components: the mean trend and the fluctuation component, where the mean trend reveals the slowly changing tendency of the original time series, and the fluctuation component represents its stochastic characteristics. The training input for SVR is a synthesis of the mean trend and the fluctuation component, which helps to ensure that these two components are independent and do not interfere with each other, leading to more accurate forecast results.

Moreover, this paper succeeds in making LSH effective and applicable in forecasting. It is used to classify the mean trend segments and find those that are similar to the forecast mean trend segment. Involving only these similar segments in the SVR improves forecast results. This has been proved by the smaller NMAEs and NMRSEs obtained by the SSA-LSH-SVR model over the SSA-SVR model.

Meanwhile, the application of LSH saves computation time remarkably than the SSA-K-means-SVR and SSA-SOM-SVR models.

The proposed model further improves over the SSA-LSH-SVR model by replacing the training input by the synthesis of the similar mean trend segments and the corresponding fluctuation component segments. Simulation studies have been conducted on two wind power data sets and two wind speed data sets, and the results have shown that the proposed model obtains remarkably smaller NMAEs and NRMSEs than the Per., SVR, SSA-SVR and SSA-LSH-SVR models, which indicates its advantage in both accuracy and stability.

Acknowledgements The project is supported by the Guangdong Innovative Research Team Program (No. 201001N0104744201) and the State Key Program of the National Natural Science Foundation of China (No. 51437006).

References

[1] Wang HZ, Wang GB, Li GQ et al (2016) Deep belief network based deterministic and probabilistic wind speed forecasting approach. Appl Energy 182:80–93
[2] Zhang Y, Liu K, Qin L et al (2016) Deterministic and probabilistic interval prediction for short-term wind power generation based on variational mode decomposition and machine learning methods. Energy Convers Manag 112:208–219
[3] GWEC. Global wind energy council (gwec)
[4] Jung J, Broadwater RP (2014) Current status and future advances for wind speed and power forecasting. Renew Sustain Energy Rev 31(2):762–777
[5] Zhang C, Wei H, Zhao X et al (2016) A gaussian process regression based hybrid approach for short-term wind speed prediction. Energy Convers Manag 126:1084–1092
[6] Zhao J, Guo Y, Xiao X et al (2017) Multi-step wind speed and power forecasts based on a WRF simulation and an optimized association method. Appl Energy 197:183–202
[7] Cardenasbarrera JL, Meng J, Castilloguerra E et al (2013) A neural network approach to multi-step-ahead, short-term wind speed forecasting. In: Proceedings of international conference on machine learning and applications, Miami, USA, 4–7 December 2013, 5 pp
[8] Al-Yahyai S, Charabi Y, Gastli A (2010) Review of the use of numerical weather prediction (NWP) models for wind energy assessment. Renew Sustain Energy Rev 14(9):3192–3198
[9] Wu JL, Ji TY, Li MS et al (2015) Multistep wind power forecast using mean trend detector and mathematical morphology-based local predictor. IEEE Trans Sustain Energy 6(4):1–8
[10] Akcay H, Filik T, Yan J (2017) Short-term wind speed forecasting by spectral analysis from long-term observations with missing values. Appl Energy 191:653–662
[11] Torres JL, Garca A, Blas MD et al (2005) Forecast of hourly average wind speed with arma models in Navarre (Spain). Sol Energy 79(1):65–77
[12] Finamore AR, Galdi V, Calderaro V et al (2017) Artificial neural network application in wind forecasting: an one-hour-ahead wind speed prediction. In: Proceedings of IET international conference on renewable power generation, London, UK, 21–23 September 2016, 6 pp
[13] Zhu L, Wu QH, Li MS et al (2013) Support vector regression-based short-term wind power prediction with false neighbours filtered. In: Proceedings of international conference on renewable energy research and applications, Madrid, Spain, 20–23 October 2013, 5 pp
[14] Yaslan Y, Bican B (2017) Empirical mode decomposition based denoising method with support vector regression for time series prediction: a case study for electricity load forecasting. Measurement 103:52–61
[15] Osrio GJ, Matias JCO, Catal JPS (2015) Short-term wind power forecasting using adaptive neuro-fuzzy inference system combined with evolutionary particle swarm optimization, wavelet transform and mutual information. Renew Energy 75:301–307
[16] Wang J, Wang J (2017) Forecasting stochastic neural network based on financial empirical mode decomposition. Neural Netw 90:8–20
[17] Zhang Y, Li C, Li L (2017) Electricity price forecasting by a hybrid model, combining wavelet transform, arma and kernel-based extreme learning machine methods. Appl Energy 190:291–305
[18] Wu JL, Ji TY, Li MS et al (2014) Multi-step wind power forecast based on similar segments extracted by mathematical morphology. In: Proceedings of IEEE PES Asia-Pacific power and energy engineering conference, Hong Kong, China, 7–10 December 2014, 6 pp
[19] Zhang X, Wang J, Zhang K (2017) Short-term electric load forecasting based on singular spectrum analysis and support vector machine optimized by cuckoo search algorithm. Electr Power Syst Res 146:270–285
[20] Afshar K, Bigdeli N (2011) Data analysis and short term load forecasting in Iran electricity market using singular spectral analysis (SSA). Energy 36(5):2620–2627
[21] Zhang Y, Lu H, Zhang L et al (2016) Video anomaly detection based on locality sensitive hashing filters. Pattern Recogn 59:302–311
[22] Alexandr A, Piotr I, Thijs L et al (2015) Practical and optimal LSH for angular distance. Computer science. arxiv: 1509.02897v1
[23] Zhang Y, Lu H, Zhang L et al (2016) Combining motion and appearance cues for anomaly detection. Pattern Recogn 51(C):443–452
[24] Lau KW, Wu QH (2008) Local prediction of non-linear time series using support vector regression. Pattern Recogn 41(5):1539–1547
[25] Feng C, Cui M, Hodge BM et al (2017) A data-driven multi-model methodology with deep feature selection for short-term wind forecasting. Appl Energy 190:1245–1257
[26] Bludszuweit H, Dominguez-Navarro JA, Llombart A (2008) Statistical analysis of wind power forecast error. IEEE Trans Power Syst 23(3):983–991

STATE GRID
STATE GRID ELECTRIC POWER RESEARCH INSTITUTE

Ling LIU received a B.Eng. degree in Electrical Engineering and Automation from University of Electronic Science and Technology of China, Chengdu, China, in 2016. She is currently pursuing an M.Sc. degree in Electric Power Systems and Automation at the School of Electric Power Engineering, South China University of Technology, Guangzhou, China. Her research interests include power/load forecasting and non-intrusive load monitoring.

Tianyao JI received a B.Sc. degree in Information Engineering in 2003, a B.A. degree in English in 2003 and an M.Sc. degree in Signal and Information Processing in 2006 from Xi'an Jiaotong University, Xi'an, China. In 2009, she obtained the Ph.D. degree in Electrical Engineering and Electronics from University of Liverpool, Liverpool, U.K. From 2010 to 2011, she worked as a research associate in University of Liverpool for 2 years. She is now an associate professor at School of Electric Power Engineering, South China University of Technology. Her research interests include mathematical morphology, signal and information processing, power system protection and evolutionary computation.

Mengshi LI received an M.Sc. (Eng.) degree with distinction in Information and Intelligence Engineering, and a Ph.D. degree in Electrical Engineering from the Department of Electrical Engineering and Electronics, the University of Liverpool, U.K. in 2005 and 2010. Then he worked as a Research Fellow in the University of Liverpool. In 2011, he became a Lecturer at South China University of Technology and a core member of Guangdong Innovation R&D Team. Currently, he is an Associate Professor at the School of Electric Power Engineering, South China University of Technology. His research interests include computational intelligence, artificial intelligence and their applications in power systems.

Ziming CHEN received the B.Eng. degree in Electrical Engineering and its Automation from South China University of Technology (SCUT), in 2012, and is currently working toward the Ph.D. degree at SCUT. His current research interests include deep learning technology and its applications in smart grid, signal processing and wind power forecasting.

Qinghua WU received a Ph.D. degree in Electrical Engineering from The Queen's University of Belfast, Belfast, U.K. in 1987, and subsequently worked as a Research Fellow. He joined the Department of Mathematical Sciences, Loughborough University, Loughborough, U.K. in 1991, as a Lecturer and subsequently as a Senior Lecturer. In 1995, he joined University of Liverpool, Liverpool, U.K. as the Chair of Electrical Engineering. Now he is with the School of Electric Power Engineering, South China University of Technology, Guangzhou, China, as a Distinguished Professor and the Director of Energy Research Institute of the University. Professor Wu has authored and coauthored more than 450 technical publications, including 240 journal papers. He is a Fellow of IEEE, Fellow of IET, Chartered Engineer and Fellow of InstMC. His current research interests include smart grid, mathematical morphology, smart energy automation chips, evolutionary computation, power system control and operation.