Indian Journal of Extension Education
Vol. 56, No. 2, 2020 (7-12)

Empirical Mode Decomposition based Support Vector Regression for Agricultural Price Forecasting

Pankaj Das[1], Girish Kumar Jha[2], Achal Lama[3], Rajender Parsad[4] and Dwijesh Mishra[5]

ABSTRACT

Price information is a crucial market information for a farmer. The price instability and uncertainty pose a significant challenge to decision makers in making proper production and marketing plans to minimize risk. Agricultural price series cannot be modelled and predicted accurately by traditional econometric models owing to its nonlinearity and nonstationary behaviour. In the present study an attempt has been made to model and predict price series using Empirical Mode Decomposition (EMD) based Support Vector Regression (SVR) model. EMD decomposes the original nonlinear and nonstationary dataset into a finite and small number of sub-signals. Then each sub-signal was modelled and forecasted by SVR method. Finally, all the forecasted values of sub-signal were aggregated to make final ensemble forecast. The effectiveness and predictability of the proposed methodology was verified using Chilli wholesale price index (WPI) dataset as sample. The results indicated that the performance of the proposed model was substantially superior as compared to the standard SVR.

Key words: Agricultural price forecasting, empirical mode decomposition, nonlinearity, nonstationary support vector regression.

INTRODUCTION

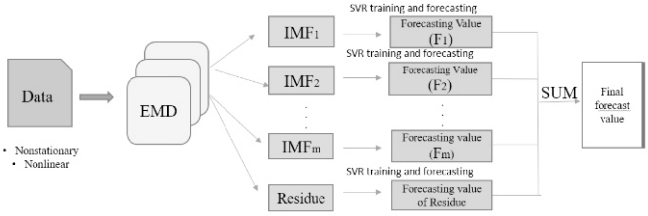
The agricultural market environment is changing with unprecedented speed both locally and globally. The dynamic nature of market affects farm prices and thereby farm income. Most of the rural farmers are unable to understand and interpret the market and price behaviour to their advantages (Anjaly *et al.,* 2010). Thus, market information and intelligence are crucial to enable the farmers and traders in making important decisions about what to grow, when to sell, and where to sell. Besides this, the price instability and uncertainty pose a restriction on decision and policy makers. Hence, agricultural price forecasting plays a vital role for both production and market strategy. Price forecasting controls the supply and demand of the commodity. Price forecasting of an agricultural product is a herculean task because it depends on too many factors which cannot be accurately predicted. Nonlinearity and nonstationary behaviour are crucial problems in agricultural price data. Traditional nonlinear time series models like Autoregressive Conditional Heteroscedastic (ARCH) models (Engle, 1982), Generalised ARCH (GARCH) model (Bollerslev, 1986) etc. fails to model the agricultural price series due to its inherent nonlinear and nonstationary characteristic. To overcome the problem, Machine learning techniques namely Artificial Neural Networks (ANNs) and Support Vector Machines (SVMs), gained significant popularity for economic time series forecasting by making data driven predictions or decisions through building a model from sample inputs. SVM has generalization capacity to obtain a unique solution (Lu *et al,* 2009). The structural risk minimization principle of SVM enhances the model performance (Duan and Stanley, 2011). SVR models

[1, 3 & 5] Scientists and [4.] Director, ICAR-Indian Agricultural Statistics Research Institute, New Delhi-12, [2.] Principal Scientist, Division of Agricultural Economics, ICAR-IARI, New Delhi-110012

require long time to train large dataset. To overcome the problem, Suykens and Vandewalle (1999) proposed least squares support vector regression (LSSVR) to transform inequality constraints into equality constraints by employing a squared loss function. Machine learning methods are usually based on the assumption that the data generation mechanism does not change over time. Dealing with non-stationarity is one of modern machine learning's greatest challenges (Sugiyama and Kawanabe, 2012).

In literature several researchers have applied different hybrid methodologies to tackle the problem of nonlinearity and nonstationary. Huang *et al.* (1998) highlighted the main advantage of the EMD *i.e.* the lack of initial assumptions on the dataset *i.e.* stationarity or linearity and not use a priori determined basis functions. Zhang (2003) proposed a hybrid ARIMA and neural network (NN) model for time series forecasting. They concluded that a hybrid methodology has advantage of the unique strength of Autoregressive Integrated Moving Average (ARIMA) and NN models in linear and nonlinear modelling. Ince and Trafails (2006) proposed a hybrid model based on Autoregressive Integrated Moving Average (ARIMA) and Support vector regression (SVR) in order to improve forecasting accuracy. The proposed methodology outperformed the logit/probit models. Chen (2007) demonstrated that superiority of Support Vector Regression (SVR) over the NN and Maximum Likelihood estimation (MLE). Brandl *et al.* (2009) used genetic algorithm for variable selection and set their model using SVR methodology. The proposed model outperformed a NN, an Ordinary Least Squares (OLS) regression and ARIMA model. An *et al.* (2012) reported that EMD can reveal the hidden pattern and trends of time series which can effectively assist in designing forecasting models for various applications. Guo *et al.* (2012) decomposed wind speed series using EMD and forecasted them using a feed-forward network. Chen *et al.* (2012) proposed an EMD approach combined with an NN model for tourism-demand forecasting. Lama *et al.* (2016) explored the superiority of GARCH based Time Delay Neural networks (TDNN) for forecasting agricultural commodity price volatility.

It is almost universally agreed in the forecasting literature that no single method is best in every situation. This is largely since a real-world problem is often complex in nature and any single model may not be able to capture different patterns equally well. This has motivated to develop an ensemble model *i.e.* combination of time series model and machine learning technique which deals with both linear and nonlinear pattern and improve forecasting accuracy. In this present study, EMD-SVR hybrid approach has been proposed. In this method, EMD was used for decomposing the nonlinear and nonstationary series into finite and small numbers of sub-signals. Then these sub-signals were individually modelled and forecasted using SVR technique. Finally, all forecasted values of sub-signals were aggregated to make final ensemble forecast. This proposed hybrid model results in improved forecasting efficiency as compared to individual model



SUM= $\sum_{k=1}^{m} F_k + forecasted\ value\ of\ final\ residue$

Figure 1: Work flow of proposed EMD-SVR ensemble learning paradigm

METHODOLOGY

Ensemble method is a machine learning approach which combined multiple base model to produce an optimal predictive model. The proposed EMD-SVR consists of three steps defined in figure 1. First step, original nonlinear and nonstationary dataset is decomposed into a finite and often small numbers of independent sub-series by EMD technique. This sub-series contain m intrinsic mode functions (IMF) and a final residue. Secondly, these IMFs and residue are modelled and predicted through SVR. Then, all the forecasted values of the IMFs and residue are summed up to produce ensemble forecast for the original series (Huang et al., 1998).

Empirical mode decomposition (EMD)

The Empirical Mode Decomposition method was introduced by Huang *et al.,* in 1998. It assumes that the data have many coexisting oscillatory modes of significantly distinct frequencies and these modes superimpose on each other and form an observable time series. EMD decomposes original non-stationary and nonlinear data into a finite and small number of independent sub-series (including intrinsic mode functions and a final residue). Intrinsic mode function (IMF) is the finite additive oscillatory component decomposed by EMD. For example, let x(t) is a dataset consisting high frequency part and low frequency part.

Data = fast oscillations superimposed to slow oscillations

$$x(t) = d_1(t) + r_1(t) \quad \ldots (2.1.1)$$

where $d_1(t)$ = high frequency part *i.e.* IMF and $r_1(t)$ = low frequency part.

EMD algorithm iterate on the slow oscillation component considered as a new signal.

$$r_1(t) = d_2(t) + r_2(t)$$

After full decomposition, $x(t) = \sum_{i=1}^{m} d_i(t) + r_i(t)$ ....(2.1.2)

Data = sum of IMFs + final residue.

Stepwise EMD algorithm procedure is mentioned below:

Step 1: Identify all extrema of $x(t)$

Step 2: Interpolate the local maxima to form an upper envelope $u(x)$

Step 3: Interpolate the local minima to form a lower envelope $l(x)$

Step 4: Calculate the mean envelope: $m(t)=[u(x)+l(x)]/2$

Step 5: Extract the mean from the signal: $h(t)=x(t)-m(t)$

Step 6: Check whether h(t) satisfies the IMF condition.

YES: $h(t)$ is an IMF, stop shifting.

NO: let $x(t)=h(t)$, keep shifting.

Support vector regression (SVR) model

Support Vector Machine (SVM) proposed by Vapnik (1998), is nonlinear algorithms used in supervised learning frameworks for data analysis and pattern recognition. Vapnik (1998) proposed introduced support vector regression (SVR) model by incorporating loss function. SVR maps input vectors into a high dimensional space and then run linear regression in the outer space. The model has been built in two steps *i.e.* the training and the testing step. In the training step, the largest part of the dataset has been used for the estimation of the function. In the testing step, the generalization ability of the model has been evaluated by checking the model's performance in the small subset that was left aside during training.

For a given data set $\{(x_1, y_1),...,(x_n, y_n)\}(x_i \in R^K, y_i \in R^1)$, SVR maps the original data into a higher or infinite dimensional space by nonlinear function φ, then seeks mapping function. $\phi : R^K \to R^1$

The general formula for linear support vector regression is given as

$$y = \phi(x) = W\varphi(x) + b \quad \ldots (2.2.1)$$

where W defines weight vector, φ denotes mapping function and b is bias. LSSVR is the least square version SVR where set of linear equations are used to find the solution.

The solution of W and b in above equation can be obtained by solving the following minimization problem (Sermpinis *et al.,* 2014)

$$\min_{W,b,\zeta,\zeta^*} \frac{1}{2}\|W\|^2 + C\sum_{i=1}^{n}(\zeta + \zeta^*) \quad \ldots (2.2.2)$$

such that $W^T\varphi(x_i) + b - y_i \le \varepsilon + \zeta_i$; $y_i - W^T\varphi(x_i) - b \le \varepsilon + \zeta_i^*$

$\zeta, \zeta^* \ge 0, i = 1,..., N$.

It is a primal function and solution of the function is quite complex in nature. So, the dual of the function can be used. Its dual will be

$$\min_{\alpha,\alpha^*} \frac{1}{2}\sum_{i,j}^{N}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)Ker(x_i, x_j) + \varepsilon\sum_{i}^{N}(\alpha_i - \alpha_i^*) + \sum_{i}^{N} y_i(\alpha_i - \alpha_i^*) \quad \ldots (2.2.3)$$

$$\text{s.t. } \sum_{i}^{N}(\alpha_i - \alpha_i^*) = 0 \text{ and } 0 \le \alpha, \alpha^* \le C, i = 1,..., N$$

where $Ker(x_i, x_j) = \varphi(x_i)^T\varphi(x_j)$ is kernel function. For getting estimated value*,αα the dual function will be used. Thus, the coefficient b will be calculated as

$$\tilde{b} = y_j - \sum_{i}^{N}(\alpha_i - \alpha_i^*)Ker(x_i, x_j) - \varepsilon.; \tilde{\alpha}_i \in (o, C) \quad \ldots (2.2.4)$$

$$\tilde{b} = y_j - \sum_{i}^{N}(\alpha_i - \alpha_i^*)Ker(x_i, x_j) + \varepsilon. \; ; \; \tilde{\alpha}_i^* \in (o, C) \quad \ldots (2.2.)$$

In the present study, Ɛ-SVR, specialised form of least squares SVR model was used. Radial basis kernel function (RBF) was used for nonlinear mapping of dataset.

RESULTS ADN DISCUSSION

Dataset description

In the present study, monthly Chilli Wholesale price index (WPI) dataset was used to evaluate the performance of proposed EMD-SVR model. The dataset was obtained from the Office of the Economic Advisor, Ministry of Commerce, Government of India. Figure 2 illustrated the monthly data of Chilli WPI from (April, 1994 to May, 2018) contained 290 data points with base year 2005. The descriptive statistics of data, stationarity test and normality were presented in Table 1. The statistics obtained through augmented *Dickey-Fuller* (ADF) and *Phillips-Perron* (PP) test were insignificant *i.e.* null hypothsis of unit root test cannot refused. It indicated that the given dataset was nonstationary. *Jarque-Bera* test (Table 1) also indicated the nonnormality of data.

Table 1: The descriptive statistics of data, stationarity test and Normality test

Observations	Minimum	Maximum	Mean	Standard deviation	Skewness	Kurtosis
290	136.5	971.3	412.8	201.415	1.042	3.252
Augmented Dickey-Fuller Test (p value)			Phillips-Perron Test (p value)		Jarque-Bera test (p value)	
0.924			0.821		2.772e-12	

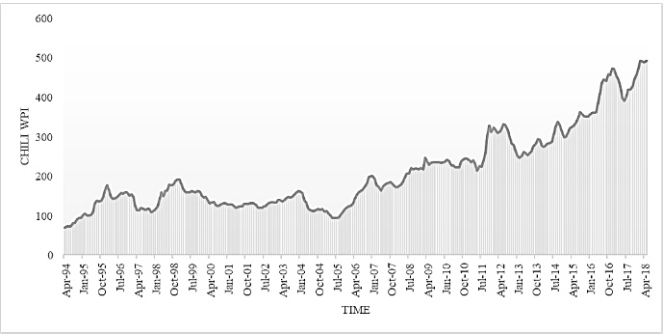


Figure 2: Time plot of monthly Chilli WPI

Brock-Dechert-Scheinkman (Brock *et al.,* 1996) test was used in the dataset for checking nonlinearity of data. The results of BDS test (Table 2) described that the test statistics were far bigger than the critical values. It provided an evidence to reject null hypothesis that the price series is linearly dependent. Therefore, the monthly Chilli WPI dataset was nonlinear and nonstationary.

Table 2: Brock- Dechert-Scheinkman (BDS) test

Embedding dimension				Conclusion
2		3		
Statistics	Probability	Statistics	Probability	Nonlinear
66.089	< 0.001	106.523	< 0.001	
51.709	< 0.001	62.327	< 0.001	
40.525	< 0.001	42.372	< 0.001	
35.129	< 0.001	34.132	< 0.001	

EMD-SVR Training and forecasting

The whole analysis was done in RStudio. The packages "EMD" (Kim *et al.,* 2009) and "e1071" (Meyer *et al.,* 2018) were used for EMD and SVR fitting respectively. Firstly, whole original dataset has been decomposed into 4 IMFs and one final residue, illustrated in figure 3. It has been observed that the frequencies and amplitudes of IMFs were different from each other.

Thus, the different hidden oscillatory modes in original dataset were separated by EMD. Unit root test was also done to check stationarity of IMFs and residue. Table 3 described the results of the test. IMF 4 and residue were nonstationary. They were transformed into stationary by differencing. Each individual component (IMFs and residue) was modelled and forecasted by SVR model. 80 per cent data were used as training and remaining 20 per cent used as testing set Then all forecasted values of IMFs and residue were summed up to get an ensemble forecast of the data.

Table 3: Unit root test of decomposed components of chilli dataset

Phillips-Perron Test (p value)	Augmented Dickey-Fuller Test (p value)	Remarks
<0.01	<0.01	Stationary
<0.01	<0.01	Stationary
<0.01	<0.01	Stationary
0.559	0.260	Non-stationary
0.99	0.985	Non-stationary

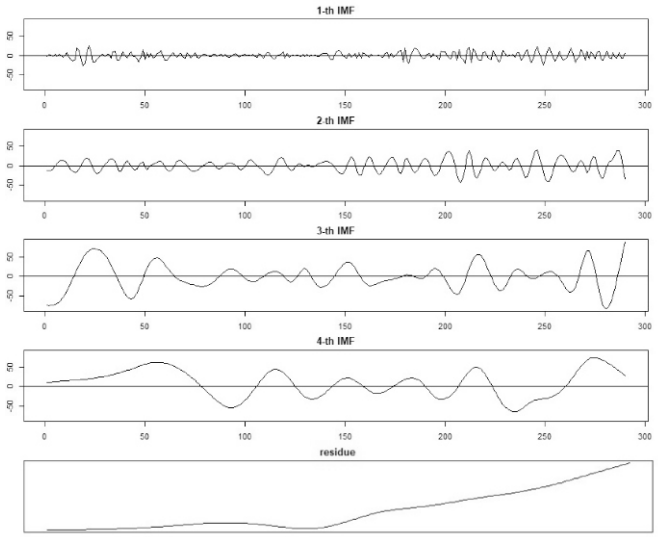


Figure 3: Decomposed components of monthly Chilli WPI

Iterative 8-step and 6-step ahead prediction was utilized in order to measure out-of-sample predictability of EMD-SVR model. The model predicted one-step ahead each time and for the next time step prediction added the current output. In the study, Ɛ-SVR with Radial Basis Function (RBF) as kernel function was employed. The optimal parameter combination was fixed using grid search method. 10-fold cross validation was done to overcome outfitting problem. Performance of the prediction was compared by Root Mean Square errors (RMSE), Mean Absolute Deviation (MAD), Mean Absolute Percentage Error (MAPE) and Maximum Error (ME). The formulas are given below

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - F_i)^2}{N}} \qquad MAD = \frac{\sum_{i=1}^{N}|y_i - F_i|}{N}$$

$$MAPE = \frac{\sum_{i=1}^{N}|y_i - F_i| / y_i}{N} \qquad ME = \max \sum_{i=1}^{N}|y_i - F_i|$$

where in above equations y_t and F_i are the i^{th} actual value and forecasted value of response variable, i= 1, …, N. N= 6 and 8. For comparison 6-step and 8-step ahead prediction was done using standard SVR. The forecasting performance of the models described in Table 4.

\ Table 4 clearly exihibit that the forecasting accuracy of the EMD-SVR model was superior compared to the standard SVR model. Even the performance of the EMD-SVR model did not limited to the forecasting period. This result strongly recommended that one can improve the accuracy of SVR by incorporating EMD. The accuracy of 8-step EMD-SVR forecasting was slightly lower than 6-step EMD-SVR forecasting.

Table 4: Forecasting performance of SVR and EMD-SVR model

Method	Forecast period	RMSE	MAPE	MAD	ME
EMD-SVR	F_1-F_6	27.228	0.021	19.627	58.519
	F_1-F_8	37.483	0.033	31.420	68.734
SVR	F_1-F_6	73.373	0.078	72.580	69.984
	F_1-F_8	78.918	0.050	46.041	116.450

CONCLUSION

The present market environment is changing abruptly, and this dynamic nature affects both the farmer and traders. Major portion of them do not understand the market and price behaviour in advance. Therefore, price forecasting plays a vital role for making efficient decisions. Perishability is another problem for the agricultural product like Chilli. Price information in advance helps the farmers and traders from market loss.

In the present study, a new ensemble technique "EMD-SVR" has been proposed. The novelty of the ensemble approach is that it can handle nonlinear and nonstationary data which was unsuitable for the traditional time series methods. Firstly, the proposed EMD-SVR model decomposed the original nonlinear and nonstationary data series into IMFs and residue. Each of decomposed components (IMFs and residue) describes distinct frequencies and amplitudes.

Then each individual component (IMFs and residue) has been forecasted by Ɛ-SVR modelling. Finally, ensembled forecasted value of EMD-SVR was obtained by summing up all individual forecasted values of IMFs and residues. The effectiveness and forecasting ability of proposed EMD-SVR was verified by using Chilli WPI dataset. Thus, it can be concluded that the proposed EMD-SVR model may be an effective tool for agricultural price forecasting.

Paper received on : November 05, 2019
Accepted on : February 13, 2020

REFERENCES

An, X., Jiang, D., Zhao, M. and Liu, C. (2012). Short time prediction of wind power using EMD and chaotic theory. Communication in Nonlinear Science and Numerical Simulation, 17(2), 1036-1042.

Anjaly, K. N., Surendran, S., Babu, S. K. and Thomas, J. K. (2010). Impact Assessment of Price Forecast: A Study of Cardamom Price Forecast by AMIC, KAU. NAIP on Establishing and Networking of Agricultural Market Intelligence Centres in India. College of Horticulture, Vellanikkara. 31.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroscedasticity. *Journal of Econometrics,* 31, 307-327.

Brandl, B, Wildburger, U. and Pickl, S. (2009). Increasing of the fitness of fundamental exchange rate forecast models. *International Journal of Contemporary Mathematical Sciences,* 4(16), 779-798.

Brock, W. A., Scheinkman. J. A., Dechert, W. D. and LeBaron, B. (1996). A test for independence based on the correlation dimension. *Econometric Reviews.* 15, 197 235.

Chen, C. F., Lai, M. and Yeh, C. C. (2012). Forecasting tourism demand based on empirical mode decomposition. Knowledge-Based Systems, 26, 281-287.

Chen. S. Y. (2007). Forecasting exchange rates: a new nonparametric support vector regression. *The Journal of Quantitative & Technical Economics,* 5, 142-150.

Dickey, D. and Fuller, W. (1981). Likelihood ratio statistics for autoregressive time series with a unit root. Econometrica, 49, 1057–1072.

Duan, W. Q., and Stanley, H. E. (2011). Cross-correlation and predictability of financial return series. Physica A, 390(2), 290-296.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of U.K. inflation. Econometrica, 50, 987-1008.

Guo, Z., Zhao, W., Lu, H. and Wang, J. (2012). Multi-step forecasting for wind speed using a modified EMD-based

artificial neural network model. Renewable Energy, 37(1), 241-249.

Huang, N. E., Shen, Z., Long, S. R., Wu, M. L., Shih, H. H., Zheng, Q., Yen, N. C., Tung, C. C., and Liu, H. H. (1998). The empirical mode decomposition and Hilbert spectrum for nonlinear and non stationary time series analysis. Proceeding of the Royal Society London A, 454, 909-995.

Ince, H. and Trafalis, T. (2006). A hybrid model for exchange rate prediction. Decision Support Systems, 42(2), 1054-1062.

Lama, A., Jha, G., Gurung, B., Paul, R. K., Bharadwaj, A. and Parsad, R. (2016). A Comparative Study on Time-delay Neural Network and GARCH Models for Forecasting Agricultural Commodity Price Volatility. *Journal of the Indian Society of Agricultural Statistics,* 70(1), 7-18.

Lu, C. J., Lee, T. S. and Chiu, C. C. (2009). Financial time series forecasting using independent component analysis and support vector machine. *Decision Support Systems*, 47(2), 115-125.

Kim, D. and Oh, H-S. (2009). EMD: a package for empirical mode decomposition and Hilbert spectrum. *The R Journal,* 1.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C. C. and Lin, C. C. (2018). e1071: a package for latent class analysis, short time Fourier transform, fuzzy clustering, support vector machines, shortest path computation, bagged clustering and naive Bayes classifier. *The R Journal.*

Phillips, P. C. B. and Perron, P. (1988). Testing for a unit root in time series regression. Biometrika, 75, 335-346.

Sermpinis, G., Stasinakis, C., Theofilatos, K. and Karath-anasopoulos, A. (2014). Inflation and unemployment forecasting with genetic support vector regression. *Journal of Forecasting,* 33, 471-487.

Sugiyama, M. and Kawanabe, M. (2012). Machine Learning in Non-Stationary Environments- Introduction to Covariate Shift Adaptation. The MIT Press, Cambridge, Massachusetts, London, England. 2nd ed.

Suykens, J. A. K. and Vandewalle, J. (1999). Least squares support vector machine classifier. Neural Processing Letters, 9(3), 293-300.

Vladimir, N. Vapnik. (1998). Statistical Learning Theory. Wiley-Interscience. 1st ed.

Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. Neur-ocomputing, 50, 159-175.