Journal of Software Engineering

ISSN 1819-4311

Academic Journals Inc.

www.academicjournals.com

Journal of Software Engineering 10 (1): 119-128, 2016
ISSN 1819-4311 / DOI: 10.3923/jse.2016.119.128

Forecasting of Telephone Traffic Based on EMD and Gaussian Process-Gray

[1]Xiuxiu Song, [1]Zhenhong Jia, [1]Xizhong Qin, [2]Chuanling Cao and [2]Hongmei Niu
[1]School of Information Science and Engineering, Xinjiang University, Urumqi, 830046, China
[2]Subsidiary Company of China Mobile in Xinjiang, Urumqi, 830063, China

Corresponding Author: Zhenhong Jia, Xinjiang University, No. 14 Sheng Li Road, Tian Shan, Urumqi, Xinjiang, China Tel: 15739575859

ABSTRACT

To improve the prediction accuracy of telephone traffic, a combined forecasting model which takes the influence of multiple factors into consideration was proposed in this study and it combined empirical mode decomposition and Gaussian process model and gray prediction model. Correlation analysis was applied to obtain the key factors which influenced the telephone traffic. Through the simulation experiments for telephone traffic data collected in practice, the simulation results showed that the proposed model has the superiority of higher prediction accuracy and easier to implement.

Key words: Forecasting of telephone traffic, multiple factors, EMD, Gaussian process, gray prediction, combined model

INTRODUCTION

According to the latest statistics, China mobile users have reached 1.5 billion. With the advent of the era of 4G in China, the future of mobile communication users will also be greatly increased. After leap-forward development and technology innovation in the mobile communication technology, the requirement of processing the mobile communication network data also gradually improved and telephone traffic data occupied very important position. Forecasting telephone traffic effectively can make the operator grasp the dynamic changes in future market, reduce the risk of decision-making encountered and then achieve the decision goal.

The traditional telephone traffic prediction usually adopt quantitative prediction method of time series, the commonly used are exponential smoothing prediction, curve fitting prediction and Markov model, the Auto-Regressive and Moving Average model (ARMA), etc. Artificial intelligence method is developing rapidly now, many scholars have succeeded in applying these methods to various kinds of prediction, the problem of rainfall prediction and wind speed prediction can be solved by using neural network (Wu *et al*., 2015; Chandra *et al*., 2014). The groundwater level can be predicted by using Support Vector Machine (SVM) (Suryanarayana *et al*., 2014). Good prediction effect have been achieved in these documents but these are insufficient, it is still difficult to choose structural parameters of neural network and convergence speed is slow, what's more, it is easy to fall into local optimum. Although the prediction effect of Support Vector Machine (SVM) is better but there is no determination method to select kernel function and regular parameters. Gaussian process based on Bayesian theory (Salcedo-Sanz *et al*., 2014) is a new machine learning method, it has significant advantages in dealing with complex problems such as high dimension

and small sample and nonlinear. This method is easy to implement, super parameters can be got adaptively and the outputs of the method have significance of probability, so this is a better method to predict telephone traffic in the current.

In addition, a lot of practice research show that the combined forecasting model is much better than single forecasting model, for telephone traffic data with complicated change rule and influenced by a variety of factors, the telephone traffic data can be described more accurate and comprehensive by using the combined forecasting model. Wavelet transform (Sudheer and Suseelatha, 2015) and Empirical Mode Decomposition (EMD) (Wang *et al*., 2014) are the widely used decomposition method (Yu *et al*., 2015). The advantages of multi-resolution analysis belong to both EMD and Wavelet transform but the difficulties in selecting on wavelet basis in Wavelet transform can be overcome in EMD, so EMD is a better choice in this study.

To sum up, considering the telephone traffic is influenced by many factors, the key factors influencing the traffic are extracted first in this study and then the Gaussian process and grey forecasting model based on the EMD decomposition is proposed to forecast the telephone traffic, through the simulation experiments for mobile telephone traffic data collected in practice, the feasibility and effectiveness of the method is confirmed.

MATERIALS AND METHODS

Empirical mode decomposition: The empirical mode decomposition method (Huang *et al*., 1998) is a method of adaptive time-domain processing and need not to convert the frequency domain, it is suitable for nonlinear and non-stationary time series processing. The basic idea of EMD is that a complex signal is decomposed into a finite number of intrinsic mode functions in different time scales (Intrinsic Mode Function, IMF), it can be written as:

$$X(t)=\sum_{i=1}^{n} c_i(t)+r_n(t) \tag{1}$$

where, X (t) is the original signal, $c_i(t)$ is the first intrinsic mode function, $r_n(t)$ is a residual term, namely the trend.

Gaussian process: The statistical characteristic of Gaussian process depends on its mean value function and covariance function (Reggente *et al*., 2014).

$$\begin{cases} m(x)=E\left[f(x)\right] \\ C(x,x')=E\left[\left(f(x)-m(x)\right)\left(f(x')-m(x')\right)\right] \end{cases} \tag{2}$$

where, x, $x' \in R^d$, as d dimensional random variables.

Considering the regression model:

$$y = f(x)+\varepsilon \tag{3}$$

where, x is the input vector, y is the observed value, noise $\varepsilon \sim N(0,\sigma n^2)$.

The prior distribution of y is $y \sim N(0, C(X,X)+\sigma_n^2 I_n)$.

The joint prior distribution of observed value y and predicted value y' is:

$$\begin{bmatrix} y \\ y' \end{bmatrix} \sim N\left(0, \begin{bmatrix} C(X,X)+\sigma_n^2 I_n & C(X,x') \\ C(x',X) & C(x',x') \end{bmatrix}\right)$$

where, C (X, X) is n×n order covariance matrix, element cij is a measure of the correlation between xi and xj, C (X, x') is the covariance matrix between the training set X and the test points x', C (x', x') is covariance of x' only, In is the unit matrix.

According to the Bayesian posterior probability formula, the posterior distribution of y' is:

$$y' \left| (X, y, x') \sim N\left(\mu_{y'}, \sigma_{y'}^2\right)\right.$$

Then expectation and variance of predicted value y'can be obtained:

$$\mu_{y'} = C(x',X)\left[C(X,X)+\sigma_n^2 I_n\right]^{-1} y \tag{4}$$

$$\sigma_{y'}^2 = C(x',x') - C(x',X)\left[C(X,X)+\sigma_n^2 I_n\right]^{-1} C(X,x') \tag{5}$$

Covariance function is equivalent to the kernel function in Gaussian process, one common covariance function is square covariance, namely:

$$C\left(x_i, x_j\right) = \sigma_f^2 \exp\left[-\frac{1}{2}\left(x_i - x_j\right)^T M\left(x_i - x_j\right)\right] + \sigma_n^2 \delta^{ij} \tag{6}$$

where, σ_f^2 is signal variance of kernel function, M = diag(l^{-2}) is the diagonal matrix of super parameter, l is variance dimension, δ^{ij} is Kronecker symbol.

$\theta = \{M, \sigma_f^2, \sigma_n^2\}$ is super parameter. The optimal super parameters are obtained usually by logarithmic function maximum likelihood method. Negative logarithmic likelihood function L (θ) and partial derivative of super parameter θ is:

$$L(\theta) = \frac{1}{2} y^T C^{-1} y + \frac{1}{2}\log|C| + \frac{n}{2}\log 2\pi \tag{7}$$

$$\frac{\partial L(\theta)}{\partial \theta_i} = \frac{1}{2} tr\left[\left(\alpha\alpha^T - C^{-1}\right)\frac{\partial C}{\partial \theta_i}\right] \tag{8}$$

where, $\alpha = C^{-1} y$.

Grey prediction: The Grey Model (GM) is used in Grey prediction (Boran, 2015) to estimate and forecast the development change rule of the behavior characteristic in the system, forecast sequence and analyze the development trend for the future on the basis of the existing data.

Make x(0) = (x(1), x(2),....,x(n)) and make accumulation to generate x(k) =Σx(m), eliminate the randomness and volatility in the data, m = 1, there is:

$$x = (x(1), x(2),....,x(n)) = (x(1), x(1),+x(2),....,x(n-1)+x(n))$$

An albino equation based on x:

$$\frac{dx}{dt} + ax = u \tag{9}$$

This is GM (1, 1), in this formula: a is a constant known as the development of grey number, u is endogenous control grey numbers and the constant input of the system.

Prediction model and process: In this study, Gaussian process and grey prediction model based on the EMD is used to forecast telephone traffic, the concrete implementation steps are as follows:

- Correlation analysis is firstly applied to the telephone traffic data and extract the key factors influencing the telephone traffic
- EMD is used to decompose the telephone traffic data and get different frequency component of the IMF1, IMF2, IMF3, IMF4, IMF5 and the trend component R
- The IMF component and the obtained key factors are loaded into Gaussian process model to predict and get the predicted value Pre1, Pre2, Pre3, Pre4, Pre5
- The trend component is loaded into gray prediction model to predict and get the predicted value Pre6
- The superposition of each predictive values is the forecasting result Pre = Pre1+Pre2+Pre3+ Pre4+Pre5+Pre6

All the five steps can be described by a diagram visually and the diagram of this prediction model is shown in the Fig. 1.

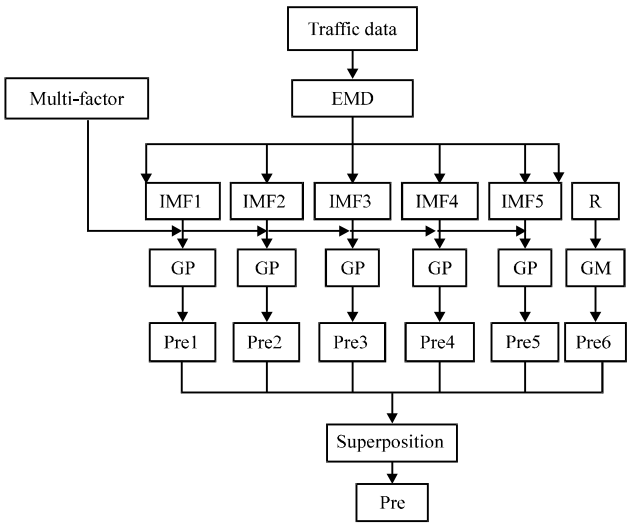

Fig. 1: Prediction model

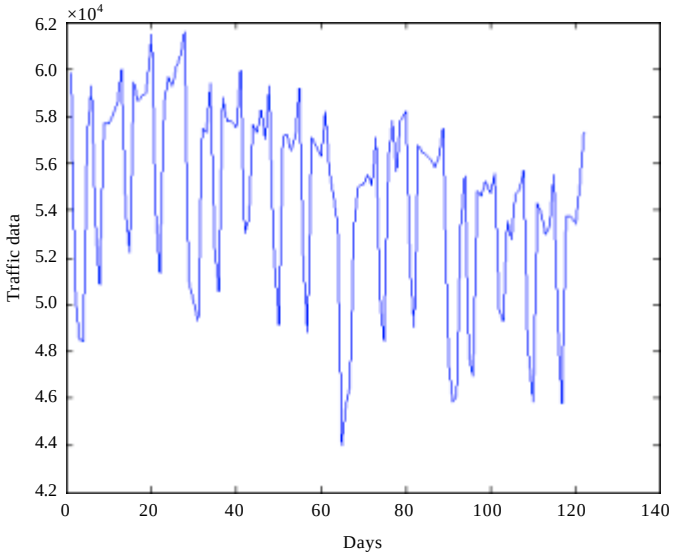

Fig. 2: Traffic data

Table 1: Correlation results

Influencing factor	Correlation coefficient
VLR subscribers	0.55
VLR boot subscribers	0.99
The total number of system response	0.52

Experimental data: The data of telephone traffic and the influencing factors are collected from April 1, 2012 to May 31, 2012 and April 1, 2013 to May 31, 2013, taking the maximum number of traffic in the 24 h a day as telephone traffic data in this study. According to the volume of telephone traffic statistics for 122 days, 122 telephone traffic data is shown in Fig. 2. It is observed that the abscissa is the time parameter from 1-122 days and the ordinate is traffic data of the 122 days. From the figure, the change rule of data at the same time during the two years is similar, so it is feasible to predict the telephone traffic using these collected data.

Considering the telephone traffic is effected by various factors, then correlation analysis is applied to the telephone traffic data to obtain the key factors which influence the telephone traffic and obtain the key factors are VLR subscribers, VLR boot subscribers, the total number of system response and the correlation coefficient of the three factors is shown in Table 1.

From Table 1, we can observe that the correlation coefficient of the key factors obtained is greater than 0.5, that is to say, the correlation between them and the telephone traffic is larger, so we select them as the key factors and the VLR boot subscribers influence the traffic data significantly.

RESULTS

First, telephone traffic data is decomposed with EMD and get each component shown in Fig. 3. The figure shows that each frequency component is gradually lower and the final residual R is the lowest frequency because of the convergence criterion of EMD and it is a monotonic function, it represents the overall trend of the change of the whole telephone traffic data.

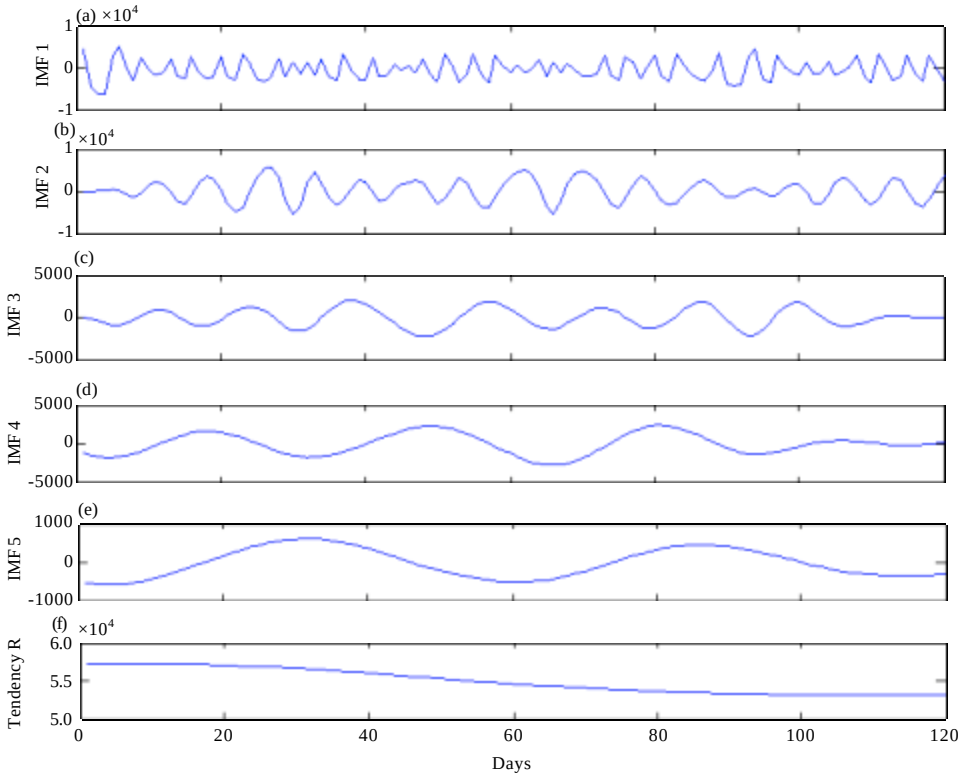

Fig. 3(a-f): Decomposition result of EMD, (a) IMF1, (b) IMF2, (c) IMF3, (d) IMF4, (e) IMF5 and (f) Tendency R

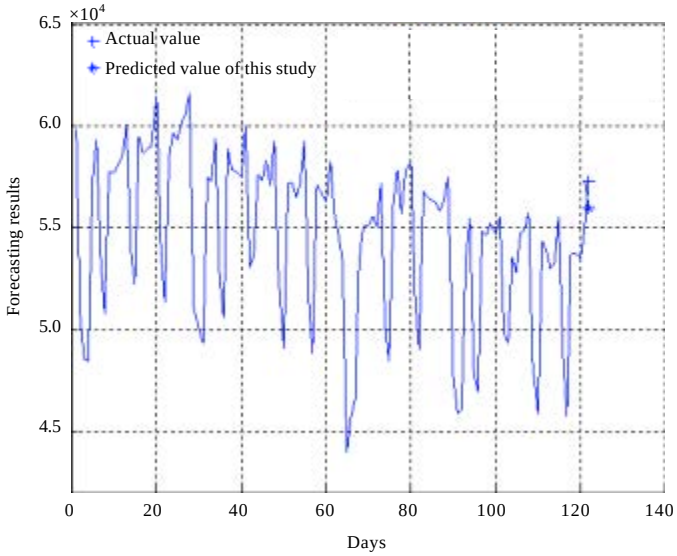

Fig. 4: Forecasting results

In Fig. 3, the telephone traffic data is decomposed into different frequency scale effectively by EMD, so it is easy to dig the internal rules of traffic as well as the analysis of influencing factors, thus it is easy to construct forecasting model of each component in order to improve the precision of the model.

The five IMF component and the three obtained key factors are loaded into Gaussian process model to predict. The last trend component R is loaded into gray prediction GM (1,1) model to predict, because R is telephone traffic trends for a period of time is characteristic of a whole component, so it does not need to consider many factors. Then the six prediction results are superimposed, so it is the predictive results of the algorithm in this study.

The approaches proposed in this study were conducted using code written in the MATLAB and through the simulation of MATLAB, we obtained the final forecasting results. The forecasting metric results are shown in Fig. 4.

In this study, according to the collected data of 122 days, we predict the telephone traffic of the last day based on the front of data. Namely our forecasting result is the traffic data in May 31, 2013. The actual value of traffic in May 31, 2013 is 57323.80, the predicted value of traffic is 55942.30. It is observed that the forecasting effect is good and the error is small. This result is the average of multiple computations, so the result is credible.

DISCUSSION

In order to further verify the predicted effect of forecasting model, the forecasting models such as Least Squares Support Vector Machine (LSSVM) based on EMD and particle swarm optimization (Zhu *et al*., 2013), LSSVM forecasting model based on Wavelet Transform (Li and Xu, 2009) are selected as the contrast models, the scholars have verified the accuracy of the two models already and when using the two models to predict, many factors are also considered in them.

The results of the comparison are shown in Fig. 5, the abscissa is the time parameter and the ordinate is traffic data and in Fig. 5, the result of each model is presented clearly, it shows that each model has errors and we can see the model of this study is the best intuitive from the figure.

The global performance of a forecasting model is evaluated by an accuracy measure, in this study, the Mean Absolute Percentage Error (MAPE) (Suryanarayana *et al*., 2014) is selected to evaluate the predictive effect. The MAPE can be written as:

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{AV - FV}{AV}\right| * 100 \tag{10}$$

In above equations, AV = Actual value, FV = Forecasted value, n = Number of operations.

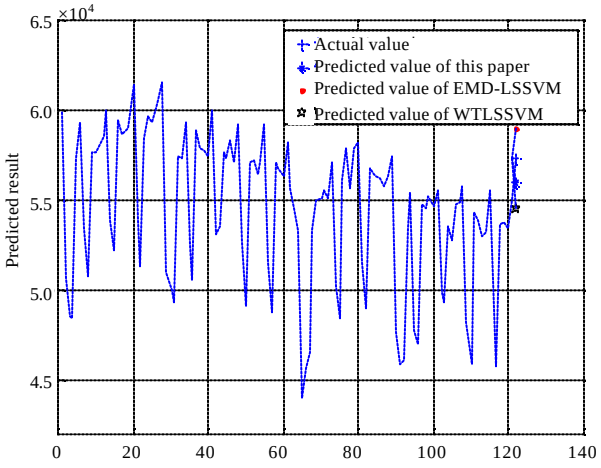

Fig. 5: Contrast figure of predicted results in 3 models

In Table 2, the MAPE of the algorithm proposed in this study is 2.41, the MAPE of the other two algorithms are 2.85 and 4.90. The algorithm proposed in this study stands out as the best forecasting method. The other two algorithms are also have good effect but the algorithm proposed in this study is much better and has higher precision. In addition, the laws of the temperature data (Cortez and Donate, 2014) is similar to the traffic data in this study and the temperature data is shown in Fig. 6. Decomposition Evolutionary Support Vector Machine (DESVM) is mentioned in this reference. In addition, the approach proposed in our study is compared against a recent Evolutionary Artificial Neural Network (EANN) (Donate *et al*., 2013) and the classical forecasting method Autoregressive Integrated Moving Average Model (ARIMA) and their MAPE results are shown in Table 3.

In Table 3, the mean absolute percentage error of EMD-GP-GM is 2.41 and the other three are greater than 3.00, it is visible that the mean absolute percentage error of the algorithm in this study is lower than the other three algorithms, the combination of Gaussian process regression and grey prediction model is better than DESVM, ARIMA and EANN model.

Table 2: Mean absolute percentage error of each model

Model	MAPE (%)
EMD-GP-GM	2.41
EMD-LSSVM	2.85
WTLSSVM	4.90

Table 3: Mean absolute percentage error of other's

Model	MAPE (%)
EMD-GP-GM	2.41
DESVM	3.85
ARIMA	3.42
EANN	3.51

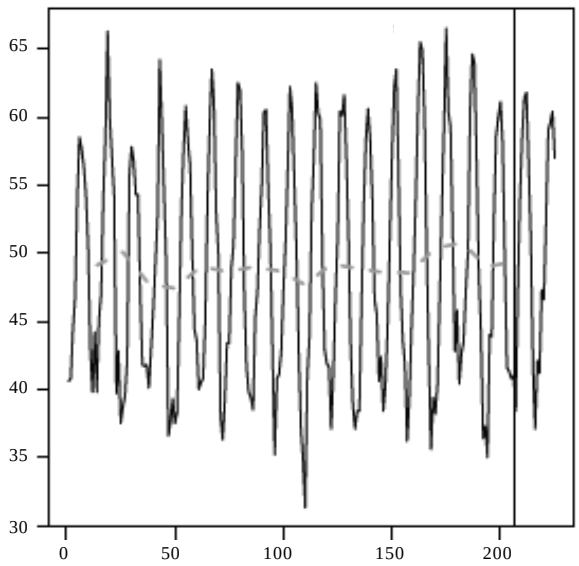

Fig. 6: Temperature data

From what has been discussed above, the model which takes the influence of multiple factors into consideration is better than the model which considers the historical data only. If different frequency component is predicted using different models, the predictive accuracy of this combined forecasting model is higher than using a single model and that the EMD method is more suitable for extracting slow deformation tendency and separating the low frequency signal compared with the wavelet method. The algorithm proposed in this study is more suitable to deal with the telephone traffic prediction.

CONCLUSION

Comparing the predicted results in this study with EMD-LSSVM, WTLSSVM and other scholars' study, the superiority of the predictive effect in the algorithm of this study can be verified and it is a relatively high accuracy forecasting method in predicting the data of the small sample, nonlinear. Good generalization ability and learning ability of the combined model which takes the influence of multiple factors into consideration are proved in this study.

In the application of Gaussian process, the predictive effect can be determined by covariance function and super parameter directly, so the proper choice of covariance function and optimum super parameter is very important, which will become the direction of further research in the future.

ACKNOWLEDGMENTS

This study is supported by the research development fund project (Grant No. XJM 2013-2788) of subsidiary company of China Mobile in Xinjiang. In addition, the authors also thank the editor and the reviewers for their valuable comments that greatly improve the quality of this paper.

REFERENCES

Boran, F.E., 2015. Forecasting natural gas consumption in turkey using grey prediction. Energy Sources Part B: Econ. Plann. Policy, 10: 208-213.

Chandra, D.R., M.S. Kumari, M. Sydulu, F. Grimaccia and M. Mussetta, 2014. Adaptive wavelet neural network based wind speed forecasting studies. J. Electr. Eng. Technol., 9: 742-751.

Cortez, P. and J.P. Donate, 2014. Global and decomposition evolutionary support vector machine approaches for time series forecasting. Neural Comput. Applic., 25: 1053-1062.

Donate, J.P., X. Li, G.G. Sanchez and A.S. de Miguel, 2013. Time series forecasting by evolving artificial neural networks with genetic algorithms, differential evolution and estimation of distribution algorithm. Neur. Comput. Applic., 22: 11-20.

Huang, N.E., Z. Shen, S.R. Long, M.C. Wu and H.H. Shih et al., 1998. The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. Proc. R. Soc. London A, 454: 903-995.

Li, X. and J.J. Xu, 2009. Landslide deformation prediction based on wavelet analysis and least square support vector machine. J. Geod. Geodyn., 29: 127-130.

Reggente, M., J. Peters, J. Theunis, M. van Poppel, M. Rademaker, P. Kumar and B. de Baets, 2014. Prediction of ultrafine particle number concentrations in urban environments by means of Gaussian process regression based on measurements of oxides of nitrogen. Environ. Modell. Software, 61: 135-150.

Salcedo-Sanz, S., C. Casanova-Mateo, J. Munoz-Mari and G. Camps-Valls, 2014. Prediction of daily global solar irradiation using temporal Gaussian processes. IEEE Geosci. Remote Sens. Lett., 11: 1936-1940.

Sudheer, G. and A. Suseelatha, 2015. Short term load forecasting using wavelet transform combined with-Holt-Winters and weighted nearest neighbor models. Int. J. Electr. Power Energy Syst., 64: 340-346.

Suryanarayana, C., C. Sudheer, V. Mahammood and B.K. Panigrahi, 2014. An integrated wavelet-support vector machine for groundwater level prediction in Visakhapatnam, India. Neurocomputing, 145: 324-335.

Wang, J., W. Zhang, Y. Li, J. Wang and Z. Dang, 2014. Forecasting wind speed using empirical mode decomposition and Elman neural network. Applied Soft Comput., 23: 452-459.

Wu, J., J. Long and M. Liu, 2015. Evolving RBF neural networks for rainfall prediction using hybrid particle swarm optimization and genetic algorithm. Neurocomputing, 148: 136-142.

Yu, S., B. Li, Q. Zhang, C. Liu and M.Q.H. Meng, 2015. A novel license plate location method based on wavelet transform and EMD analysis. Pattern Recognit., 48: 114-125.

Zhu, Q.Y., X.Z. Qin, Z.H. Jia, L. Sheng and L. Chen, 2013. Network traffic prediction based on EMD and partcle swarm optimization of LS-SVM. Comput. Eng. Des., 34: 4104-4108.