

FORECASTING THAI DURIAN EXPORTS USING A HYBRID TIME SERIES SARIMA-SVR APPROACH

Thanakon Sutthison*

Received: May 14, 2024; Revised: September 19, 2024; Accepted: October 07, 2024

Abstract

Thailand is a prominent frontrunner in Southeast Asia in terms of its production and exportation of durian. Each year, the production of durian is a significant source of income for farmers. The consumption of durian has continued to increase throughout the years, both within the country and internationally. Consequently, cultivators have increased their durian cultivation to meet the expanding customer demand. However, there is a chance of excessive production, which has the ability to disrupt future market trends. Precise forecasts of durian export volumes are crucial for planning production strategies. This work aims to develop a novel SARIMA-SVR model for forecasting future export quantities of Thai durian, thereby fulfilling this need. The approach involves using the empirical mode decomposition (EMD) technique to reduce data fluctuations. The work conducts an analysis using the SARIMA model and calculates Gaussian white noise, which we then use as an input variable for SVR modeling. The study creates and evaluates four hybrid models, specifically SARIMASVR1, SARIMASVR2, SARIMASVR3, and SARIMASVR4. The study's findings suggest that SARIMASVR4 demonstrates superior accuracy compared to other models. Furthermore, they show the effectiveness of improving forecast accuracy by utilizing data filtering techniques such as EMD and incorporating Gaussian white noise.

Keywords: Durian export; Empirical mode decomposition; Gaussian white noise; Hybrid model; SARIMA; SVR

*Program of Mathematics, Faculty of Science, Ubon Ratchathani Rajabhat University, Ubon Ratchathani Province, 34000, Thailand.
Tel. 0-4535-2000; Fax. 0-4535-2070; E-mail: thanakon.s@ubru.ac.th*

** Corresponding author*

DOI: <https://doi.org/10.55766/sujst-2024-05-e04780>

Suranaree J. Sci. Technol. 31(5):030229(1-13)

Introduction

Durian is a tropical fruit with distinctive characteristics and a high price, typically cultivated in Southeast Asia (Sharma *et al.*, 2022). Thailand, as one of the leading producers and exporters of durian in Southeast Asia, has the largest total cultivation area, equivalent to 370,258.354536 acres (Office of Agricultural Economics, 2022). The majority of Thailand's durian exports consist of 90% fresh durian and approximately 10% frozen, pureed, and dried durian. Currently, Thailand's durian exports are continuously increasing, partly due to the growing popularity of Thai durian consumption among international consumers. As demand for durian increases, farmers must accelerate production by expanding cultivation areas to meet the growing demand. With a large quantity of durian produced, ensuring high-quality durian that meets the standards of foreign markets becomes challenging. Therefore, anticipating trends in durian demand in advance is crucial for relevant agencies to utilize as information for management and production planning throughout Thailand's durian supply chain in the future.

Time series forecasting is an intriguing technique with widespread applications for predicting future events. This study applies this technique to forecast the volume of Thai durian exports. Given the limited research on forecasting durian export quantities, it is an interesting research gap that warrants further investigation to provide crucial information for relevant agencies in planning and managing future durian export quantities. Currently, various researchers have developed highly accurate forecasting methods, some utilizing linear and non-linear models independently or in combination (Büyüksahin and Ertekin, 2019; Yu *et al.*, 2022). Linear forecasting models commonly used include Autoregressive Integrated Moving Average (ARIMA) and Seasonal Autoregressive Integrated Moving Average (SARIMA) (Büyüksahin and Ertekin, 2019; Xu *et al.*, 2019; de Araújo Moraes and da Silva Gomes, 2022). These models excel at accurately forecasting linear data but exhibit lower accuracy when forecasting non-linear data (Büyüksahin and Ertekin, 2019). Neural Networks (ANNs) (Dai *et al.*, 2022) and Support Vector Regression (SVR) (Xian and Che, 2022), on the other hand, are non-linear forecasting models that are flexible and accurate, but they might not work as well with linear time series data. To address these limitations, researchers have developed hybrid models that combine the strengths of linear and non-linear models, resulting in higher predictive

accuracy (Amini and Khashei, 2019; Xu *et al.*, 2019; Yang and Chen, 2019).

However, traditional forecasting models may make a lot of mistakes when they are used on complicated time series data, like data that is not stationary or linear, changes over time, and random noise (Liang *et al.*, 2019; Guo *et al.*, 2022). To mitigate this, researchers have applied data preprocessing techniques to reduce data variability before forecasting, leading to more accurate predictions (AL-Musaylh *et al.*, 2018; Zhang *et al.*, 2022). Empirical mode decomposition (EMD) is one of these methods. It breaks down time series data into parts that are not linear or stationary. These parts are called intrinsic mode functions (IMF) and residue. Analyzing data after applying EMD results in smoother data, leading to more accurate predictions. Researchers have successfully applied this approach, combining EMD with ensemble models, to forecast various phenomena (Xiang *et al.*, 2018; Fan *et al.*, 2020; Wang *et al.*, 2020). However, using EMD can be time-consuming for model training and may not be practical for current applications (Yaslan, 2017).

We propose a new hybrid model based on the above literature review to develop a forecasting model with reduced complexity and high accuracy for forecasting the quantity of Thai durian exports. This model utilizes empirical mode decomposition (EMD) to reduce signal noise in time-series data. The rationale behind this approach is to subtract the average value of the intrinsic mode functions (IMFs), analyzed by the EMD method, from the original time series data. After noise reduction, a hybrid SARIMA-SVR model forecasts the time series data. This model takes the best parts of SARIMA and SVR. SARIMA excels in accurately predicting seasonal time series data, while SVR excels in accurately predicting non-linear time series data, with no restrictions on the amount of data available for forecasting (Xu *et al.*, 2019). This research is particularly noteworthy as it is the first to apply a hybrid model to forecast Thai durian exports. Researchers have integrated Gaussian white noise into the forecasting model, enhancing the accuracy of predictions. The proposed model development process proceeds as follows:

- 1) To build a suitable SARIMA model, use time series data from Thai durian export quantities that have undergone noise reduction.
- 2) Forecast the time series data for Thai durian export quantities.
- 3) Compute the Gaussian white noise from step 2.
- 4) Utilize the values computed from steps 2 and 3

as input variables for forecasting Thai durian export quantities using an SVR model. This is the final step in the model creation and application of forecasted values for future use.

Materials and Methods

Data Preparation

This study employed time series data, capturing monthly durian export volumes from January 2011 to November 2023, encompassing a 156-month period. The study variables included Fresh Durian (unit = metric ton) or the ExFDTH variable, Paste Durian (unit = metric ton) or the ExPDTH variable, Frozen Durian (unit = metric ton) or the ExFroTH variable, and Freeze Dried Durian (unit = metric ton) or the ExFrDDTH variable. We divided the durian export data into two sets: a training dataset that covered 124 months from January 2011 to April 2021 for model creation, and a testing dataset that covered 32 months from May 2021 to November 2023 for model testing. We established the criteria for data division in cross-validation at 80% for training and 20% for testing, following established practices (Gholamy *et al.*, 2018).

Empirical Mode Decomposition (EMD)

The application of the EMD method is a technique used to diminish the oscillation signal or frequency of time series data prior to forecasting it through various methods (Xiang *et al.*, 2018; Fan *et al.*, 2020; Wang *et al.*, 2020). EMD's fundamental concept involves breaking down time series into a collection of oscillatory functions known as intrinsic mode functions (IMF). Each IMF is

required to meet two specified conditions (Abadan and Shabri, 2014).

- Across the complete dataset, the quantity of extrema (both maximum and minimum) and the number of zero-crossings must either match or differ by a maximum of one.

- At each point, the mean value of the envelope established by the local maxima and that formed by the local minima is always zero.

The Intrinsic Mode Functions (IMFs) can be extracted from the time series dataset through an iterative decomposition process, described by the following steps:

Step 1: Identify all the local extrema of time series Y_t .

Step 2: Connect all the local maxima by a cubic spline line as the upper envelope $e_{\max}(t)$, and repeat the procedure for the local minima to produce the lower envelope $e_{\min}(t)$.

Step 3: Calculate the mean of the upper and lower envelopes and let it as $m(t)$.

$$m(t) = (e_{\max}(t) + e_{\min}(t)) / 2 \quad (1)$$

Step 4: Extract the mean from the data and define the difference as $d(t)$.

$$d(t) = Y_t - m(t). \quad (2)$$

Step 5: Check $d(t)$: if it satisfies those two conditions of IMF, output it as the i th IMF and replace Y_t with the residue.

$$r(t) = Y_t - d(t), \quad (3)$$

if $d(t)$ is not an IMF, replace Y_t with $d(t)$.

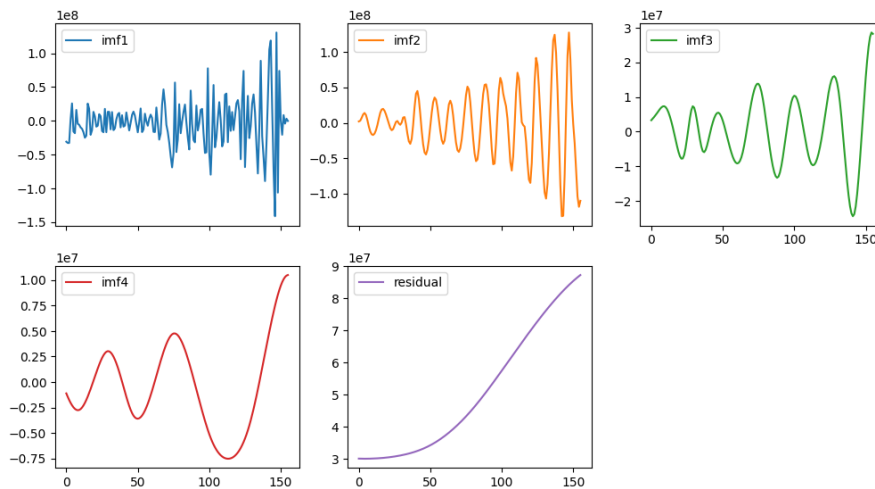


Figure 1. Shows the EMD analysis of Thailand's fresh durian export

Step 6: Repeat steps 1-5 until the residue becomes a constant, a monotonic function, or a function with only one maximum and one minimum from which no more IMF can be extracted.

Then the original time series Y_t can be expressed as the sum of these IMFs and residue.

$$Y_t = \sum_{i=1}^N imf(t) + r(t). \quad (4)$$

Where N is the number of IMFs, and $r(t)$ is the final residue. A demonstration highlighting the analysis of Thailand's fresh durian export volume through time series data, employing the EMD method is depicted in Figure 1.

Reduction of Oscillation Signal of Time Series Data

Step 1: Utilizing the EMD method for data analysis yields intrinsic mode functions (IMFs) and residual information.

Step 2: Average IMF (aIMF) computation is performed using Equation (5).

$$aIMF(t) = \frac{1}{N} \sum_{i=1}^N imf(t). \quad (5)$$

Step 3: Compute the noise-free signal using Equation (6).

$$X_t = Y_t - aIMF(t). \quad (6)$$

Use the X_t value, which has been reduced to a fast-swing signal, as a key part of creating the next forecast model.

Data Normalization

A forecasting model utilizing support vector regression (SVR) requires proper preparation of the data. We use the SARIMA model to determine the appropriate time lag, which aids in feature selection. The min-max scaler technique standardizes the data before applying the model,

ensuring that all values fall within the range of 0 to 1.

The transformation is explained as follows:

$$\tilde{x}_i^c = \frac{x_i^c - x_{min}^c}{x_{max}^c - x_{min}^c}, \quad (7)$$

where:

$\tilde{x}_i^c$ is the normalized value.

x_i^c is the original value of the I -th data point for the c -th feature.

x_{min}^c is the minimum value of the c -th feature.

x_{max}^c is the maximum value of the c -th feature.

SARIMA Model

The SARIMA model is a highly accurate statistical model suitable for predicting time series with seasons (Ren *et al.*, 2022). Currently, SARIMA models are often used for forecasting agricultural export data (Vonglao *et al.*, 2023). For example, Thai durian export data usually exhibits seasonal elements, i.e., exports are high in volume from May to June and repeat in the following year. The use of the SARIMA model has an important condition: the time series data used for forecasting must be stationary (constant mean and constant variance). There are 6 components of the SARIMA model: Autoregressive (AR), Integration (I), Moving Average (MA), Seasonal Autoregressive (SAR), Seasonal Integration (SI), and Seasonal Moving Average (SMA). The corresponding parameters are p , d , q , P , D , and Q , respectively. The general SARIMA model is denoted as SARIMA (p , d , q) (P , D , Q) [m], where m represents the order of seasonal differencing; for monthly time series, m is set to 12. Detailed explanations of these variables are listed in Table 1 (Xu *et al.*, 2019). The expression is shown in Equation (8).

$$\phi_p(B)\Phi_P(B^m)\nabla^d\nabla_m^D Y_t = \theta_q(B)\Theta_Q(B^m)a_t \quad (8)$$

Table 1. Name and definition of each parameter in SARIMA

Name	Parameter	Operator	Equation
Autoregressive (AR)	p	$\phi_p(B)$	$1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p$
Integration (I)	d	∇^d	$(1 - B)^d$
Moving Average (MA)	q	$\theta_q(B)$	$1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q$
Seasonal AR (SAR)	P	$\Phi_P(B^m)$	$1 - \phi_1 B^m - \dots - \phi_P B^{P \cdot m}$
Seasonal Integration (SI)	D	∇_m^D	$(1 - B^m)^D$
Seasonal MA (SMA)	Q	$\Theta_Q(B^m)$	$1 - \theta_1 B^m - \dots - \theta_Q B^{Q \cdot m}$
Seasonality	m	m	m
Back Shift	B	B	$B^n \cdot Y_t = Y_{t-n}$
Gaussian White Noise	σ^2	a_t	$a_t(0, \sigma^2)$

Where $\nabla^d \nabla_m^D Y_t$ is the seasonally differenced series, B is the back-shift operator, and a_t is the Gaussian White Noise with zero mean and variance σ^2 . The development of a SARIMA model encompasses the following particulars:

1. Preparation of time series data for durian export quantities (details in step 1: data preparation)
2. Analyze data using the EMD method to reduce the signal fluctuations in time series data (detail in steps 2 and 3: reduction of oscillation signal of Time Series Data).

3. The Dickey-Fuller test is utilized to assess stationarity in time series data; in instances of non-stationarity, logarithmic transformation, and differencing are implemented before model specification.

4. From step 3, utilizing stationary time series data, the researchers programmed Python commands to iteratively select parameters p , q , P , and Q within the range of 0 to 2. The selection of a suitable model was performed iteratively, employing criteria such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). The model with the lowest AIC and BIC values was deemed appropriate.

5. After selecting the model in step 4 forecasting and parameter estimation were performed. Diagnostic checks of the model were then conducted using the 'stats models' package in Python for automatic model adequacy assessment. Additional details are available in the Results and Discussion section.

6. Using the model approved during the diagnostic check in step 5, we proceeded to forecast and evaluate the performance of the subsequent model.

SVR Model

The Support Vector Regression (SVR) model is a versatile and high-precision tool used for predicting nonlinear time series data. Its fundamental principle involves conducting a nonlinear mapping $\Phi(\cdot)$ of the training dataset $\{x_i, y_i\}_{i=1}^l$ from the original feature space to a higher-dimensional space through the utilization of the kernel trick (Adaryani *et al.*, 2022; Laouafi *et al.*, 2022). Let $\{x_i, y_i\}_{i=1}^l$ denote the training set, where $x_i \in \mathcal{R}^d$ represents the i -th input vector, $y_i \in \mathcal{R}$ denotes the i -th prediction output of x_i , d signifies the embedding dimension of the time series, and l indicates the number of training samples. The primary objective of SVR is to identify the optimal function from a set of potential functions, as expressed in Equation (9):

$$f(x) = w^T x + b, w \in \mathcal{R}^d, b \in \mathcal{R} \quad (9)$$

Where: $x \in \mathcal{R}^d$ is the input feature vector representing the features of a sample. Each component of x corresponds to a specific feature in the dataset, with d being the number of features or dimensions. w is the weight vector estimated by minimizing the regularized risk function shown in Equation (10) and b is a bias or threshold.

$$\frac{1}{2} \|w\|^2 + C \sum_{i=1}^l L(y_i, f(x_i)) \quad (10)$$

In order to find the best function f , it is necessary to minimize the regularized risk, where $C > 0$ is a regularization factor, $\|\cdot\|$ is a 2 norm and $L(\cdot)$ is a loss function. In order to induce sparsity in SVR the ε -insensitive loss function is presented in Equation (10), which creates an ε -tube allowing some predictions to fall within the boundaries of this ε -tube:

$$L(y, f(x)) = \begin{cases} 0, & \text{if } |f(x) - y| < \varepsilon \\ |f(x) - y| - \varepsilon, & \text{otherwise} \end{cases} \quad (11)$$

With the ε -insensitive function the SVR is formulated as shown in the following Equation:

$$\begin{aligned} &\text{Minimize } \frac{1}{2} \|w\|^2 + C \sum_{i=1}^l (\xi_i - \xi_i^*) \\ &\text{Subject to } \begin{cases} w^T x_i + b - y_i \leq \varepsilon + \xi_i \\ y_i - w^T x_i - b \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0, i = 0, \dots, l \end{cases} \end{aligned} \quad (12)$$

where ξ and ξ^* are slack variables used to measure the errors that occurred by values outside the boundaries determined by the ε -tube. The use of kernels allows the SVR to perform nonlinear mapping into a higher dimensionality space. The regression function takes the form:

$$f(x) = \sum_{i=1}^l (\alpha_i, \alpha_i^*) k(x_i, x_j) + b \quad (13)$$

where α and α^* are Lagrange multipliers and $k(x_i, x_j)$ is a kernel function. The Gaussian radial basis function (RBF) is widely employed in time series forecasting (Adaryani *et al.*, 2022; Laouafi *et al.*, 2022)

$$k(x_i, x_j) = \exp \left(-0.5 \|x_i - x_j\|^2 / \sigma^2 \right),$$

where σ is a parameter of the RBF kernel. To forecast the quantity of Thai durian exports using an SVR model, the data must be transformed into the $d = \{x_i, y_i\}_{i=1}^l$ form. Subsequently, it is divided into two parts: a Training dataset and a Testing dataset (refer to Figure 2) for model creation and

testing purposes. This research proposes a method for selecting a novelty training dataset from the SARIMA model and employs the EMD technique to reduce signal fluctuations, thereby enhancing forecast accuracy. The details of the model creation are as follows:

1. Prepare the time series data for the durian export volume and perform EMD analysis to reduce high-frequency oscillations. (described in Sections EMD Analysis and Reduction of Oscillation Signal of Time Series Data)

2. Divide time series data into two sets: a training dataset and a testing dataset.

3. Analyze the SARIMA model on the training dataset to determine optimal parameters (details are presented in the section on SARIMA model analysis). Then, the sequence of time series predictions generated by the SARIMA model is taken. This sequence is used as the input data for training the SVR model.

4. Reverse elements are calculated based on the parameters of the SARIMA model obtained in step 3. This calculation is crucial for establishing the input variable in the subsequent SVR model. For example, if a trained SARIMA model has parameters (1,1,1) and (0,1,1), the reverse elements are determined using the SARIMA Equation.

$$\phi_1(B)\Phi_0(B^{12})\nabla^1\nabla_{12}^1Y_t = \theta_1(B)\Theta_1(B^{12})a_t \quad (14)$$

According to Table 1, Equation (14) can be rewritten as:

$$(1 - \phi_1 B)(1 - B)(1 - B^{12})Y_t = (1 - \theta_1 B)(1 - \theta_1 B^{12})a_t \quad (15)$$

Then,

$$[1 - (1 + \phi_1)B + \phi_1 B^2 - B^{12} + (1 + \phi_1)B^{13} - \phi_1 B^{14}]Y_t = (1 - \theta_1 B - \theta_1 B^{12} + \theta_1 \theta_1 B^{13})a_t \quad (16)$$

$$Y_t - (1 + \phi_1)Y_{t-1} + \phi_1 Y_{t-2} - Y_{t-12} + (1 + \phi_1)Y_{t-13} - \phi_1 Y_{t-14} = a_t - \theta_1 a_{t-1} - \theta_1 a_{t-12} + \theta_1 \theta_1 a_{t-13}. \quad (17)$$

Finally,

$$Y_t = (1 + \phi_1)Y_{t-1} - \phi_1 Y_{t-2} + Y_{t-12} - (1 + \phi_1)Y_{t-13} + \phi_1 Y_{t-14} + a_t - \theta_1 a_{t-1} - \theta_1 a_{t-12} + \theta_1 \theta_1 a_{t-13}. \quad (18)$$

$$a_t = Y_t - (1 + \phi_1)Y_{t-1} - \phi_1 Y_{t-2} + Y_{t-12} - (1 + \phi_1)Y_{t-13} + \phi_1 Y_{t-14} + \theta_1 a_{t-1} + \theta_1 a_{t-12} - \theta_1 \theta_1 a_{t-13}. \quad (19)$$

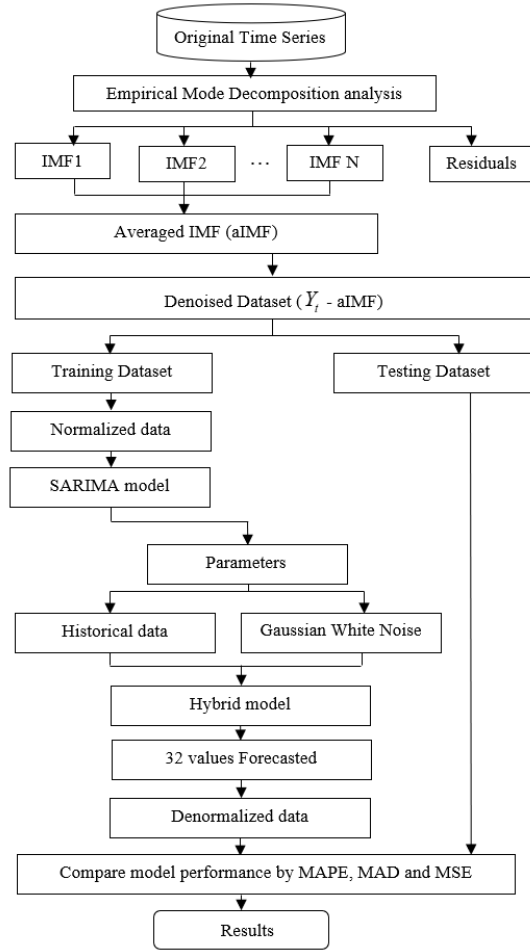


Figure 2. Procedure of the hybrid SARIMASVR method for forecasting Thai durian export volumes

From Equation (18), it can be seen that Y_t is related to Y_{t-1} , Y_{t-2} , Y_{t-12} , Y_{t-13} , Y_{t-14} , a_t , a_{t-1} , a_{t-12} , a_{t-13} . And Equation (19) provides a way to calculate the value of a_t in reverse, known as Gaussian White Noise, which can be introduced into the SVR forecasting model. The details of the vector input (x_i) configuration to forecast the vector output (y_i) consist of four configurations described below.

Configuration 1: SARIMASVR1

$$\hat{Y}_t = SVR(Y_{t-1}, Y_{t-2}, Y_{t-12}, Y_{t-13}, Y_{t-14})$$

Where $\hat{Y}_t$ is forecasting value, SVR is the SVR model, and Y_{t-1} , Y_{t-2} , Y_{t-12} , Y_{t-13} , Y_{t-14} are the historical part of Equation (18).

Configuration 2: SARIMASVR2

$$\hat{Y}_t = SVR(Y_{t-1}, Y_{t-2}, Y_{t-12}, Y_{t-13}, Y_{t-14}, SARIMA)$$

Where $\hat{Y}_t$ is forecasting value, SVR is the SVR model, Y_{t-1} , Y_{t-2} , Y_{t-12} , Y_{t-13} , Y_{t-14} are historical parts of Equation (18), and SARIMA is forecasting results from the SARIMA model.

Configuration 3: SARIMASVR3

$$\hat{Y}_t = SVR(Y_{t-1}, Y_{t-2}, Y_{t-12}, Y_{t-13}, Y_{t-14}, a_{t-1}, a_{t-12}, a_{t-13})$$

Where $\hat{Y}_t$ is forecasting value, SVR is the SVR model, Y_{t-1} , Y_{t-2} , Y_{t-12} , Y_{t-13} , Y_{t-14} are historical parts of Equation (18), and a_{t-1} , a_{t-12} , a_{t-13} are the Gaussian White Noise part of Equation (18).

Configuration 4: SARIMASVR4

$$\hat{Y}_t = SVR(Y_{t-1}, Y_{t-2}, Y_{t-12}, Y_{t-13}, Y_{t-14}, a_{t-1}, a_{t-12}, a_{t-13}, SARIMA)$$

Where $\hat{Y}_t$ is forecasting value, SVR is the SVR model, Y_{t-1} , Y_{t-2} , Y_{t-12} , Y_{t-13} , Y_{t-14} are historical parts of Equation (18), and a_{t-1} , a_{t-12} , a_{t-13} are the Gaussian White Noise part of Equation (18), and SARIMA is forecasting results from SARIMA model.

5. When formatting data in step 4, the researcher normalizes it to reduce its influence and then employs it for forecasting with SVR models in step 6.

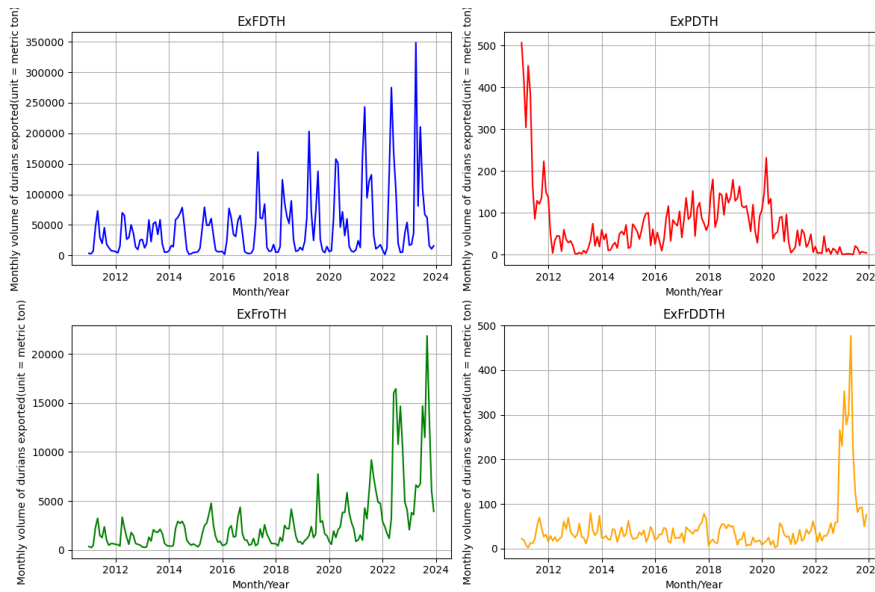
6. The SVR analysis forecasts statistical indicators for monthly durian exports in Thailand. In this study, we utilize the “sklearn.svm” package in “Python” to build the SVR model. We use a grid search method to determine the optimal parameters of the SVR model, as detailed in Table 2. The SVR model employs the RBF as its kernel function. To prevent overfitting, we implement a 5-fold cross-validation technique on the training set. Upon determining the appropriate parameter value, a forecast is generated in advance. Subsequently, the dataset is transformed into real values prior to its utilization in computing the subject’s performance measurement at step 7.

Table 2. The rangers of parameters of SVR

Parameter	Range
Cost	100, ..., 1000
Epsilon	0.001, ..., 0.01
Gamma	0.001, ..., 0.01

Table 3. The performance criteria for the proposed approaches

Criteria	Equation
Mean Absolute Percentage Error	$MAPE = \frac{1}{n} \sum_{i=1}^n \left \frac{y_i - f(x_i)}{y_i} \right $
Mean Absolute Deviation	$MAD = \frac{1}{n} \sum_{i=1}^n y_i - f(x_i) $
Mean Squared Error	$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2$

**Figure 3. Shows a time-series dataset illustrating the export volume of durian from Thailand**

7. The testing dataset is applied to the trained hybrid model to obtain forecasting values. Then, the forecasting accuracy is calculated to measure the performance of each model. The performance criteria are shown in Table 3, where $\hat{f}(x_i)$ is the forecasting result, and y_i is the real value. This study uses Mean Absolute Percentage Error (MAPE), Mean Absolute Deviation (MAD), and Mean Squared Error (MSE) as its three metrics for evaluating forecast performance. When the MAPE, MAD, and MSE values are small, the forecast model is highly accurate. However, the choice of the three benchmarks must be made cautiously when data outliers occur, resulting in inaccurate performance measurement values (Koondee *et al.*, 2023). For this reason, the emphasis is on preparing data before forecasting. Currently, there are many ways to deal with anomalous data, such as normalization minmax-scaler and using EMD methods to reduce data fluctuations before forecasting. Given the limitations of the aforementioned performance metrics, several additional metrics are presented to assess the accuracy of the durian export forecast model, which is more comprehensive.

Results and Discussion

Basic Statistical Analysis

According to the analysis of four durian export datasets (Data Preparation), Thailand's durian exports exhibit clear seasonal patterns and fluctuations (Figure 3). Export quantities peak

during mid-year (April-July) and decline towards year-end (October-December). This seasonality supports the development of a SARIMA forecasting model (see details provided in the section on SARIMA model analysis). Basic statistics (Table 4) show significant differences among durian types, with fresh durians (ExPDTH) having the highest average quantity (45,360.67 metric tons) and considerable month-to-month variations. These findings align with data from the Office of Agricultural Economics (2022), indicating that 90% of Thailand's durian exports are fresh, with the remaining 10% consisting of frozen, paste, and dried durians (Table 4).

SARIMA Analysis Summary

1. Analyze data using the EMD method to reduce the signal fluctuations in time series data

2. Evaluation of Data Stationarity.

We observed non-stationarity in the data after analyzing the quantity of Thai durian exports. To address this, we applied a logarithmic transformation to reduce variance, followed by the implementation of both non-seasonal ($d = 1$) and seasonal ($D = 1$) differencing techniques to achieve stationarity. Further examination using the Augmented Dickey-Fuller test (see Table 5) confirmed that all four datasets exhibited stationary properties after transformation. Subsequently, the data underwent analysis, resulting in the selection of a model as outlined in the section on SARIMA model analysis.

Table 4. Summarizes the descriptive statistics of the export quantities of Thai durian

Data set	data	Number	Mean	Std.	Max.	Min.	Median
ExPDTH	All sample	156	45,360.67	55,131.75	348,527.74	1,432.73	24,785.34
	Training	124	37,026.26	38,937.24	203,283.23	1,742.22	23,705.65
	Testing	32	77,656.53	88,473.18	348,527.74	1,432.73	36,657.69
ExFDTH	All sample	156	71.95	81.23	506.38	0.21	50.27
	Training	124	86.55	84.77	506.38	1.74	67.30
	Testing	32	15.36	17.88	59.65	0.21	6.90
ExFroTH	All sample	156	2,769.70	3,459.32	21,860.33	230.44	1,638.42
	Training	124	1,585.58	1,236.13	7,755.91	230.44	1,267.16
	Testing	32	7,358.18	5,141.01	21,860.33	1,173.09	6,043.22
ExFrDDTH	All sample	156	45.69	61.93	476.7	2.09	31.21
	Training	124	30.63	17.12	79.42	2.09	26.62
	Testing	32	104.04	116.59	476.7	14.72	53.3

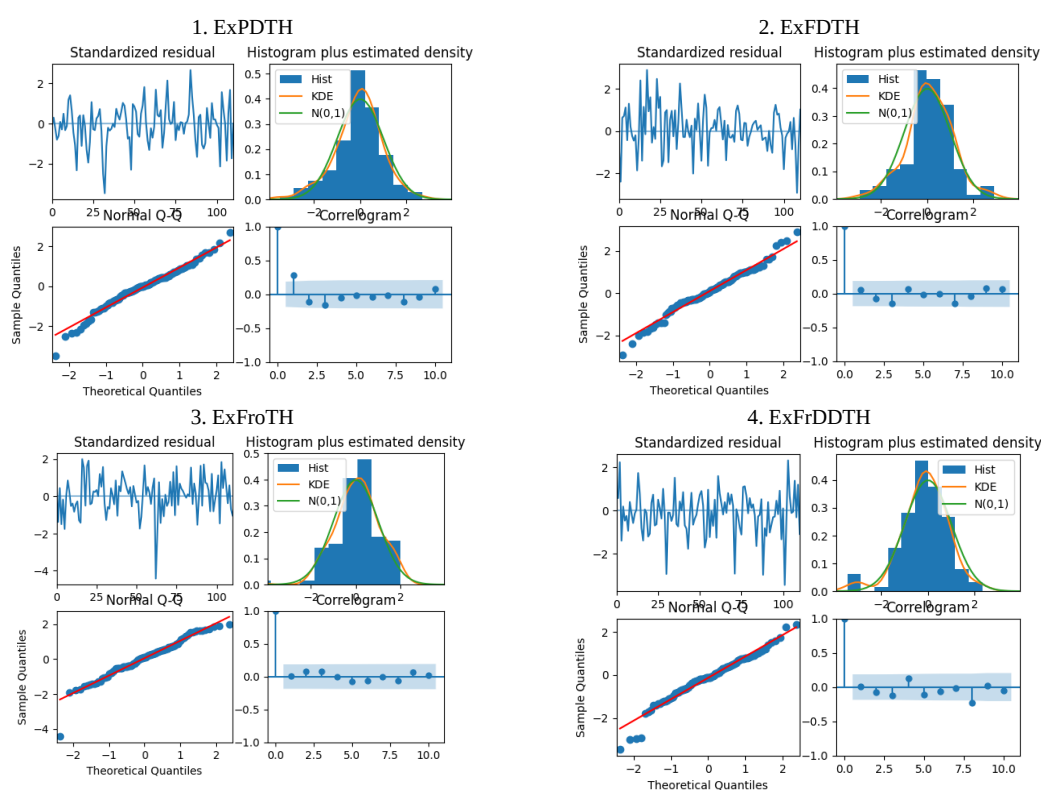
Note: unit = metric ton

Table 5. Augmented dickey-fuller test for data stationarity

Data set	Pre-transformed Data		Post-transformed Data	
	ADF	p-value	ADF	p-value
ExPDTH	1.073	0.994	-7.654	<0.001
ExFDTH	-2.015	0.279	-10.139	<0.001
ExFroTH	2.304	0.998	-9.953	<0.001
ExFrDDTH	0.423	0.982	-6.159	<0.001

Table 6. Parameter estimation for the SARIMA model

Data set	Model	Parameter	Coefficient	Std. Error	Z-Statistics	p-value
ExPDTH	SARIMA (0,1,1)(0,1,1) ₁₂ AIC = 101.687 BIC = 109.816	MA1	-0.9170	0.061	-14.606	<0.001
		SMA1	-0.5468	0.113	-4.179	<0.001
ExFDTH	SARIMA (1,1,1)(0,1,1) ₁₂ AIC = 132.583 BIC = 140.712	AR1	-0.3003	0.095	-3.158	0.002
		MA1	-0.3820	0.082	-4.683	<0.001
		SMA1	-0.8103	0.120	-6.780	<0.001
ExFroTH	SARIMA (0,1,1)(0,1,1) ₁₂ AIC = 148.438 BIC = 159.276	MA1	-0.6131	0.083	-7.354	<0.001
		MA2	-0.3229	0.080	-4.055	<0.001
		SMA1	-0.5945	0.097	-6.130	<0.001
ExFrDDTH	SARIMA (0,1,1)(0,1,1) ₁₂ AIC = 287.782 BIC = 295.911	MA1	-0.9204	0.051	-18.075	<0.001
		SMA1	-0.7881	0.130	-6.065	<0.001

**Figure 4. SARIMA model diagnosis for Thailand's durian export quantity**

3. Selection of SARIMA Models.

The researchers followed the approach described in the section on SARIMA model analysis to determine the suitable SARIMA model. The selected model is presented in Table 6.

4. Model Inspection and Diagnostics.

The statistically significant parameter estimates in Table 6 validate the forecasting model. Figure 4 depicts the model validation procedures. In summary, the model accurately predicts Thailand's durian exports across all datasets, with a normal distribution in residual values and no significant lag shown in the correlogram graph.

5. Forecasting and Performance Evaluation.

We conducted forecasting and performance evaluations after obtaining a suitable model. Table 8 presents the results.

Summary of SVR Model

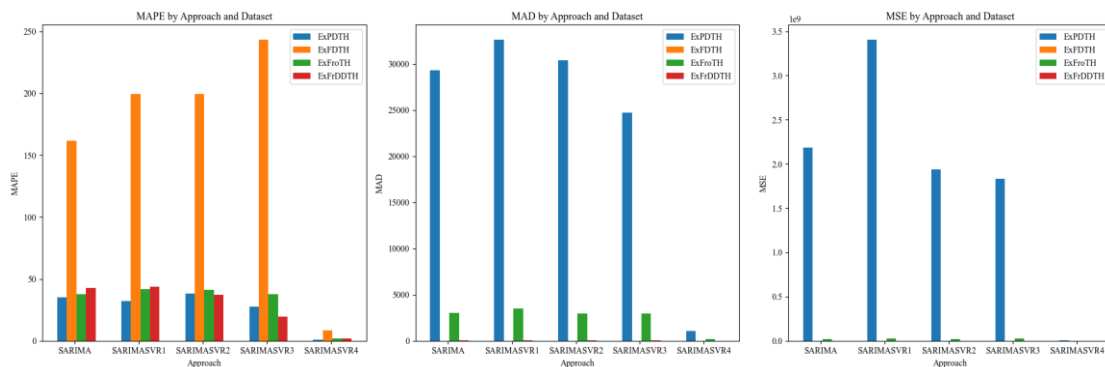
Following Step 6, we built the Support Vector Regression (SVR) model by examining the EMD technique to remove and reduce data fluctuations. Subsequently, we performed a forecasting study utilizing SARIMA modeling (refer to Table 6) and determined the input variables for training Support Vector Regression (SVR) (refer to Table 7). Smaller values indicate better model performance, as shown

Table 7. Displays the input and output variables for each model

Data set	Model	Input	Output
ExPDTH	SARIMA	Y_t	Y_t
	SARIMASVR1	$Y_{t-1}, Y_{t-12}, Y_{t-13}$	Y_t
	SARIMASVR2	$Y_{t-1}, Y_{t-12}, Y_{t-13}, \text{SARIMA}$	Y_t
	SARIMASVR3	$Y_{t-1}, Y_{t-12}, Y_{t-13}, a_t, a_{t-1}, a_{t-12}, a_{t-13}$	Y_t
	SARIMASVR4	$Y_{t-1}, Y_{t-12}, Y_{t-13}, a_t, a_{t-1}, a_{t-12}, a_{t-13}, \text{SARIMA}$	Y_t
ExFDTH	SARIMA	Y_t	Y_t
	SARIMASVR1	$Y_{t-1}, Y_{t-2}, Y_{t-12}, Y_{t-13}, Y_{t-14}$	Y_t
	SARIMASVR2	$Y_{t-1}, Y_{t-2}, Y_{t-12}, Y_{t-13}, Y_{t-14}, \text{SARIMA}$	Y_t
	SARIMASVR3	$Y_{t-1}, Y_{t-2}, Y_{t-12}, Y_{t-13}, Y_{t-14}, a_t, a_{t-1}, a_{t-12}, a_{t-13}$	Y_t
	SARIMASVR4	$Y_{t-1}, Y_{t-2}, Y_{t-12}, Y_{t-13}, Y_{t-14}, a_t, a_{t-1}, a_{t-12}, a_{t-13}, \text{SARIMA}$	Y_t
ExFroTH	SARIMA	Y_t	Y_t
	SARIMASVR1	$Y_{t-1}, Y_{t-12}, Y_{t-13}$	Y_t
	SARIMASVR2	$Y_{t-1}, Y_{t-12}, Y_{t-13}, \text{SARIMA}$	Y_t
	SARIMASVR3	$Y_{t-1}, Y_{t-12}, Y_{t-13}, a_t, a_{t-1}, a_{t-12}, a_{t-13}$	Y_t
	SARIMASVR4	$Y_{t-1}, Y_{t-12}, Y_{t-13}, a_t, a_{t-1}, a_{t-12}, a_{t-13}, \text{SARIMA}$	Y_t
ExFrDDTH	SARIMA	Y_t	Y_t
	SARIMASVR1	$Y_{t-1}, Y_{t-12}, Y_{t-13}$	Y_t
	SARIMASVR2	$Y_{t-1}, Y_{t-12}, Y_{t-13}, \text{SARIMA}$	Y_t
	SARIMASVR3	$Y_{t-1}, Y_{t-12}, Y_{t-13}, a_t, a_{t-1}, a_{t-12}, a_{t-13}$	Y_t
	SARIMASVR4	$Y_{t-1}, Y_{t-12}, Y_{t-13}, a_t, a_{t-1}, a_{t-12}, a_{t-13}, \text{SARIMA}$	Y_t

Table 8. Presents the evaluation of statistical indicators for the quantity of durian exports in Thailand

Dataset	Approach	MAPE	MAD	MSE
ExPDTH	SARIMA	35.037	29,297.228	2,181,654,563.619
	SARIMASVR1	32.289	32,586.792	3,402,411,724.730
	SARIMASVR2	38.347	30,353.777	1,941,093,895.240
	SARIMASVR3	27.545	24,720.353	1,834,941,094.510
	SARIMASVR4	0.827	1,063.552	7,148,381.320
ExFDTH	SARIMA	161.778	10.297	148.144
	SARIMASVR1	199.487	13.361	252.932
	SARIMASVR2	199.584	13.702	270.123
	SARIMASVR3	243.056	13.904	251.834
	SARIMASVR4	8.344	0.427	0.242
ExFroTH	SARIMA	37.415	2,996.264	17,623,484.472
	SARIMASVR1	41.840	3,466.910	28,580,801.210
	SARIMASVR2	41.210	2,944.890	19,697,750.820
	SARIMASVR3	37.44	2,933.990	23,575,597.310
	SARIMASVR4	1.720	184.320	116,996.050
ExFrDDTH	SARIMA	42.867	66.162	13,172.123
	SARIMASVR1	43.790	49.670	8,431.910
	SARIMASVR2	37.180	37.660	3,976.430
	SARIMASVR3	19.500	31.210	3,238.170
	SARIMASVR4	2.100	4.430	84.870

**Figure 5. Shows a comparison of performance measurement criteria for forecasting models**

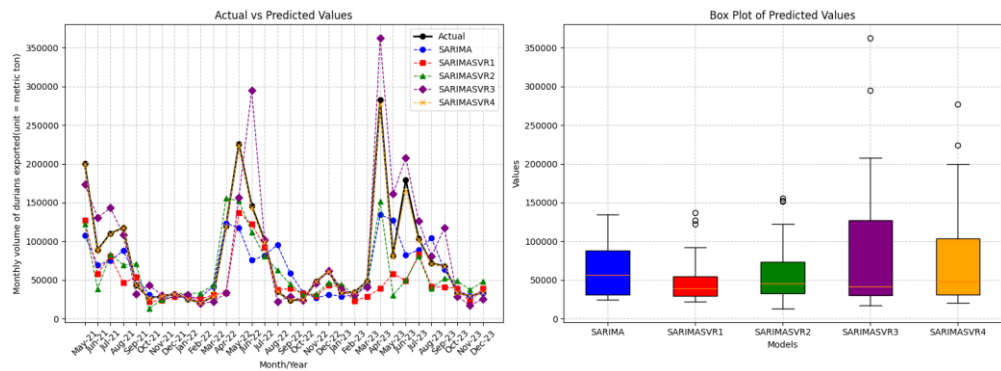


Figure 6. Shows a line graph and box plot that compare the actual and forecasted quantities of fresh durian exports

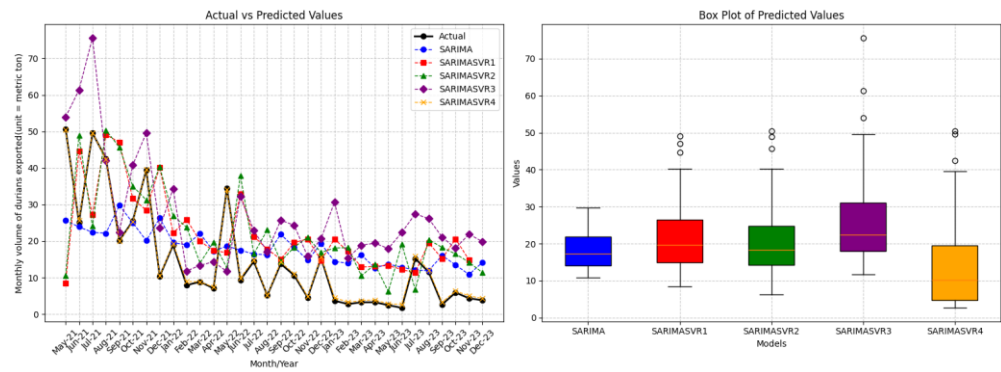


Figure 7. Shows a line graph and box plot that compare the actual and forecasted quantities of paste durian exports

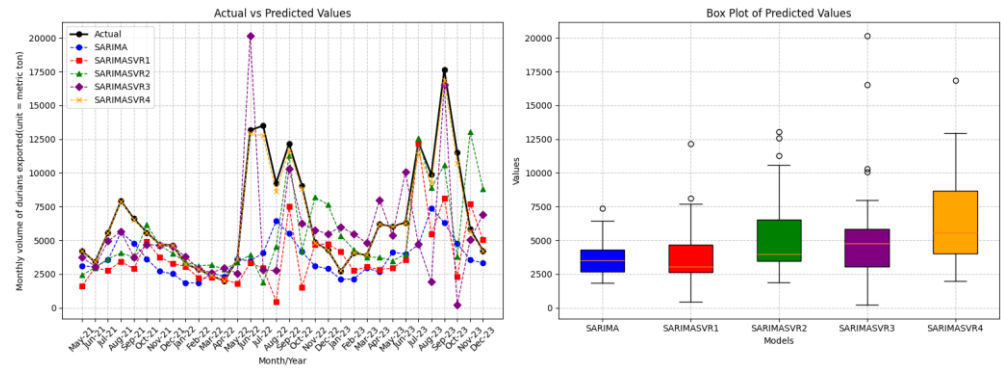


Figure 8. Shows a line graph and box plot that compare the actual and forecasted quantities of frozen durian exports

in Table 8 and Figure 5. Table 8 reveals that the hybrid model selects input variables from deoscillated time series data using the EMD method. Subsequently, the SARIMA model analyzes the data to determine input variables based on historical observation values' ranking. The forecast results from the SARIMA and Gaussian White

models consistently outperform those from single time series models across all four datasets, as evidenced by all forecast performance metrics. This superior accuracy stems from leveraging historical observational values to select pertinent factors related to current time series data, augmented by the addition of Gaussian White

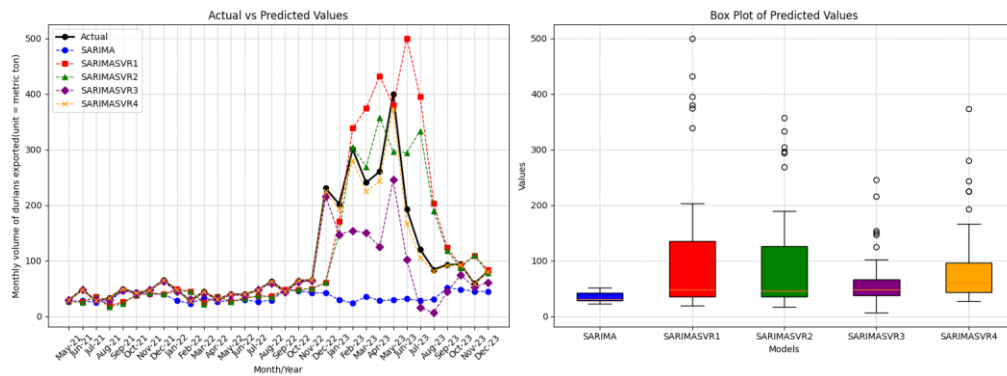


Figure 9. Shows a line graph and box plot that compare the actual and forecasted quantities of freeze-dried durian exports

Noise. This inclusion introduces another non-linear relationship factor, enriching the prediction process for time series data (SARIMASVR3, SARIMASVR4). By incorporating the SVR model, renowned for its flexibility and high accuracy (Xian and Che, 2022), the hybrid model adeptly discerns the underlying structure of time series data.

The SARIMASVR4 model has lower prediction errors compared to the SARIMASVR1, SARIMASVR2, and SARIMASVR3 models when considering how well the hybrid models from the set of Thai durian export quantity data work in all four datasets (Table 8, Figures 6-9). For example, in the ExPDTH dataset, the SARIMASVR4 model has a MAPE (mean absolute percentage error) of 0.827%, while SARIMASVR1, SARIMASVR2, and SARIMASVR3 have MAPE values of 32.289%, 38.347%, and 27.545%, respectively. This is consistent with Figure 6, which shows 32-month forecasts. The SARIMASVR4 model provides the closest forecast to the actual value, while the other datasets also yield consistent results. This indicates that combining data from SARIMA model forecasts and Gaussian white noise improves the accuracy of predicting Thai durian export quantities. These research findings contradict Xu *et al.* (2019), found that adding SARIMA input variables and Gaussian white noise to SVR models did not improve prediction accuracy. The different results may be due to the fact that this study used EMD to reduce data volatility before analyzing it with SARIMA models. We then selected appropriate lag orders from the SARIMA models and Gaussian white noise as input variables for SVR model learning. Considering the MSE values in Table 8 and Figure 5, these values are significantly higher than other performance metrics, especially in the ExPDTH dataset. This discovery suggests that MSE is sensitive to extreme values; even a single

extreme value may significantly increase the value (Kavuncuoglu *et al.*, 2018; Koondee *et al.*, 2023). Furthermore, the study revealed that a significant deviation in the forecast from the actual value leads to a significant increase in the MSE. As a result, selecting input variables from the SARIMA model has an impact on hybrid models' accuracy.

Conclusions

This study presents and evaluates four hybrid models, namely SARIMASVR1, SARIMASVR2, SARIMASVR3, and SARIMASVR4, which combine SARIMA and SVR models for forecasting the export quantity of Thai durians. The dataset comprises four sets of data: ExPDTH, ExFDTH, ExFrOTH, and ExFrDDTH. The primary contribution of this study lies in proposing a novel forecasting model that integrates SARIMA and SVR in a different manner from previous studies. Specifically, we aim to reduce the data's signal variability using EMD before analyzing the time series with the SARIMA model. Moreover, we enhance the forecasting accuracy by introducing Gaussian white noise as a variable input for the SVR model. We utilize the SVR model to capture both linear and nonlinear patterns in the time series data, with linear patterns being a special case of nonlinear patterns. SVR, known for its ability to effectively predict nonlinear time series data, improves forecasting accuracy. To evaluate the forecasting performance, three criteria are employed: MAPE, MAD, and MSE. The results indicate that the SARIMASVR4 model outperforms the other models across all four datasets. Filtering the data to reduce signal variability significantly impacts the SARIMA model's performance, while the inclusion of Gaussian white noise as an input variable

enhances the SVR model's accuracy. In conclusion, to enhance SVR model efficiency, careful selection of input variables and restructuring of data are crucial for improving forecasting accuracy.

Acknowledgement

We would like to express our sincere gratitude to the Faculty of Science at Ubon Ratchathani Rajabhat University for supporting the research with funding, as well as to the Office of Agricultural Economics at the Ministry of Agriculture and Cooperatives for providing the data used in this experiment.

References

- Abadan, S. and Shabri, A. (2014). Hybrid empirical mode decomposition-ARIMA for forecasting price of rice. *Applied Mathematical Sciences*, 8(63):3133-3143. <https://doi.org/10.12988/ams.2014.43189>
- Adaryani, F.R., Mousavi, S.J., and Jafari, F. (2022). Short-term rainfall forecasting using machine learning-based approaches of PSO-SVR, LSTM and CNN. *Journal of Hydrology*, 614(part A):128463. <https://doi.org/10.1016/j.jhydrol.2022.128463>
- AL-Musaylh, M.S., Deo, R.C., Li, Y., and Adamowski, J.F. (2018). Two-phase particle swarm optimized-support vector regression hybrid model integrated with improved empirical mode decomposition with adaptive noise for multiple-horizon electricity demand forecasting. *Applied Energy*, 217:422-439. <https://doi.org/10.1016/j.apenergy.2018.02.140>
- Amini, P. and Khashei, M. (2019). A soft intelligent allocation-based hybrid model for uncertain complex time series forecasting. *Applied Soft Computing*, 84:105736. <https://doi.org/10.1016/j.asoc.2019.105736>
- Büyüksahin, Ü.Ç. and Ertekin, Ş. (2019). Improving forecasting accuracy of time series data using a new ARIMA-ANN hybrid method and empirical mode decomposition. *Neurocomputing*, 361:151-163. <https://doi.org/10.1016/j.neucom.2019.05.099>
- Dai, Y., Yang, X., and Leng, M. (2022). Forecasting power load: A hybrid forecasting method with intelligent data processing and optimized artificial intelligence. *Technological Forecasting and Social Change*, 182:121858. <https://doi.org/10.1016/j.techfore.2022.121858>
- de Araújo Morais, L.R. and da Silva Gomes, G.S. (2022). Forecasting daily Covid-19 cases in the world with a hybrid ARIMA and neural network model. *Applied Soft Computing*, 126:109315. <https://doi.org/10.1016/j.asoc.2022.109315>
- Fan, G.-F., Wei, X., Li, Y.-T., and Hong, W.-C. (2020). Forecasting electricity consumption using a novel hybrid model. *Sustainable Cities and Society*, 61:1-16. <https://doi.org/10.1016/j.scs.2020.102320>
- Gholamy, A., Kreinovich, V., and Kosheleva, O. (2018). Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation. Departmental Technical Report (CS), Technical Report: UTEP-CS-18-09. https://scholarworks.utep.edu/cs_techrep/1209
- Guo, Y., Guo, J., Sun, B., Bai, J., and Chen, Y. (2022). A new decomposition ensemble model for stock price forecasting based on system clustering and particle swarm optimization. *Applied Soft Computing*, 130:109726. <https://doi.org/10.1016/j.asoc.2022.109726>
- Kavuncuoglu, H., Kavuncuoglu, E., Karatas, S.M., Benli, B., Sagdic, O., and Yalcin, H. (2018). Prediction of the antimicrobial activity of walnut (*Juglans regia* L.) kernel aqueous extracts using artificial neural network and multiple linear regression. *Journal of Microbiological Methods*, 148:78-86. <https://doi.org/10.1016/j.mimet.2018.04.003>
- Koondee, P., Niemsakul, J., Supeekit, T., Somboonwivat, T., and Chanpuypetch, W. (2023). Forecasting and analysing the gap between Thailand's wood pellet supply and global demand. *Engineering and Applied Science Research*, 50(2):107-120. <https://doi.org/10.14456/EASR.2023.12>
- Laouafi, A., Laouafi, F., and Boukelia, T.E. (2022). An adaptive hybrid ensemble with pattern similarity analysis and error correction for short-term load forecasting. *Applied Energy*, 322:119525. <https://doi.org/10.1016/j.apenergy.2022.119525>
- Liang, Y., Niu, D., and Hong, W.-C. (2019). Short term load forecasting based on feature extraction and improved general regression neural network model. *Energy*, 166:653-663. <https://doi.org/10.1016/j.energy.2018.10119>
- Office of Agricultural Economics. (2022). *Agricultural statistics of Thailand 2022*. Bangkok: Office of Agricultural Economics, 48p.
- Ren, F., Tian, C., Zhang, G., Li, C., and Zhai, Y. (2022). A hybrid method for power demand prediction of electric vehicles based on SARIMA and deep learning with integration of periodic features. *Energy*, 250:123738. <https://doi.org/10.1016/j.energy.2022.123738>
- Sharma, S., Sumesh, K.C., and Sirisomboon, P. (2022). Rapid ripening stage classification and dry matter prediction of durian pulp using a pushbroom near infrared hyperspectral imaging system. *Measurement*, 189:110464. <https://doi.org/10.1016/j.measurement.2021.110464>
- Vonglao, P., Somnat, K., Thepchim, S., and Sutthison, T. (2023). Enhancing accuracy in predicting Thailand's rice exports: a hybrid modeling approach. *Naresuan University Journal: Science and Technology (NUJST)*, 31(4):1-21.
- Wang, J., Wang, X., Lei, X.H., Wang, H., Zhang, X.H., You, J.J., Tan, Q.F., and Liu, X.L. (2020). Teleconnection analysis of monthly streamflow using ensemble empirical mode decomposition. *Journal of Hydrology*, 582:124411. <https://doi.org/10.1016/j.jhydrol.2019.124411>
- Xian, H. and Che, J. (2022). Unified whale optimization algorithm based multi-kernel SVR ensemble learning for wind speed forecasting. *Applied Soft Computing*, 130:109690. <https://doi.org/10.1016/j.asoc.2022.109690>
- Xiang, Y., Gou, L., He, L., Xia, S., and Wang, W. (2018). A SVR-ANN combined model based on ensemble EMD for rainfall prediction. *Applied Soft Computing*, 73:874-883. <https://doi.org/10.1016/j.asoc.2018.09.018>
- Xu, S., Chan, H.K., and Zhang, T. (2019). Forecasting the demand of the aviation industry using hybrid time series SARIMA-SVR approach. *Transportation Research Part E: Logistics and Transportation Review*, 122:169-180. <https://doi.org/10.1016/j.tre.2018.12.005>
- Yang, H.-F. and Chen, Y.-P.P. (2019). Hybrid deep learning and empirical mode decomposition model for time series applications. *Expert Systems with Applications*, 120:128-138. <https://doi.org/10.1016/j.eswa.2018.11.019>
- Yaslan, Y. (2017). Empirical mode decomposition based denoising method with support vector regression for time series prediction: A case study for electricity load forecasting. 103:52-61. <http://dx.doi.org/10.1016/j.measurement.2017.02.007>
- Yu, L., Liang, S., Chen, R., and Lai, K.K. (2022). Predicting monthly biofuel production using a hybrid ensemble forecasting methodology. *International Journal of Forecasting*, 38(1):3-20. <https://doi.org/10.1016/j.ijforecast.2019.08.014>
- Zhang, K., Cao, H., Thé, J., and Yu, H. (2022). A hybrid model for multi-step coal price forecasting using decomposition technique and deep learning algorithms. *Applied Energy*, 306(Part A):1-21. <https://doi.org/10.1016/j.apenergy.2021.118011>