Forecasting Average Rice Prices at Milling Level According to Quality Using Support Vector Regression

Fauziah Roshafara

Department of Statistics, Faculty of Mathematics and Natural Science, Universitas Islam Bandung

Jln Ranggagading No.8, Tamansari, Kecamatan Bandung Wetan, Kota Bandung, Jawa Barat
e-mail: fauziahroshafara@unisba.ac.id

ABSTRAK

Indonesia merupakan negara agraris yang mayoritas penduduknya berprofesi sebagai petani dan salah satu komoditas besar yang dihasilkan sekaligus sebagai makan utaman adalah beras. Harga beras yang fluktuatif akan berpengaruh terhadap daya beli Masyarakat, sehingga perlu upaya untuk menyiapkan kebijakan peningkatan daya beli masyarakat terhadap beras melalui peramalan. Penelitian ini menggunakan SVR untuk pemodelan harga rata-rata beras menggunakan 114 dataset yang diperoleh dari Januari 2013 hingga Juni 2023, kemudian mengevaluasi kinerjanya menggunakan *Mean Absoute Percetage Error* (MAPE). Model terbaik terbentuk dari kernel linier dengan parameter $\varepsilon = 0.078$ dan $C = 3.1$. Model tersebut menghasilkan nilai MAPE terkecil yaitu 2.32% pada data pengujian dan 1,2% pada data pelatihan yang juga kurang dari 10% artinya kinerja model untuk meramalkan harga rata-rata beras sangat tinggi.
Kata kunci: peramalan; *support vector regression*; rata-rata harga beras

ABSTRACT

Indonesia is an agricultural country where the majority of the population work as farmers and one of the humongous commodities produced is rice. Rice is a very important commodity for the Indonesian people, because it is the main food of them. This is why rice production in Indonesia is the big concern to the government, including of the average rice prices at milling level. The fluctuative of the rice prices will be affect to the purchasing power of the people. One of the efforts that can be made to prepare a policy to increase people's purchasing power of the rice is by forecasting. This study used SVR to modeling the average rice prices using 114 datasets obtained from January 2013 to June 2023, then evaluating its performance using Mean Absoute Percetage Error (MAPE). The best model formed from a linear kernel with parameters $\varepsilon = 0.078$ *and* $C = 3.1$*. The model produced the smallest MAPE value of 2.32% in testing data and 1.2% in training data which also less than 10% meaning that the performance of the model to forecast the average price of rice is very high.*
Keywords: *Forecasting; Support Vector Regression; Average Rice Prices*

INTRODUCTION

Indonesia is an agricultural country where the majority of the population work as farmers and one of the humongous commodities produced is rice. Rice is a very important commodity for the Indonesian people, because it is the main food of them. According to United States Department of Agriculture (USDA), Indonesia is the top fourth rice consuming contry after China, India and Bangladesh. From the same source also Indonesia is the top third rice producing countries from 2017 to 2022. This is why rice production in Indonesia is the big concern to the government, including of the average rice prices at milling level. The fluctuative of the rice prices will be affect to the purchasing power of the people [13]. One of the efforts that can be made to prepare a policy to increase people's purchasing power of the rice is by forecasting [4].

Forecasting is an activity to predict future events using time series data [10]. The basic core of the forecasting process is to predict future events based on patterns of events that have occurred in the past. The purpose of forecasting is as a source of information about conditions that will occur, so the suitable action can be taken in dealing with these conditions. One of the time series models that often used for forecasting is the autoregressive integrated moving average (ARIMA) [3]. However there is requirement that need to be fulfilled, the data have to stationary to the mean and variance [17]. If the time series are non-stationary, then it needs to be transformed first to be able to obtained the model used for forecasting.

In recent years, there has been growing interest in using machine learning techniques to forecast in many sectors. The advantage of the machine learning model with other conventional time series models is that it can be applied to both linear and non-linear data. In the tourism and hospitality sector machine learning models can be used for forecasting international tourist arrivals and allso the daily room rates [16], [2]. Artificial Neural Network (ANN) can also be applied in predicting active family planning participants through government channel [6]. Another machine learning models that can be used for forecasting is support vector regression (SVR). There are several studies that use SVR for modelling and predicting wetland rice production [11], [1] which the data is non-linear patterns. On the studies about forecasting electricity consumption comparing Neural Networks and Support Vector Regression by Oğcu, et al shows that the SVR give the better performance in the forecasting accuracy than the NN [11].

This study aims to forecast the average price of rice using 114 datasets obtained from January 2013 to June 2023 based on BPS (Badan Pusat Statistik) official website. The data indicated that it was non stationary to the average due to seasonal effects. Therefore, SVR is more suitable to used to overcome this problem where the data does not need to be transformed first to produce a model used for forecasting. After obtaining the results by the optimum parameter, the performance is then evaluated using the Mean Absoute Percentage Error (MAPE).

METHOD

Support Vector Regression is the application of machine learning concepts to classify regression models using the Support Vector Machine method [9]. The goal of SVR is to find the $f(y)$ function as a hyperplane in the form of a regression function that corresponds to all input data by making the smallest possible error (ε) [14]. Suppose there is a $y(t)$ time series datasets, where $t = \{0,1,2,\dots,N-1\}$. Then by using regression analysis the prediction function in linear and non-linear regression can be written as follows:

$$\widehat{Y_t} = (\boldsymbol{w} \cdot y) + b \quad (1)$$

$$\widehat{Y_t} = \left(\boldsymbol{w} \cdot \varphi(y)\right) + b \quad (2)$$

where $\boldsymbol{w}$ is the weighting vector, $\varphi(y)$ is a function that maps to a higher dimensional space and b is biased. Then, to maximize the hyperplane can be done by minimizing the regularized risk function defined by equation (3).

$$R_{reg}(f) = R_{emp}(f) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2 \tag{3}$$

where,

$$R_{emp}(f) = \frac{1}{N}\sum_{t=1}^{N} L(\gamma(t) - f(y(t), \boldsymbol{w})) \tag{4}$$

with λ is a constant used to reduce overfitting of the data and minimize the adverse effects of generalization, i is the time series index $t = \{0,1,2,\dots,N-1\}$, $\gamma(t)$ is the actual data of the predicted value that to be found and $L(.)$ is the loss function [15].

Loss function is a function that shows the relationship between errors and how errors are penalized [5]. The most commonly used of loss function is the ε-insensitive loss function. To obtain the optimal weight by minimizing regularized risk, a quadratic equation is formed using the ε-insensitive loss function as follows:

$$\min\left\{\frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{t=1}^{N} L(\gamma(t) - f(y(t), \boldsymbol{w}))\right\} \tag{5}$$

where

$$L(\gamma(t), f(y(t), \boldsymbol{w})) = \begin{cases} |\gamma(t) - f(y(t), \boldsymbol{w})| - \varepsilon, & |\gamma(t) - f(y(t), \boldsymbol{w})| \geq \epsilon \\ 0 \qquad , & other. \end{cases} \tag{6}$$

The C constant is a normalizing factor of sum by $1/N$ and ε is a measure of the precision used to approximate the function. Solving for optimal weights and refractive values uses the Lagrange multiplier and kernel function, so the prediction function of the regression can be explicitly written as follows:

$$\widehat{Y}_t = \sum_{t=1}^{N}(\alpha_t - \alpha_t^*)\langle y, y(t)\rangle + b \tag{7}$$

where α_i is a Lagrange multiplier and $\langle y, y(t)\rangle$ is a kernel function. To perform non-linear regression using SVR, it is necessary to map the $y(t)$ input space into higher-dimensional features $\varphi(y(t))$. Then a kernel function that satisfies *Mercer's condition* can be generated as:

$$k(y, y') = \langle \varphi(y), \varphi(y')\rangle \tag{8}$$

There are several kernel functions that can be used to solve (8) equation, namely:

1. Linear Kernel Function: $K(y_i, y) = y^T y$
2. Polynomial Kernel Function: $K(y_i, y) = (y^T y + 1)^d$
3. Radial Basis Fuction *(RBF)* Kernel Function: $K(y_i, y) = exp\left(-\frac{\|y_i - y\|^2}{2\gamma^2}\right)$

While a method that can be used to evaluate the effectiveness of a forecast is the Mean Absoute Percetage Error (MAPE). Error percentage measurement has the advantage of being scale independent and is often used to compare the performance of a forecast on different datasets [8]. Performance measurement using MAPE produces a percentage value. The smaller the MAPE value, the better the level of accuracy of a forecast. The calculation of MAPE can be written as follows:

$$MAPE = \frac{1}{N}\sum_{t=1}^{N}\left|\frac{Y_t - \hat{Y}_t}{Y_t}\right| \times 100\% \tag{9}$$

Y_t is actual value of sales in the t-month, $\hat{Y}_t$ is value of sales forecasting results in the t-month and N is an amount of data. Based on Lewis [8], MAPE values can be interpreted in four categories, highly accurate (<10%), good (10 – 20%), reasonable (20 – 50%) and not accurate (>50%).

RESULT AND DISCUSSION

According from BPS (Badan Pusat Statistik) official website the average rice prices at milling level differ based on its quality, respectively premium, medium and low quality. This research use the average rice prices at milling level with the medium quality obtained from January 2013 to June 2023, and then forecast for the next six months using support vector regression.

The time series plot of the average rice praces can be shown in Figure 1. It can be seen that the data in non-stasionary. There is an increasing trend and also the variance of the data is not constant meaning the regular time series methods that needs stationary of the mean and variance of the data can not be used without making transformation and differencing to remove an unconstant variance and trend. Therefore support vector regression can be used as an alternative method at the data with those problem.

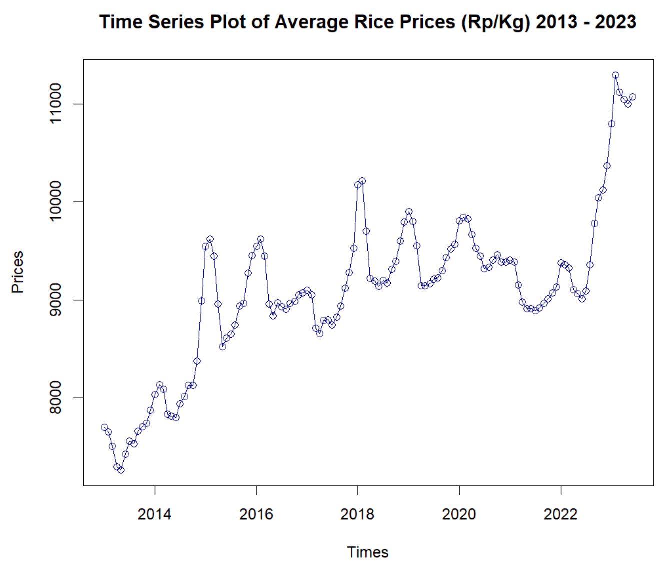


Figure 1. The Series Plot of The Average Rice Prices at Milling Levels

The time series plot of the average rice prices from Figure 1 is not clear whether there is a seasonality or not. To reveal this issue, needs to be plotted by the periods of 12 months in Figure 2. It is very clear that there is seasonality in the average of rice prices which almost every year almost have the same pattern of price shifting in the each of months.

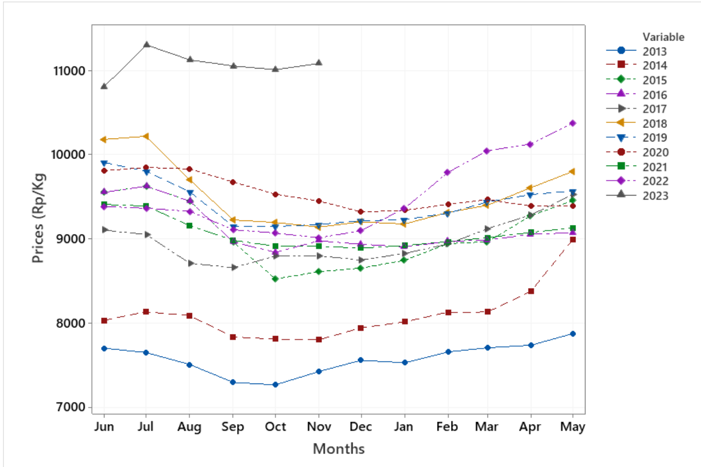


Figure 2. The Average Rice Prices of Each Months

The traditional univariate time series methods for forecasting will only use the average rice prices level as the variable. However in the literature of the forecasting using SVR always used multivariate data where the average rice prices as dependent varible and other variables related to it as independent variables. Here, in this research, total three inputs were used where all derived from the average rice prices data as output. These inputs consist of 'lag 1', 'lag 2', and 'seasonal index'.

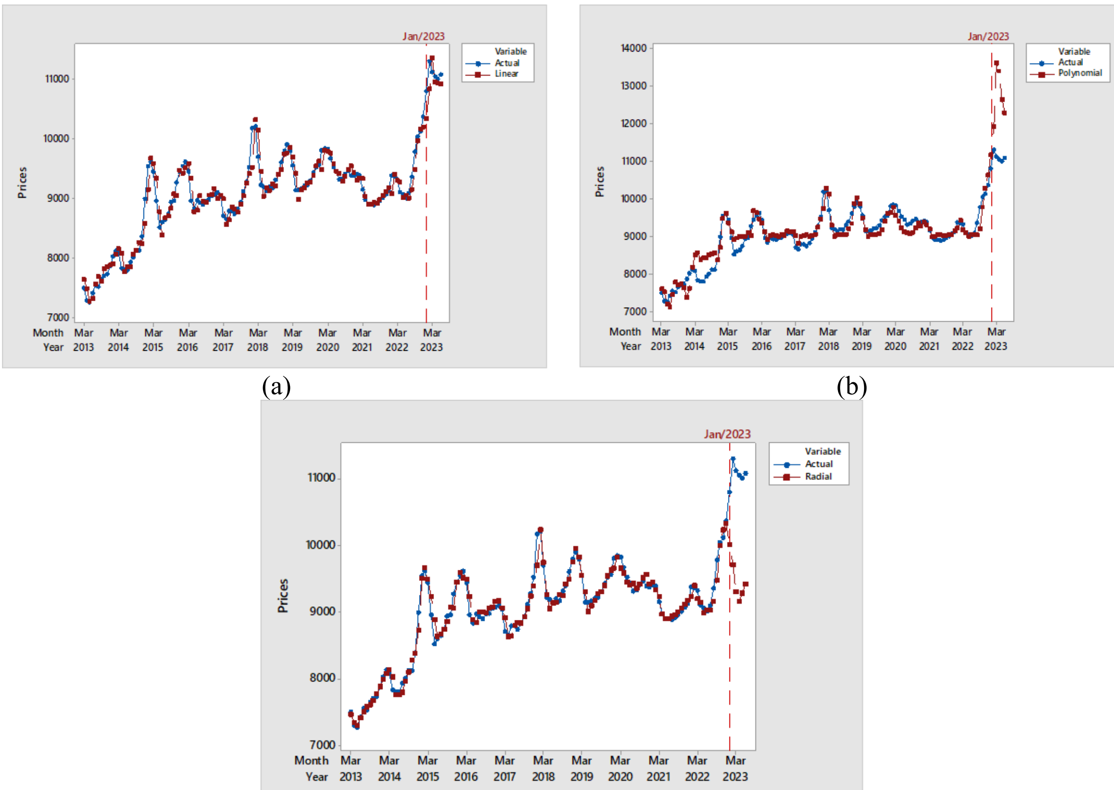
Before modeling the data, it is necessary to do partition of the datasets into training and testing datasets. Initially, the model is trained according to the training data from January 2013 to December 2022 which consist 108 months of the datasets. The 6 months rest of the datasets would be used to evaluate the forecasting performance of the model formed from January 2023 until June 2023. The testing datasets were used for assessment of the model.

The kernel function used in this study is a linear kernel. For comparison, polynomial kernel functions are used to find out the best model results. The best SVR modeling results are generated from optimal parameter values. The selection of optimal parameters is carried out using the grid search method. The combination of cost parameters used is between 1 to 100 with a difference in value of 1, epsilon parameters are used between 0.001 to 0.1 with a difference in values of 0.001. While the parameter of degree for the polynomial kernel used combination between 1 to 15 with a difference value of 1. In order to measure the performance of the model the Mean Absolute Percentage of Error (MAPE) is used following of the (1.8) equation.

Table 1. Forecasting Performance

Kernel	Parameter				MAPE	
	Epsilon	Cost	Degree	Gamma	Testing	Training
Linear	0.080	3.0	-	-	2.330887	1.200998
	0.078	3.0	-	-	2.317659	1.200759
	0.078	**3.1**	**-**	**-**	**2.315519**	**1.200605**
Polynomial	0.100	1.0	1	-	2.375195	1.208691
	0.037	1.5	2	-	63.43985	3.977249
	0.098	13.6	3	-	11.17574	2.030829
Radial	0.069	1	-	0.1996	17.52425	1.11822
	0.069	15	-	0.1996	16.73884	0.834056
	0.001	15	-	0.1596	18.20053	0.776371

The result of the forecasting performance of the several models using different kernels and its parameter shown in Table 2.1. The smallest MAPE score from testing data obtained from the model with a linear kernel using epsilon and cost parameter of 0.078 and 3.1 respectively. The smallest MAPE score for the polynomial kernel found in the parameter degree is 1 and the score is near with the MAPE score of linear kernel. Meaning that the model is following the linear form. This is evidenced by the smallest MAPE scores found in model with linear kernel resulting from the epsilon parameter and cost parameter of 0.078 and 3.1.



(a) (b)

(c)

Figure 3. Comparison of Actual Time Series Plots with Forecasting Results using SVR in (a) Linear Kernel (b) Polynimial Kernel (c) Radial Kernel

Forecasting results on the test data using linear kernels in the Figure 3 (a) shows a value of the rice prices closest to its actual value compared to using polynomial kernels and radial kernels. The results of forecasting using polynomial and radial kernels show that they are able to model on training data well but when evaluated using test data the model is not able to predict well the price of rice, which is called overfitting. This is not happend when forecasting using the linear kernel shown in Figure 3 (a) where the forecasting results in both the training data and the test data display values close to the actual values. Then the model with linear kernel and parameters ε = 0.078; C = 3.1 can be said to be the best model. Then the SVR-linear model (ε = 0.078; C = 3.1) is formed as follows:

$$\hat{Y}_t = 0.002176049 + 1.319091Y_{t-1} - 0.3605231Y_{t-2} + 0.07345937I_s \quad (10)$$

Equation (10) above can be explained that if there is an increase in rice prices by 1 rupiah in the previous month. the price of rice I t-month will also increase by 1.32 rupiahs. Meanwhile. if there is an increase in rice prices in the previous two months. the price of rice in the t-month will decrease by 0.36 rupiahs. As for the seasonal index where the seasonal period occurs every month as shown in figure (2), meaning that for every increase in one month period. the price of rice in the t-month will increase by 0.07 rupiahs.

CONCLUSION

Based on the results of the research that has been done above. it can be concluded on forecasting the average prices of rice using support vector regression formed the best model from a linear kernel with parameters ε = 0.078 and C = 3.1. The model produced the smallest MAPE value of 2.32% in testing data and 1.2% in training data. where this model was able to capture data patterns without ovefitting. The MAPE value of the best model is also already less than 10% which according to Lewis [8] meaning that the performance of the model to forecast the average price of rice is very high. Then this model can be used for forecast the average prices of rice for next period.

REFERENCE

[1] Alkaff, M., Khatimi, H., Puspita, W., & Sari, Y. "Modelling and predicting wetland rice production using support vector regression." *TELKOMNIKA (Telecommunication Computing Electronics and Control)*. vol. 17. no.2. pp. 819-825. 2019.

[2] Andariesta, D. T., & Wasesa, M. "Machine learning models for predicting international tourist arrivals in Indonesia during the COVID-19 pandemic: a multisource Internet data approach." *Journal of Tourism Futures*. 2022.

[3] De Gooijer, J.G. & Hyndman, R.J. "25 Years of Time Series Forecasting." *International journal of forecasting*. vol. 22. no. 3. pp. 443-473. 2006.

[4] Firdaus, Y.A., Ngatini, N. & Wijaya, S.U. "Pemodelan Regresi Data Panel Harga Beras di Wilayah Indonesia Bagian Barat." *J Statistika: Jurnal Ilmiah Teori dan Aplikasi Statistika*, vo.16. no. 2, pp. 486-498. 2023

[5] Gunn, S. R. "Support vector machines for classification and regression." *ISIS Technical Report*. vol. 14. no. 1. pp. 5-16. 1998.

[6] Hayadi, B. H., Sudipa, I. G. I., & Windarto, A. P. "Model peramalan pada peserta KB aktif jalur pemerintahan menggunakan Artificial Neural Network Back-propagation." *Matrik: Jurnal Managemen, Teknik Informatika, dan Rekayasa Komputer*, vol. 21. no.1, pp 11-20. 2021.

[7] Hou, X., Zhu, S., Xia, L., & Wu, G. "Stock price prediction based on Grey Relational Analysis and support vector regression." *Chinese Control And Decision Conference (CCDC), IEEE*. 2018. pp. 2509-2513.

[8] Lewis, C. D. *Industrial and business forecasting methods: A practical guide to exponential smoothing and curve fitting*. Butterworth-Heinemann. 1982.

[9] Lu, C. J., Lee, T. S., & Chiu, C. C. "Financial time series forecasting using independent component analysis and support vector regression." *Decision support systems*. vol. 47. no. 2. pp. 115-125. 2009.

[10] Montgomery, D.C., Jennings, C.L. & Kulahci, M. *Introduction to Time Series Analysis and Forecasting*, Second Edition. John Wiley & Sons Inc. 2015.

[11] Oğcu, G., Demirel, O. F., & Zaim, S. "Forecasting electricity consumption with neural networks and support vector regression." *Procedia-Social and Behavioral Sciences*. 2012. pp. 1576-1585.

[12] Paidipati, K.K., Chesneau, C., Nayana, B.M., Kumar, K.R., Polisetty, K. & Kurangi, C. "Prediction of rice cultivation in india—support vector regression approach with various kernels for non-linear patterns." *AgriEngineering*. vol. 3. no. 2. pp. 182-198. 2021.

[13] Resnia, R. “Fluktuasi Harga Bahan Pangan Pokok (Bapok) dan Daya Beli Kelompok Masyarakat Berpendapatan Rendah.” *Buletin Ilmiah Litbang Perdagangan*, vol. 6. no.2. pp. 169-188. 2012.

[14] Smola, A. J. & Schölkopf, B. “A tutorial on support vector regression.” *Statistics and computing*. vol. 14. no.3. pp. 199-222. 2004.

[15] Sapankevych, N. & Sankar, R. “Time series prediction using support vector machines: A survey.” *IEEE Computational Intelligence Magazine*. vol. 4. no.2, pp.24–38. 2009.

[16] Wiryawanto, Tiara Melati Putri, Zuyyina Hawani, & Muhamad Attar Ramadhani. "Comparison of Support Vector Machine (SVM) and Autoregressive Integrated Moving Average (ARIMA) Methods for Predicting Air Quality Using Python and KNIME." *J Statistika: Jurnal Ilmiah Teori dan Aplikasi Statistika*. vol. 16. no. 1. pp. 384-394. 2023.

[17] Zheng, T., Liu, S., Chen, Z., Qiao, Y., & Law, R. “Forecasting daily room rates on the basis of an LSTM model in difficult times of Hong Kong: evidence from online distribution channels on the hotel industry.” *Sustainability*. vol. 12. no. 18. pp. 7334. 2020.