



Research on a Hybrid EMD-SVR Model for Time Series Prediction

Qiangqiang Yang^{1,2}, Dandan Liu^{2(✉)}, Yong Fang¹, Dandan Yang³, Yi Zhou⁴,
and Ziheng Sheng⁵

¹ School of Communication and Information Engineering,
Shanghai University, Shanghai, China

² College of Electronics and Information Engineering, Shanghai University
of Electric Power, Shanghai, China
liudandan@shiep.edu.cn

³ Tianjin Navigation Instruments Research Institute, Tianjin, China

⁴ MXSUN Software Company, Guangzhou, China

⁵ School of Electrical Engineering and Telecommunications,
The University of New South Wales, Sydney, NSW 2052, Australia

Abstract. Time series prediction methods were widely used in various fields. The prediction method for non-stationary and nonlinear time series was studied in this paper. This method decomposed non-stationary time series into stationary sub-sequences using the Empirical Mode Decomposition method. And then an appropriate time-step was chosen and the Support Vector Regression algorithm was applied to predict each stationary sub-sequence. The sum of predicted values was the forecasting results of the original sequence. The method was applied to building energy consumption datasets, which were collected in some buildings. The experimental results showed that the hybrid algorithm of Support Vector Regression and Empirical Mode Decomposition had higher accuracy and was suitable for predicting non-linear and non-stationary time series. Moreover, this hybrid algorithm was used to predict the time series with outliers and to test its noise-resistant performance. The forecasting results also illustrated EMD-SVR algorithm was more robust than SVR algorithm.

Keywords: Time series · Empirical Mode Decomposition · Support Vector Regression · Building energy consumption · Prediction

1 Introduction

Time series is a set of data collected sequentially usually at fixed intervals of time. It is very common in real life. Basically, the goal of time series prediction is to estimate future values based on current and past data [1]. For example, forecasting models of building energy consumption can provide more reasonable building management solutions for building managers. Prediction models for financial time series will help people find out

the macro market operation rules. Therefore, many studies focused on how to develop accurate and effective time series prediction models.

Classic time series prediction models are based on stochastic process theory and mathematical statistics, which are divided into two categories: Auto-Regressive Moving Average Model (ARMA) and Autoregressive Integrated Moving Model (ARIMA). These classic time series prediction methods have achieved good prediction results in many fields [2, 3]. However, some literature also pointed out that it is difficult to develop accurate prediction models for nonlinear and non-stationary time series using the classic methods. Therefore, machine learning algorithms were applied to predicting nonlinear and non-stationary time series [4–6]. Among them, the support vector machine (SVM) algorithm is one of the most effective machine learning algorithms and has been widely verified in various fields [1].

However, it is still complicated to develop accurate prediction models for nonlinear and non-stationary time series even using SVM algorithm in some fields [7]. Then some hybrid algorithms that can decompose and predict non-stationary time series were discussed [8–11]. These algorithms turned non-stationary series into stationary subsequences and then analyzed decomposed subsequences to improve prediction accuracy.

In this paper, a time series prediction model for building energy consumption was established based on the EMD-SVM hybrid algorithm. In order to verify the robustness of the algorithm, it was applied to datasets with outliers, which proved that the algorithm had anti-noise ability and was suitable for the prediction of building energy consumption.

2 Methodology

2.1 Theory of Empirical Mode Decomposition (EMD)

Most of the natural phenomena are nonlinear and non-stationary systems. The Empirical Mode Decomposition (EMD) method has proposed by N.E. Huang et al. [12], which is effective in analyzing the non-linear and non-stationary time series. There are also some other transformation methods including Fourier transform, wavelet transform and so on. Fourier transform is always used to decompose linear and stationary signal and the Wavelet analysis depends on the Fourier transform even though it can be used to analyze non-stationary signals. Moreover, the basic wavelet function should be given before the wavelet analysis and it has a non-adaptive characteristic. Once the basic wavelet is selected, one will have to use it to analyze all the data. In order to solve these problems, EMD was adopted in lots of experiments and the results showed it has better performance.

The EMD method can be summarized as follows [12, 13]:

- Given a signal $x(t)$, identify all extrema of $x(t)$.
- Interpolate between minima or maxima, ending up with some envelope $s_{min}(t)$ or $s_{max}(t)$ and compute the mean of envelope $m_1(t)$:

$$m_1(t) = (s_{max}(t) + s_{min}(t))/2 \quad (1)$$

- Extract the component $c_1(t) = x(t) - m_1(t)$, $c_1(t)$ is an Intrinsic Mode Function (IMF).
- Iterate on the residual $r_1(t) = x(t) - c_1(t)$.
- Repeat the steps until the decomposition results satisfy the stopping criterion. The original time series is decomposed into multiple IMFs and one corresponding residual:

$$x(t) = \sum_{i=1}^N c_i(t) + r_N(t) \tag{2}$$

where the residual $r_N(t)$ is computed as formula (3):

$$\begin{cases} r_1(t) - c_2(t) = r_2(t) \\ r_2(t) - c_3(t) = r_3(t) \\ \dots \\ r_{N-1}(t) - c_N(t) = r_N(t) \end{cases} \tag{3}$$

2.2 Theory of Support Vector Regression (SVR)

The support vector machines algorithm is a machine learning method that was proposed by Vapnik [14]. SVM is called Support Vector Regression (SVR) when it is used to model and predict.

Actually, the purpose of the SVM algorithm is to classify the data points. If the data points are not linearly separable, they will be mapped into the N-dimensional feature space to make them linearly separable. That is to say, SVM will find an (N-1)-dimensional hyperplane in an N-dimensional space to classify the data points.

So kernel functions and optimizer algorithm are two parts of SVMs. The non-linear data is divided into high-dimensional space by kernel function and the optimizer algorithm is used to find the hyperplane in high dimensional space. SVR method minimizes the empirical risk and identifies an optimum hyperplane to maximize the distance separating the training data into subsets and minimize training error [15].

In this paper, the kernel function, radial basis function (RBF), was chosen and optimizer parameters were searched by the Particle swarm optimization (PSO) algorithm.

2.3 EMD-SVR Method for Time Series Analysis

EMD-SVR is a hybrid algorithm. The complete procedure of developing prediction model for time series using EMD-SVR algorithm can be described as Fig. 1.

1. Build the new time series for prediction models. Suppose that $y = \{y_1, \dots, y_n\}$ is an original time series, then a new series $\bar{y}_i = \{(\bar{x}_i, \bar{z}_i)\}$ will be reconstructed, where $\bar{x}_i = \{y_i, y_{i+1}, \dots, y_{i+d-1}\}$, $\bar{z}_i = y_{i+d}$ and d is time-steps. \bar{x}_i is the input of SVR model and \bar{z}_i is the output of SVR model [16]. \bar{x} is described by formula (4):

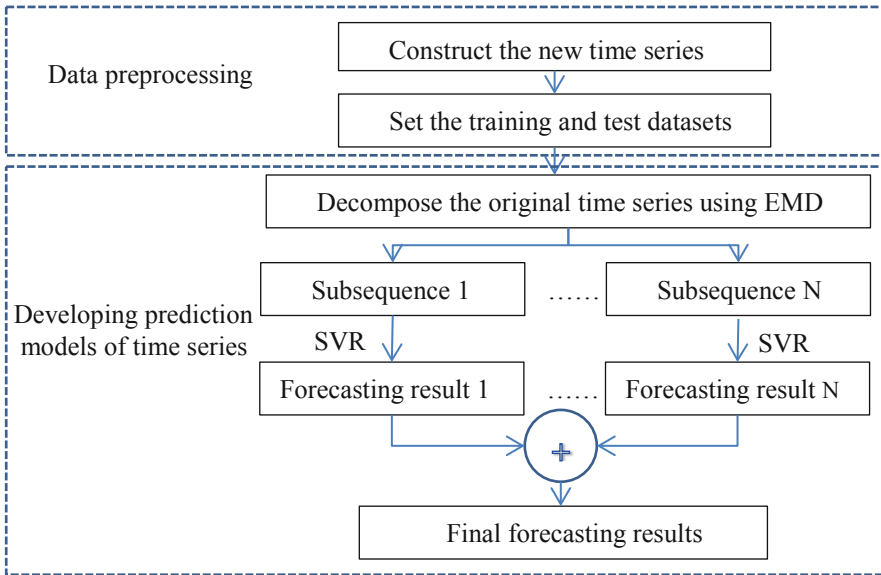


Fig. 1. The experimental procedure of SVR-EMD algorithm

$$\bar{x} = \begin{bmatrix} y_1 & y_2 & \cdots & y_d \\ y_2 & y_3 & \cdots & y_{d+1} \\ & & \cdots & \\ y_{n-d} & y_{n-d+1} & \cdots & y_{n-1} \end{bmatrix} \tag{4}$$

\bar{z} is showed by formula (5):

$$\bar{z} = \begin{bmatrix} y_{d+1} \\ y_{d+2} \\ \cdots \\ y_n \end{bmatrix} \tag{5}$$

namely \bar{x}_i is a $(n-d) \times d$ matrix and \bar{z}_i is a $(n-d) \times 1$ matrix.

2. Set the training dataset and test dataset of prediction models. The number of data in the training set and test set can be specified according to the following formula (6).

$$\begin{cases} N = (n - d) \times 3/4 \\ M = (n - d) \times 1/4 \end{cases} \tag{6}$$

Then training dataset and test dataset were confirmed by formula (7):

$$\begin{cases} \bar{y}_{\text{train}} = \{(\bar{x}_i, \bar{z}_i)\}(i = 1, \cdots N) \\ \bar{y}_{\text{test}} = \{(\bar{x}_i, \bar{z}_i)\}(i = (N + 1), \cdots (N + M)) \end{cases} \tag{7}$$

- Determine the evaluation criteria of prediction models. The two parameters, mean squared error (MSE) and squared correlation coefficient (R^2), were selected to evaluate the models. They are defined by formula (8) and (9), where R^2 is a number between 0 and 1. The lower MSE and higher R^2 would be expected.

$$MSE = \frac{1}{l} \sum_{i=1}^l (f(x_i) - y_i)^2 \tag{8}$$

$$R^2 = \frac{\left(l \sum_{i=1}^l f(x_i)y_i - \sum_{i=1}^l f(x_i) \sum_{i=1}^l y_i \right)^2}{\left(l \sum_{i=1}^l f(x_i)^2 - \left(\sum_{i=1}^l f(x_i) \right)^2 \right) \left(l \sum_{i=1}^l y_i^2 - \left(\sum_{i=1}^l y_i \right)^2 \right)} \tag{9}$$

- Develop the forecasting model using training dataset. The steps of developing forecasting model for time series using EMD-SVR was shown as follows:

- Decompose time series into some subsequences.
- For every subsequence, the range of d was determined. For example, d can be chosen between 2 to 24 for the time series of building energy consumption.
for d = 2 to 24

Apply the SVR algorithm and develop subsequence prediction models.

Compute the sum of prediction results to get the final results.

Compute MSE and R^2 .

end for

- Choose the best time-steps according to MSE and R^2 .
- Develop the forecasting model with the chosen time-steps for the original time series.

3 Experiments and Results

In an office building, the energy consumption data was collected. There are 120 data in one series because there are 120 working hours in one week.

After the ADF test, the non-stationary time series was selected to verify the validity of the EMD-SVR algorithm. Then the input and output datasets of SVM were illustrated by formula (10) and (11):

$$\bar{x} = \begin{bmatrix} y_1 & y_2 & \cdots & y_{24} \\ y_2 & y_3 & \cdots & y_{25} \\ \cdots & & & \\ y_{96} & y_{97} & \cdots & y_{119} \end{bmatrix} \tag{10}$$

$$\bar{z} = \begin{bmatrix} y_{25} \\ y_{26} \\ \cdots \\ y_{120} \end{bmatrix} \tag{11}$$

So training dataset and test dataset were confirmed by formula (12):

$$\begin{cases} \bar{y}_{\text{train}} = \{(\bar{x}_i, \bar{z}_i)\}(i = 1, \dots, 72) \\ \bar{y}_{\text{test}} = \{(\bar{x}_i, \bar{z}_i)\}(i = 73, \dots, 96) \end{cases} \quad (12)$$

3.1 The EMD-SVR Forecasting Model of Non-stationary Time Series for Building Energy Consumption

First, the non-stationary time series of energy consumption was decomposed into some subsequences as shown in Fig. 2. Then the prediction models were developed using SVR algorithm for every subsequence.

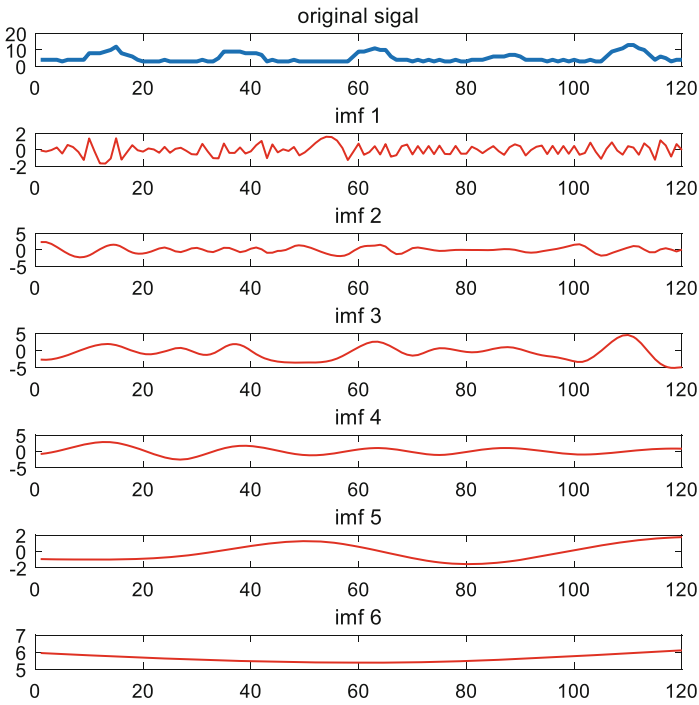


Fig. 2. The original time series and decomposed subsequences of building energy consumption

The comparison of prediction results between EMD-SVR and SVR was shown in Table 1. It illustrated that prediction results of EMD-SVR algorithm were better than SVR algorithm when different time-steps was chosen.

From Table 1 we knew the prediction performance was better when time-steps was 4. In this case, R^2 is 0.925743 and 0.943764 for the training dataset and the test dataset, respectively. As we knew the closer R^2 is to 1, the better the prediction results. Further, the value of R^2 for the test set is greater than the value of R^2 for the training set and they were all greater than 0.9, which indicated that the SVR predicting model was not

Table 1. The Comparison of Prediction performance between SVR and EMD-SVM models using different time-steps for building energy consumption

Time-steps	Algorithm	Training dataset		Testing dataset	
		MSE	R2	MSE	R2
2	EMD-SVR	0.571094	0.906797	1.196776	0.89809
	SVR	1.597639	0.750962	2.13689	0.803385
3	EMD-SVR	0.492307	0.922563	1.01905	0.918678
	SVR	0.836858	0.881241	1.807069	0.868493
4	EMD-SVR	0.482795	0.925743	0.861123	0.943764
	SVR	0.877369	0.874117	1.997539	0.884993
5	EMD-SVR	0.447164	0.931439	1.488371	0.90603
	SVR	1.404345	0.777708	1.728945	0.853921
6	EMD-SVR	0.458997	0.932865	1.27686	0.921485
	SVR	1.34057	0.791787	1.827347	0.851383
7	EMD-SVR	0.295075	0.965816	1.019978	0.929475
	SVR	1.343158	0.794188	1.724895	0.851919
8	EMD-SVR	0.28713	0.967779	1.12577	0.924394
	SVR	0.584819	0.919103	3.101792	0.73007
9	EMD-SVR	0.291151	0.970087	1.068173	0.924757
	SVR	1.378072	0.786893	1.786711	0.855755
10	EMD-SVR	0.28603	0.969325	1.307217	0.919837
	SVR	0.630584	0.932211	2.345651	0.87277
11	EMD-SVR	0.288488	0.968742	1.532112	0.894996
	SVR	0.501229	0.937123	2.324805	0.878927
12	EMD-SVR	0.295423	0.964094	1.948183	0.863141
	SVR	0.569224	0.934161	2.194049	0.887287
13	EMD-SVR	0.324873	0.955616	2.113517	0.834063
	SVR	0.984289	0.873426	2.512899	0.889048
14	EMD-SVR	0.390129	0.945416	1.692577	0.887154
	SVR	0.571432	0.942153	3.098651	0.855208
15	EMD-SVR	0.372134	0.941069	1.612551	0.892333
	SVR	0.448016	0.94655	3.103344	0.829688
16	EMD-SVR	0.39435	0.935235	2.180305	0.828951
	SVR	0.452569	0.952432	3.558848	0.837837
17	EMD-SVR	0.313769	0.948636	2.086048	0.830763
	SVR	0.451582	0.949555	3.592484	0.853068
18	EMD-SVR	0.26255	0.956753	1.460961	0.913228
	SVR	0.431549	0.948027	3.543799	0.843864
19	EMD-SVR	0.31237	0.951181	1.487005	0.905245
	SVR	0.476713	0.956992	5.24637	0.780169
20	EMD-SVR	0.459039	0.953733	5.217063	0.789639
	SVR	0.320427	0.959652	7.124436	0.371715

(continued)

Table 1. (continued)

Time-steps	Algorithm	Training dataset		Testing dataset	
		MSE	R2	MSE	R2
21	EMD-SVR	0.2697	0.957281	1.82646	0.885663
	SVR	0.922285	0.832968	3.082533	0.838191
22	EMD-SVR	1.044031	0.825763	3.315175	0.829745
	SVR	0.297115	0.962591	4.956615	0.584612
23	EMD-SVR	0.228852	0.964429	1.356089	0.913678
	SVR	0.905582	0.843615	2.261825	0.854461
24	EMD-SVR	0.293686	0.962969	3.510583	0.718327
	SVR	0.416496	0.943969	4.50047	0.765655

over-fitted. The prediction results of energy consumption models were seen in Fig. 3 and Fig. 4. It can also be seen that the forecasting curves were very similar to the actual curves.

3.2 Comparison of Noise-Resistant Capabilities for SVR and EMD-SVR

In fact, it is normal that forecasting results are disturbed by all kinds of noise, which is one of the mainly reasons that forecasting accuracy cannot be improved. Forecasting models were developed using SVR and EMD-SVR based on datasets with outliers in this section. The forecasting results responded the noise-resistant capability of EMD-SVR.

To learn the noise-resistant capability of different algorithms, the datasets that added outlier instead of original datasets were generated to simulate abnormal energy consumption. Suppose that the dataset contains A% outliers and there are N data in this dataset. The $N \cdot A\%$ integers will be generated randomly between 1 and N and these integers are the serial number of outliers.

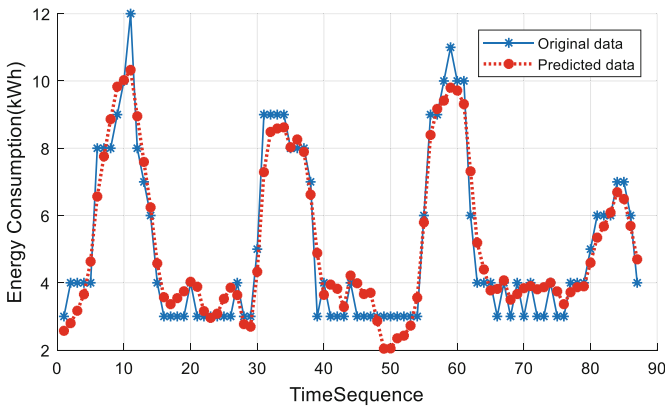


Fig. 3. Prediction results of building energy consumption using ESD-SVR method for the training dataset

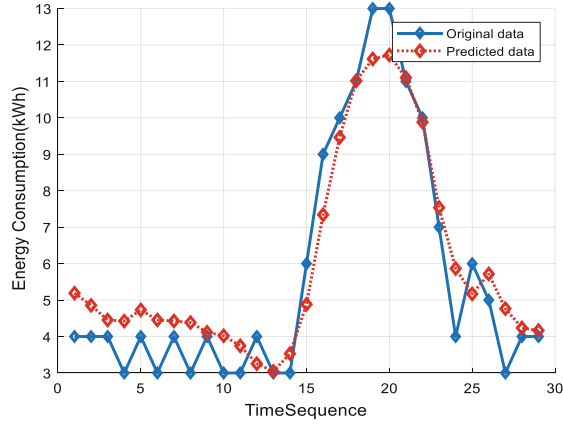


Fig. 4. Prediction results of building energy consumption using ESD-SVR method for the test dataset

$$j = randi([1, N], 1, N * A\%) \tag{13}$$

Then the dataset with outliers can be described by formula (14):

$$y(i) = y(i) + a\delta_i \cdot y(i) \quad i = 1, 2 \dots N \tag{14}$$

where a is a constant that can be set to 0.5 and δ_i is an unit function,

$$\delta_i = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases} \tag{15}$$

An energy consumption time series with outliers was shown in Fig. 5.

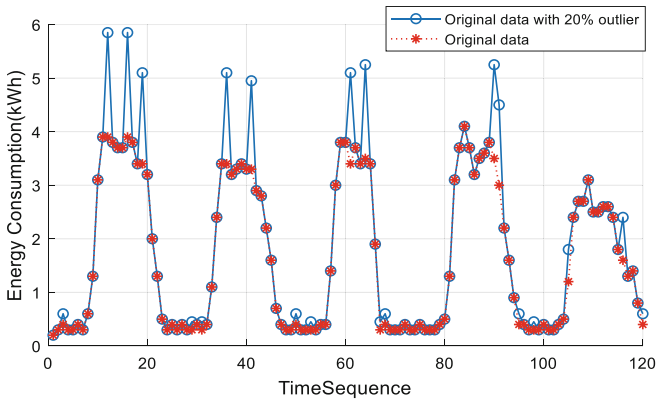


Fig. 5. Comparison of original dataset and original dataset with outliers

The SVR and EMD-SVR algorithms were applied to the datasets with 10%, 20%, 30% and 40% outliers, and the experimental results were shown in the Table 2, Table 3, Table 4 and Table 5. The comparisons among predicted values, original values and original values added to outliers were seen in Fig. 6, Fig. 7, Fig. 8 and Fig. 9. From these tables and figures we knew EMD-SVR algorithm was more robust than SVR algorithm. It meant that EMD-SVR algorithm had better noise-resistant capabilities.

Table 2. The Comparison of prediction performance between SVR and EMD-SVM models for building energy consumption dataset with 10% outliers

Time-steps	Algorithm	Training dataset		Testing dataset	
		MSE	R2	MSE	R2
2	EMD-SVR	1.049133	0.879419	0.998073	0.913055
	SVR	0.287353	0.899854	0.404267	0.657795
3	EMD-SVR	1.042892	0.894517	1.059364	0.909933
	SVR	0.282261	0.895995	0.392198	0.668
4	EMD-SVR	0.910117	0.900216	1.146224	0.901543
	SVR	0.250799	0.899624	0.398196	0.711845
5	EMD-SVR	0.869306	0.882654	1.554877	0.889715
	SVR	0.249604	0.898202	0.259128	0.787533

Table 3. The Comparison of prediction performance between SVR and EMD-SVM models for building energy consumption dataset with 20% outliers

Time-steps	Algorithm	Training dataset		Testing dataset	
		MSE	R2	MSE	R2
13	EMD-SVR	0.947983	0.891675	1.6985	0.865301
	SVR	1.34509	0.80497	3.979598	0.667633
14	EMD-SVR	1.115337	0.87121	1.662653	0.869568
	SVR	1.251211	0.818155	3.754576	0.676766
15	EMD-SVR	0.87896	0.892769	1.687701	0.865985
	SVR	0.901377	0.916191	4.943163	0.698825
16	EMD-SVR	0.7454	0.90294	2.266961	0.818882
	SVR	0.88956	0.887116	4.555295	0.75042

Table 4. The Comparison of prediction performance between SVR and EMD-SVM models for building energy consumption dataset with 30% outliers

Time-steps	Algorithm	Training dataset		Testing dataset	
		MSE	R2	MSE	R2
2	EMD-SVR	1.165155	0.901028	2.544622	0.927529
	SVR	2.206972	0.683425	3.55639	0.708367
3	EMD-SVR	1.096337	0.889661	2.460228	0.937368
	SVR	2.294979	0.693566	3.901687	0.694548
4	EMD-SVR	1.115835	0.882032	2.323355	0.92792
	SVR	2.316277	0.686902	3.950681	0.707086
5	EMD-SVR	1.334761	0.870688	2.62816	0.912298
	SVR	2.287445	0.719084	2.795342	0.79984

Table 5. The Comparison of prediction performance between SVR and EMD-SVM models for building energy consumption dataset with 40% outliers

Time-steps	Algorithm	Training dataset		Testing dataset	
		MSE	R2	MSE	R2
2	EMD-SVR	2.256755	0.773657	4.016026	0.840588
	SVR	2.304331	0.671405	2.667304	0.769721
3	EMD-SVR	2.042788	0.793416	2.935345	0.885079
	SVR	2.41279	0.677468	2.677986	0.776113
4	EMD-SVR	2.03426	0.793219	2.616165	0.886622
	SVR	2.471562	0.660606	2.563792	0.77963
5	EMD-SVR	2.093942	0.797823	2.7458	0.915917
	SVR	2.621712	0.695381	2.673072	0.818439

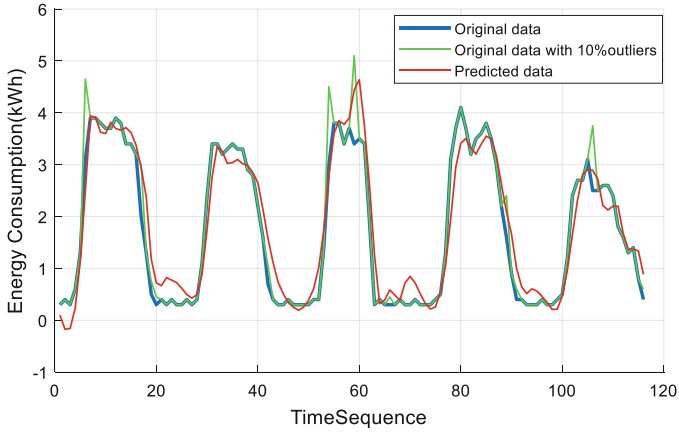


Fig. 6. Prediction results of building energy consumption using ESD-SVR method for the dataset with 10% outliers

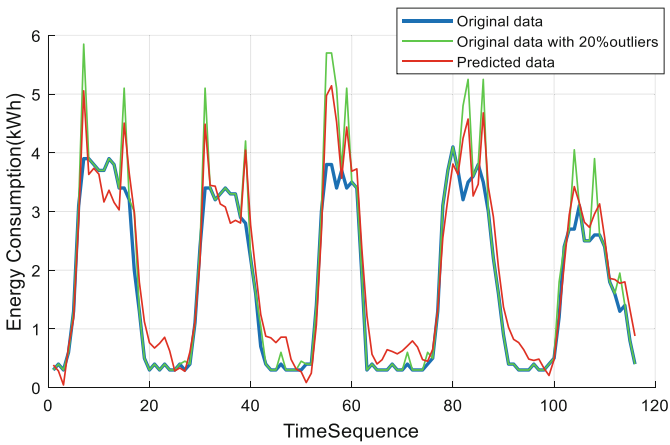


Fig. 7. Prediction results of building energy consumption using ESD-SVR method for the dataset with 20% outliers

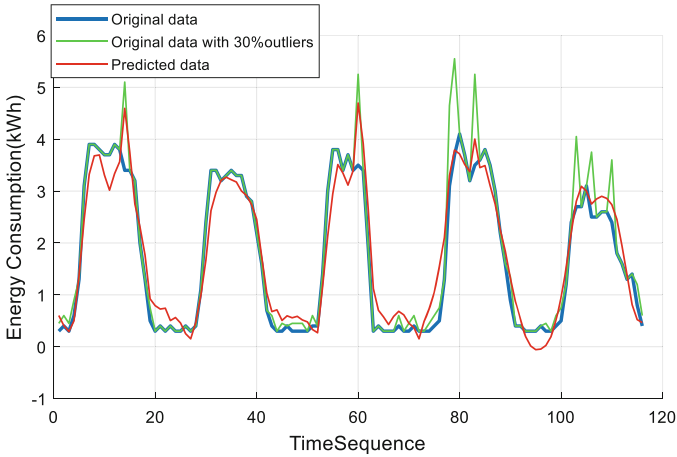


Fig. 8. Prediction results of building energy consumption using ESD-SVR method for the dataset with 30% outliers

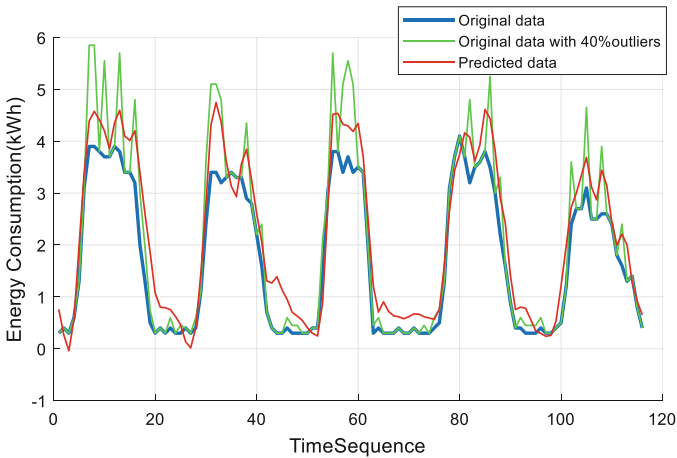


Fig. 9. Prediction results of building energy consumption using ESD-SVR method for the dataset with 40% outliers

4 Conclusion

In this paper, a method for predicting nonlinear and non-stationary time series was studied. The prediction models of time series were developed based on a hybrid EMD-SVR algorithm. Firstly, the nonlinear and non-stationary time series was decomposed into some subsequences using EMD. And then the prediction model for every subsequence was developed based on SVR. The sum of the prediction value of each subsequence is the final prediction result.

The forecasting model of time series for building energy consumption can be used to learn the law of building energy consumption and save energy. The EMD-SVR algorithm

was applied to time series of building energy consumption. The results showed that the EMD-SVR algorithm was better than SVR algorithm. Moreover, the method was used to the dataset with outliers for verifying the robustness of the algorithm. The forecasting results of datasets with outliers illustrated EMD-SVR algorithm was more robust than SVR algorithm.

References

1. Sapankevych, N., Sankar, R.: Time series prediction using support vector machines: a survey. *IEEE Comput. Intell. Mag.* **4**(2), 24–38 (2009)
2. Wang, Q., Li, S.Y., Li, R.R., et al.: Forecasting U.S. shale gas monthly production using a hybrid ARIMA and metabolic nonlinear grey model. *Energy* **160**(10), 378–387 (2018)
3. Rounaghi, M.M., Zadeh, F.N.: Investigation of market efficiency and financial stability between S&P 500 and London stock exchange: monthly and yearly forecasting of time series stock returns using ARMA model. *Phys. A Stat. Mech. Appl.* **456**(15), 10–21 (2016)
4. Mohapatra, U.M., Majhi, B., Satapathy, S.C.: Financial time series prediction using distributed machine learning techniques. *Neural Comput. Appl.* **31**(8), 3369–3384 (2017). <https://doi.org/10.1007/s00521-017-3283-2>
5. Chen, S., Mihara, K., Wen, J.: Time series prediction of CO₂, TVOC and HCHO based on machine learning at different sampling points. *Build. Environ.* **146**(12), 238–246 (2018)
6. Li, Z., Ye, L., Zhao, Y., Song, X., Teng, J., Jin, J.: Short-term wind power prediction based on extreme learning machine with error correction. *Protect. Control Mod. Power Syst.* **1**(1), 1–8 (2016). <https://doi.org/10.1186/s41601-016-0016-y>
7. Shiu, M.C., Wei, L.Y., Liu, J.W., et al.: A hybrid one-step-ahead time series model based on GA-SVR and EMD for forecasting electricity loads. *J. Appl. Sci. Eng.* **20**(4), 467–476 (2017)
8. Yuanfang, X., Yuanyuan, J., Xuemei, Z.: Gas outburst prediction model based on empirical mode decomposition and extreme learning machine. *Recent Adv. Electr. Electron. Eng.* **8**(1), 50–56 (2015)
9. Yaslan, Y., Bican, B.: Empirical mode decomposition based denoising method with support vector regression for time series prediction: a case study for electricity load forecasting. *Measurement* **103**(6), 52–61 (2017)
10. Wang, Z.Y., Qiu, J., Li, F.F.: Hybrid models combining EMD/EEMD and ARIMA for long-term streamflow forecasting. *Water* **10**(7), 1–12 (2018)
11. Yabing, J., Changwen, L., Fengrong, B., et al.: Sensitivity analysis on physical noise sources of small generator based on EMD-SVM. *J. Tianjin Univ.* **50**(10), 1077–1083 (2017)
12. Huang, N.E., Shen, Z., Long, S.R.: The empirical mode decomposition and Hilbert spectrum for nonlinear and nonstationary time series analysis. *Proc. Roy. Soc. London A* **454**(1), 903–995 (1998)
13. Rilling, G., Flandrin, P., Goncalves P.: On empirical mode decomposition and its algorithms. In: 2003 IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing, pp. 8–11. IEEE, Grado (2003)
14. Vapnik, V.: *The Nature of Statistical Learning Theory*. Springer, New York (1995). <https://doi.org/10.1007/978-1-4757-2440-0>
15. Yu, J.Q., Nan, Y.L., Zhang, Y., Yang, X.: Research on fractal characteristics of building energy consumption time series. In: 4th International Conference on Energy Equipment Science and Engineering, pp. 1–8. IOP, Xi'an, China (2019)
16. Ghelardoni, L., Ghio, A., Anguita, D.: Energy load forecasting using empirical mode decomposition and support vector regression. *IEEE Trans. Smart Grid* **4**(1), 549–556 (2013)