JOURNAL OF STATISTICAL MODELING & ANALYTICS (JOSMA)
(ISSN: 2180-3102)
UNIVERSITI MALAYA

Performance of Chili Price Forecasting Models in Johor: A Comparative Study

Aimi Athirah Ahmad[1*], Nadiah Ruza[2], Syahrin Suhaimee[3], Hafidha Azmon[4] and Teoh Chin Chuang[5]

[1,3]*Socio-Economy, Market Intelligence and Agribusiness Research Centre, Malaysian Agricultural Research and Development Institute, Malaysia*
[2]*School of Business and Economics, Universiti Putra Malaysia*
[4,5] *Engineering Research Centre, Malaysian Agricultural Research and Development Institute, Malaysia*

**Corresponding author: aimiathirah@mardi.gov.my*

Received 30 March 2025
Accepted 05 May 2025
Published 26 Dec 2025

RESEARCH ARTICLE

Abstract

A substantial portion of household income in Malaysia is allocated to food expenditures, and chili is a staple ingredient in Malaysian cuisine. Fluctuations in chili prices directly affect the cost of living for individuals and families, impacting their purchasing power and overall well-being. Forecasting chili prices helps in effective supply chain management. Producers, distributors, and retailers can plan and adjust their operations based on anticipated price trends. This, in turn, contributes to the country's efficiency and stability of the chili supply chain. This study emphasises the importance of comparing various forecasting models to identify the most accurate predictors of chili prices. The goal is to develop a model that can contribute to more informed decision-making in crop production and market interventions, ultimately promoting stability in the chili industry and ensuring sustainable practices. Statistical models, time series forecasting models and machine learning models which include multiple linear regression (MLR), Auto Regressive Integrated Moving Average with exogenous inputs (ARIMAX), and machine learning models that consist of Support Vector Regression (SVR) were tested and compared using ex-farm prices in Johor with the duration of 5 years, starting from 2018 to 2022. This study reveals that SVR under machine learning algorithms performed best as the forecasted model followed by ARIMAX and MLR. However, ARIMAX models, an extension of the ARIMA model, effectively capture and predict patterns by incorporating significant exogenous variables. Overall, the results show that the price of fertilisers, Movement Control Order (PKP) season and chili production significantly affect the prices of chilies.

Keywords: ARIMAX; Chili prices; Forecasting; Multiple linear regression; Support vector regression

1. Introduction

Given the substantial allocation of household income towards food expenditures, assessing price volatility through forecasting is particularly crucial in developing nations. This significance arises from the direct impact of food prices' uncertainties on overall well-being. Furthermore, the repercussions of instability in food prices extend to affect individuals of low-income status and small-scale agricultural producers who rely on the sales of their crops.

As acknowledged, the price of chili exhibits inherent instability and fluctuation, posing challenges for stakeholders in arriving at consistent and reliable decisions regarding chili pricing. Additionally, chili production is susceptible to threats arising from agricultural issues, potentially leading to a decrease in supply. The consequential decline in supply and increased demand results in a corresponding rise in chili prices. Consequently, there is a pertinent need for information concerning projected fluctuations in chili price trends to ascertain market demand.

Furthermore, a comprehensive analysis of chili price volatility and the impact of shocks on chili price fluctuations, attributed to factors such as production dynamics, input prices, the pricing and quantity of imported chili, climate, the Movement Control Order (PKP), and festive demands, is imperative. This analytical pursuit aims to elucidate the multifaceted determinants influencing chili prices. Consequently, a heightened interest emerges in comparing various forecasting models, seeking to identify those that yield the most accurate forecasted values.

The envisaged outcomes include the enhancement of decision-making processes in both crop production strategies and market interventions. Adopting a robust forecasting model is anticipated to alleviate uncertainties surrounding chili prices, thereby contributing to more informed decision-making within agriculture and market interventions.

Furthermore, the model that is most suitable for assessing the fluctuation of chili price needs to be identified through data analysis approach using statistical methods and machine learning. Therefore, this study aims to predict the chili price using the Statistical Modelling and Machine Learning approach and emphasises the importance of comparing various forecasting models to identify the most accurate predictors of chili prices. This study uses chili price data in the state of Johor. The selection of Johor state is because Johor is the largest chili producer in Malaysia.

2. Literature Review

Chili prices are influenced by a combination of supply, demand, and external factors. Climate and weather conditions play a critical role, as chilies are highly sensitive to variations in temperature and rainfall. Extreme weather events such as droughts or floods can reduce yields, leading to supply shortages and price hikes. Input costs such as fertilizers, seeds, and labor also directly impact production costs, with increases in these costs often being passed on to market prices. Pest and disease outbreaks can significantly reduce crop quality and quantity, further straining supply.

From the input factors, fertilizer costs give a significant impact on the cost of production on the commodities especially for Chili. In Indonesia, studies from (Rachmaniah et al., 2022) identified that for both red chili and cayenne pepper, higher fertilizer costs negatively affect production. This is reflected in the statistically significant negative coefficients in the production equations for both chili types. Reduced production due to higher input costs leads to lower supply, which, in turn, puts upward pressure on prices, exacerbating chili price volatility. Similar findings from Alfred et al. (2022) highlighted that fertilizer prices are noted as a critical input cost that can constrain supply and indirectly affect chili prices by influencing production costs in Malaysia. Another recent study by Arndt et al. (2023) also showed that fertilizer price increases significantly impacted agricultural production costs and productivity, leading to higher food prices. Their study focuses on the impact of higher fertilizer prices caused by the Ukraine-Russia War on the agriculture commodities prices where the source of fertilizer mostly came from Russia.

On the demand side, seasonal and cultural consumption patterns drive fluctuations, particularly during festivals or periods of high culinary demand. Additionally, international trade dynamics, including export restrictions, import tariffs, or changes in global production, affect local market prices. Lastly, transportation and storage costs, coupled with market infrastructure inefficiencies, influence

final chili prices. Together, these factors create a complex interplay that determines the pricing of chilies in the market.

Monitoring the volatility of commodity prices, especially agricultural commodities, can be essential in evaluating the country's economic performance. Commodity price forecasts can help the government to make and develop appropriate economic policies and strategies in the future. Therefore, price forecasting is an alternative approach for reducing the negative effects of uncertainty, which can further decrease the risk of the producers' agricultural commodities.

Shahizan et al. (2023) investigated the effectiveness of the seasonal regression and the quadratic trend models with seasonal indices in predicting the price of red chili in Johor, Malaysia, using historical price data from 2018 to 2022. The results revealed that the seasonal regression model outperformed the quadratic trend model, with seasonal indices predicting red chili prices. Nguyen et al. (2022) mentioned that red chili is an agricultural commodity with high price volatility. Their research aimed to analyze the price volatility of red chili in Semarang Regency from January 2019 to February 2020. They applied the ARCH-GARCH method, showing that the price volatility of red chili occurred at the beginning, middle, and end of the year due to climate change, changes in public consumption patterns on religious holidays, and oversupply.

On the other hand, (Basnayake et al., 2022) forecasted the prices of green chili peppers in Sri Lanka using artificial neural networks. The Time Delay Neural Network (TDNN), Feedforward Neural Network (FFNN) with Levenberg-Marquardt (LM) algorithm, and FFNN with Scaled Conjugate Gradient (SCG) algorithm were employed on weekly average retail prices of green chili in Sri Lanka from the 1st week of January 2011 to the 4th week of December 2018. Based on the Mean Squared Error (MSE), Mean Absolute Error (MAE), and Normalized Mean Squared Error (NMSE), FFNN with the LM algorithm gave the best performance compared to other methods.

In general, fluctuations in the price of agricultural commodities occur mainly due to shocks in supply. These disturbances, combined with demand elasticity and short-term supply, cause sudden price instability, which can cause farmers and consumers to experience uncertainty, risk, and commodity price fluctuations. Pratiwi et al. (2020) used the time series data price of large red and curly red chili from July 2016 to October 2019 in Yogyakarta, Indonesia. They forecasted chili prices using Auto-regressive Integrated Moving Average (ARIMA) for 12 periods, beginning in November 2019 and ending in October 2020. Putriasari et al. (2022) forecasted a weekly red chili price in Bengkulu City using the ARIMA and Singular Spectrum Analysis (SSA) Methods. They found that ARIMA performed better than SSA.

ARIMA was observed to be the most accurate forecasting model for curly red chili prices in Indonesia (Sukiyono & Janah, 2019). Their study applied five forecasting models: Moving Average, Single Exponential Smoothing, Double Exponential Smoothing, Decomposition, and ARIMA. Despite focusing on univariate ARIMA, external regressors are recommended to improve the accuracy of the forecasting model. Hamjah (2014) developed the best Box-Jenkins Auto-Regressive Integrated Moving Average with external regressor, that is, the ARIMAX model, for measuring the temperature and rain-fall effects on major spice crops productions in Bangladesh and forecasting the output using the same model. Their study found that ARIMAX (2,1,2), ARIMAX (2,0,1), and ARIMAX (2,1,1) are the best models for chili, garlic, and ginger crops, respectively. The finding contributes to agricultural forecasting by emphasizing the importance of comparing different models for price prediction and considering multiple crops in the analysis.

3. Methodology

3.1 Data collection

The prices of chilies were obtained from FAMA based on the ex-farm prices. The prices are taken monthly with a duration of 5 years, starting from 2018 to 2022. Federal Agricultural Marketing Authority (FAMA) is a governmental authority under the Ministry of Agriculture and Food Industries that oversees the marketing of agricultural products. The life span of chili is between 6 and 12 months, while the maturity time is between 60 and 120 days. Table 1 lists all the variables in the analysis.

Table 1. Predictor variables

Type of variables	Notation	Sources
Monthly production (kg)	Prod	Department of Agriculture (DOA), Malaysia
Fertlisers price (RM)	DAP	(Index Mundi, 2023)
	Potassium Chloride	
	Phosphate Rock	
	Super Phosphates	
	Urea	
Climate	Temperature	https://power.larc.nasa.gov/data-access-viewer/
	Rainfall	
Season	Movement control order (PKP)	National Security Council (MKN), Malaysia
	Festive	Author's own calculation

3.2 Multiple Linear Regression

A multiple linear regression model is an extension of the simple linear regression model for data with multiple predictor variables and one outcome, $(x_1, x, \dots, x_n)$, where n is the number of observations. It formalizes a simultaneous statistical relation between the single continuous outcome Y and the predictor variables, $X_1, X_n, \dots, X_n$.

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon_i \, , \, \varepsilon_i \sim N(0, \sigma^2) \quad (1)$$

where β_0 represents the intercept, the mean of Y when the predictor variables $X_1, X_n, \dots, X_n = 0$, and $\beta_1, \beta_2, \dots, \beta_n$ represents a slope with respect to $X_1, X_n, \dots, X_n$. The assumptions are thus the same as for simple linear regression:

I. y_i are independent of each other;
II. y_i follows a normal distribution;
III. mean of that distribution is a linear function of $x_1, x, \dots, x_n$;
IV. variance of that distribution is the same for all y (constant variance, or homoscedasticity).

3.3 ARIMAX Modeling

ARIMAX is a statistical modeling technique used in time series analysis and forecasting. It is an extension of ARIMA (Auto-Regressive Integrated Moving Average) model, which is designed to capture and predict patterns in time series data. The only difference between ARIMA and ARIMAX is

the addition of an exogeneous (external) variable. The ARIMA model works on a single time series data (univariate) whereas ARIMAX uses multiple variables to include the external feature.

Implementing an ARIMAX model involves several key steps. First, thorough data preparation is crucial, which includes addressing missing values, managing outliers, and ensuring that the exogenous variables are relevant and correctly aligned with the dependent variable. Stationarity testing follows, helping to determine whether differencing is necessary. Common tests for stationarity include the Augmented Dickey-Fuller (ADF) test and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test. Once stationarity is established, the model identification process involves selecting appropriate orders for the AR, I, and MA components, often guided by Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots. When incorporating exogenous variables, it is important to account for their lagged effects and address potential multicollinearity.

Parameter estimation in ARIMAX models is typically performed using Maximum Likelihood Estimation (MLE) or other optimization techniques to achieve the best fit for the observed data. Model diagnostics are then conducted to validate the model's assumptions and suitability. These include analyzing residuals to confirm they exhibit white noise characteristics and using tests like the Ljung-Box test to detect autocorrelation. Once the model is validated, it can be applied to forecast future values, with predictions incorporating anticipated values of the exogenous variables. The Ljung-Box test the magnitudes of the residuals autocorrelation for significance:

H_0: The data are independently distributed with no serial correlation
H_1: The data are not independently distributed exhibit serial correlation

The Ljung-Box test can be calculated as:

$$Q = n(n+2)\sum_{k=1}^{h}\frac{\hat{p}_k^2}{h-k} \tag{2}$$

where n is the sample size, $\hat{p}_k^2$ is the sample autocorrelation at lag k and h is the number of lags being tested. Under H_0 the statistic Q follows a chi-square distribution, $\chi^2(k-p-q)$.

ARIMAX $(m, n, k+p)$ stands for Autoregressive Integrated Moving Average process with exogenous input. The model can be written as:

$$\begin{aligned} y(t) = a_1 y(t-1) + \cdots + a_m y(t-m) + e(t) + \cdots + c_n e(t-n) + b_0 u(t-k) \\ + b_1 u(t-k-1) + b_p u(t-k-p) \ldots + \beta_n X_n + \varepsilon_i \end{aligned} \tag{3}$$

$e(t) \sim W_n(\mu, \lambda^2)$ where $e(t)$ is white noise (with mean μ and variance λ^2), $c_0, c_1, \ldots c_n \in R$ are the MA model's parameters, $a_1, \ldots a_m \in R$ are the AR model's parameters, n is the order of the MA portion, m is the order of AR portion, $u(t)$ is the exogenous input, $b_0, b_1, \ldots, b_p$ are the input portion parameters, p is the order of the exogenous portion, and k is the positive input delay.

If we zeta transforms the process in the domain,

$$Y(z) = (a_1 z^{-1} + \cdots + a_m z^{-m})Y(z) + (1 + c_1 z^{-1} + \cdots + c_n z^{-n})E(z) + \left(b_0 + b_1 z^{-1} + \cdots + b_p z^{-p}\right)z^{-k}U(z)$$

$$Y(z) = \frac{z^{-n}}{z^{-m}} \frac{z^n + c_1 z^{n-1} + \cdots + c_n}{z^m - a_1 z^{m-1} - \cdots - a_m} E(z) + \frac{z^{-p}}{z^{-m}} \frac{b_0 z^p + b_1 z^{p-1} + \cdots + b_p}{z^m + a_1 z^{m-1} + \cdots + a_m} z^{-k} U(z)$$

$$Y(z) = z^{m-n} \frac{z^n + c_1 z^{n-1} + \cdots + c_n}{z^m - a_1 z^{m-1} - \cdots - a_m} E(z) + z^{m-p-k} \frac{b_0 z^p + b_1 z^{p-1} + \cdots + b_p}{z^m + a_1 z^{m-1} + \cdots + a_m} U(z)$$

$$Y(z) = W(z)E(z) + G(z)U(z) \tag{4}$$

where z represents the complex frequency variable in the zeta-transform domain. $W(z)$ and $G(z)$ are asymptotically stable if their poles (roots of the denominator) are inside the unit circle. So, the stationarity of an ARIMAX depends on the inputs as well.

Let $y(t)$be an ARIMAX process. Then $y(t)$ is stationary if the poles $W(z)$ are inside the unit circle $u(t)$ is stationary, this happens when $u(t) = U$ constant for all t. In general, an ARIMAX process is not stationary, however, its stochastic portion (ARMA) is stationary. The nonstationary derives from the input $u(t)$.

An equivalent form of representing an ARIMAX is:

$$y(t) = \frac{C(z)}{A(z)} e(t) + \frac{B(z)}{A(z)} z^{-k} u(t), \qquad e(t) \sim w_n(\mu, \lambda^2) \tag{5}$$

where

$$C(z) = 1 + c_1 z^{-1} + \cdots + c_n z^{-n} \tag{6}$$

$$A(z) = 1 - a_1 z^{-1} - \cdots - a_m z^{-m} \tag{7}$$

$$B(z) = b_0 + b_1 z^{-1} + \cdots + b_p z^{-p} \tag{8}$$

Note that applying the superposition principle

$$y_e(t) = \frac{C(z)}{A(z)} e(t)$$

$$y_u(t) = \frac{B(z)}{A(z)} z^{-k} u(t)$$

$$\therefore y(t) = y_e(t) + y_u(t) \tag{9}$$

We can analyze $y_e(t)$ as an ARMA process and $y_u(t)$ as a dynamic system.

3.4 Support Vector Regression

Support Vector Regression (SVR) is a supervised machine learning method developed for regression tasks (Drucker et al., 1997; Vapnik & Lerner, 1963). This approach is suitable for analysing the relationship between a dependent variable and one or more predictor variables. SVR solves this task by

formulating an optimisation problem to learn a regression function that maps the input predictor variables to the observed response values. This technique is particularly effective as it strikes a balance between model complexity and prediction error and performs strongly on high-dimensional data. SVR is an extension of the Support Vector Machine (SVM) classification algorithm (Boser et al., 1992). Originally proposed by Vapnik & Chervonenkis (1964), SVM can handle both classification and regression tasks by estimating a variable based on the behaviour of certain explanatory variables.

In contrast to SVM classification, which provides binary results, SVR deals with regression problems and enables the estimation of real-valued functions. SVR uses the basic concept of SVM — a sparse kernel machine that performs classification using a hyperplane defined by a few support vectors. Consequently, optimisation in SVR is expressed in terms of support vectors, making the solution dependent on the number of support vectors rather than the dimension of the input data.

SVR offers several advantages over other regression methods. By using a kernel, SVR effectively handles nonlinear regression problems by projecting the original features into a kernel space where the data can be linearly separated (Ben-Hur et al., 2008). Another advantage of SVR is that it develops a model that describes the importance of the variables in characterising the relationship between input and output, unlike traditional regression methods that often require the assumption of a potentially inaccurate model (Noh et al., 2024; Orru et al., 2012).

In SVM classification, each labelled sample in a training dataset is treated as a data point in a multidimensional feature space. A hyperplane is then calculated to correctly classify as many training samples as possible. New samples are classified based on their position relative to this hyperplane in the feature space. The optimisation process aims to maximise the distance between the support vectors — the data points closest to the hyperplane. For regression tasks, instead of identifying a hyperplane that separates the training samples, an ε-insensitive loss function is introduced to find a hyperplane where the predicted response values of the training samples deviate by at most ε from their actual values. This forms an ε-insensitive tube (or band) that is used to calculate the generalisation limits for the regression.

The optimisation in SVR minimises the ε-insensitive tube so that it is as flat (narrow) as possible while encompassing most of the training patterns. Consequently, the hyperplane is defined by some support vectors — training samples outside the ε-insensitive tube. The SVR training process results in a regression model that is used to predict the response output for new samples.

Consider a dataset $x\text{-}i.$, $y\text{-}i..|i=1,2,\ldots,m.$, where, $x\text{-}i.$ and, $y\text{-}i. \in \mathbb{R}$. Here, $x\text{-}i.$ represents the features, and, $y\text{-}i.$ represents the labels, both normalized to (+1,−1). The classifier categorises the data based on the labels, which requires the creation of a hyperplane to separate the data. Given the nature of the data, many hyperplanes can be created within the same dataset. The selected hyperplane should maximise the space defined by the closest points to the space, called support vectors.

Thus, Lestari et al. (2022) formulated the SVM as follows:

$$minimize\ \omega, b, \xi = \frac{1}{2}\omega^T\omega + C\sum_{i=1}^{m}\xi_i \tag{10}$$

subjected to

$$y_i(\omega^T x_i + b) \geq 1 - \xi_i, \qquad \xi_i \geq 0, \qquad for\ 1 \leq i \leq m$$

The margin width is represented by ω, while b denotes the bias. The slack variable ξ is associated with the soft SVM, permitting some values to fall within the margin. C indicates the trade-off between margin width and misclassification tolerance.

3.5 Forecasting accuracy

In comparing the model efficiencies of several selected forecasting models, several criteria, such as Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Root Mean Square Error (RMSE), need to be evaluated. The smaller the values of the three criteria, the better the forecasting model used (Ahmad et al., 2018). The following equation describes the criteria used:

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \widehat{y_t}|}{n} \tag{11}$$

$$MAPE = \frac{\sum_{i=1}^{n} \frac{|y_i - \widehat{y_t}|}{y_i}}{n} \times 100\% \tag{12}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} \frac{(y_i - \widehat{y_t})^2}{y_i}}{n}} \tag{13}$$

All analyses performed in this study were using R Studio software.

4. Result and discussion

4.1 Trend analysis

The monthly chili prices from January 2018 to December 2022 are plotted as shown in Figure 1. The series exhibits cyclical effects, and no long-term linear trend or seasonality is identified from the plot. The stationarity assumption of the time series is tested by the Augmented Dickey-Fuller (ADF) and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) tests. Although the ADF (*p*-value=0.14) accepts the null hypothesis, the KPSS (*p*-value= 0.1) confirms the time series has stationary trend.

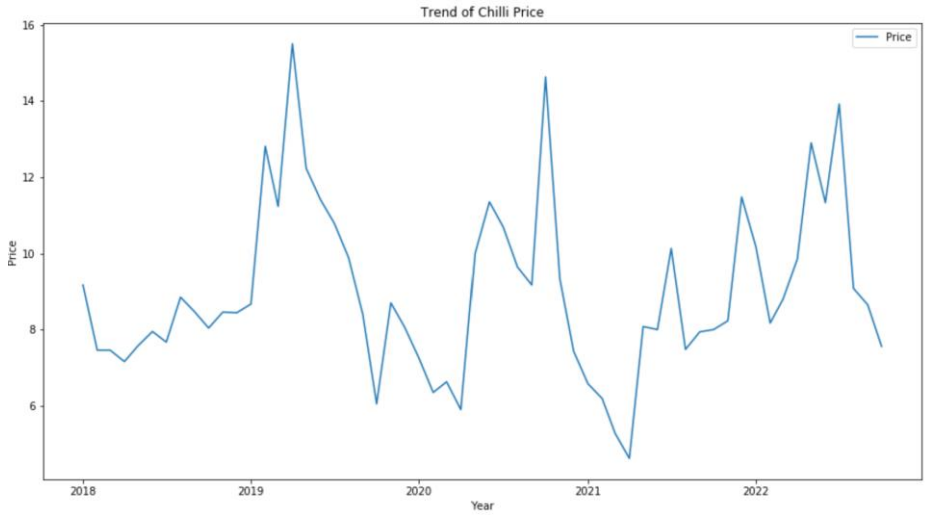

Figure 1. Trend of Chili Price

The development of chili prices in Malaysia from 2018 to 2022 shows a high degree of volatility with several significant peaks and troughs. The highest prices were observed around mid-2019 and mid-2021, with prices reaching around 15 and 14 units respectively. The lowest prices were observed around mid-2020, falling to around 5 units. Overall, the price trend depicts a pattern of sharp increases followed by significant decreases, indicating possible influences of factors such as seasonal changes, fluctuations

in market demand and supply, and other external economic or climatic conditions that affect chili production and pricing.

Between 2021 and 2023, chili prices in Malaysia surged due to a combination of adverse weather, labor shortages, and rising costs. Unusually heavy and prolonged rains in key farming areas, such as Cameron Highlands and Johor, damaged crops and reduced yields by up to 30%. Pandemic-related restrictions led to labor shortages, disrupting farming operations, while global supply chain issues drove up input costs for fertilizers and pesticides, which were passed on to consumers. Additionally, inflation and increased demand, particularly during festive periods and the post-pandemic reopening of the food service industry, further intensified price pressures.

In general, the dynamics of supply and demand on the market further exacerbate price fluctuations. Changes in consumer preferences, dietary trends and population growth can lead to fluctuating demand, while export and import policies influence the availability of chilies in the domestic market. Economic factors such as inflation, currency fluctuations and labour costs affect production costs and consequently prices. Government policies, including subsidies, support programmes and regulatory changes, also influence the agricultural sector and price stability. In addition, global market trends and supply chain disruptions, such as during the COVID-19 pandemic, affect domestic chili prices by altering availability and logistics. These intertwined factors combine to drive the volatility of chili prices, making them subject to rapid and unpredictable change.

4.2 Multiple Linear Regression

In general, the inclusion of lag variables in a time series model is a remedial approach to the autocorrelation problem presented in the residuals. In this study, the lag-predictor variables are included in the modeling. The visual and numerical assessments supported the intuitive approach based on the red chili cultivation experience, which shows that planting chilies are not seasonal. The harvesting could be done when chili plants reach three months old in addition to the effects of the fertilizer components added by the rainfall and temperature from the surrounding area.

Multiple Linear Regression (MLR) is implemented to determine the factors that significantly influence the price of chili. The model with the highest adjusted R^2 is chosen as the best model. The results are presented in Table 2. In this case, Model 3 is chosen as the best model, indicating that chili prices and fertilizer components show significant relationships at least at lag 2. It can be summarized that the variables that significantly contribute to the chili price are Production, PKP versus non-PKP period, and the price of fertilizer, namely DAP, Potassium Chloride, Phosphate Rock, and Super Phosphate.

Table 2. Multiple Linear Regression based on Analysis of Variance (ANOVA) for y = chili price

Model	**Predictors**	p **−value**	R^2	**Adjusted** R^2
1	x_1 = Production x_2 = Fertilizer_Potassium Chloride x_3 = Fertilizer_Phosphate Rock x_4 = Fertilizer_Super Phosphate x_5 = PKP	0.003	45.4%	30.5%
2	x_1 = Production x_2 = Lag 1_Fertilizer_Phosphate Rock x_3 = Lag 1_Fertilizer_Super Phosphate	0.003	45.7%	30.9%
3	x_1 **= Production** x_2 **= Lag 2_Fertilizer_Potassium Chloride**	**0.002**	**46.8%**	**32.3%**

	x_3 = **Lag 2_Fertilizer_Phosphate Rock** x_4 = **Lag 2_Fertilizer_Super Phosphate** x_5 = **Lag 2_Fertilizer_DAP** x_6 = **PKP**			
4	x_1 = Production x_2 = Lag 3_Fertilizer_Potassium Chloride x_3 = Lag 3_Fertilizer_Phosphate Rock x_4 = Lag 3_Fertilizer_Super Phosphate x_5 = Lag 3_Fertilizer_DAP	0.003	45.7%	30.9%

4.3 ARIMAX

ARIMAX model, or regression with ARIMA errors, generally conducts a time series regression with covariates. The remainder (error values) between the observed y_t and the fitted $\hat{y}_t$ from the time series, regression is then remodelled using the ARIMA model, as shown in Figure 2.

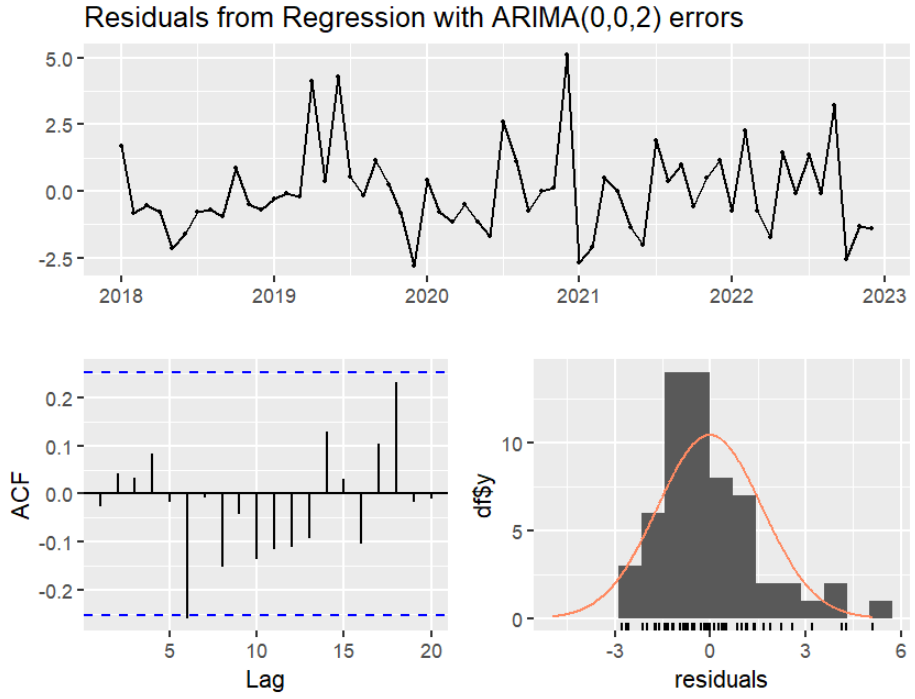

Figure 2. Residuals analysis with ARIMA (0,0,2)

Based on the MLR approaches, the selected factors, i.e. fertilizers, production, and PKP season, act as the regressor, and the remainder terms are best fitted with the MA (2) model, hence ARIMAX (0, 0, 2) is the best model. The fitted model is written as:

$$\hat{Y}(t) = 9.8821 + \varepsilon_t + 0.4487\varepsilon_{t-1} + 0.4232\varepsilon_{t-2} - 0.02297P_t - 0.7888K_{tj} + 0.0064F_{1t} + 0.0006F_{2t} + 0.0027F_{3t} - 0.0084F_{4t} \quad (14)$$

where, the terms $0.4487\varepsilon_{t-1} + 0.4232\varepsilon_{t-2}$ indicate an MA(2) process with lag 2 and P_t K_t and F_t represent exogenous variables which are production, PKP period and fertilizers price (Potassium Chloride, Phosphate Rock, Super Phosphate and DAP).

Production negatively affects the chili price and further supports the findings from MLR. Similarly, the effect of the fertilizer on chili prices is similar to those found in MLR. Based on Figure

2, the Ljung-Box diagnostic test yields a p-value of 0.4002, which is greater than 0.05, supporting the fact that the ARIMAX (0, 0, 2) has no autocorrelation and can be considered adequate. The comparison between observed and fitted values using ARIMAX (0,0,2) is displayed in Figure 3.

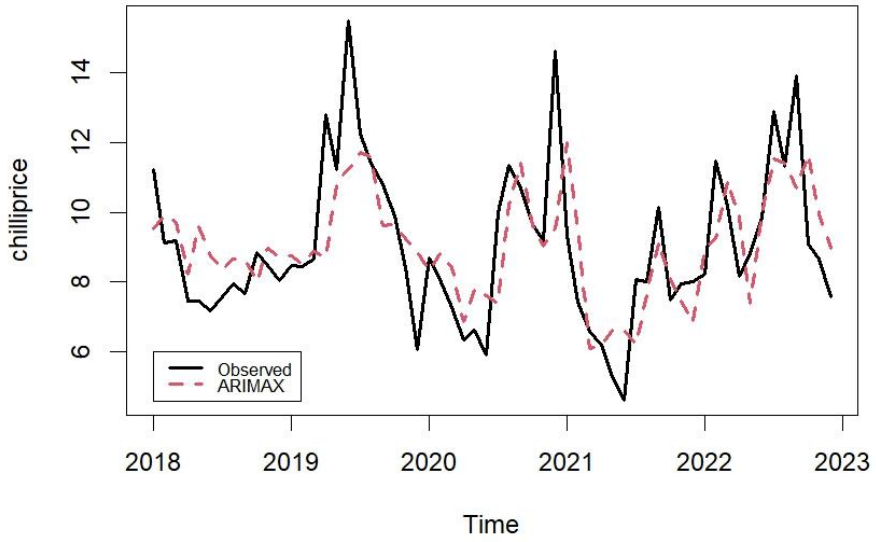

Figure 3. Observed vs Fitted Chili Price with ARIMAX (0,0,2) model

4.4 Support Vector Regression

Figure 4 illustrates the observed versus fitted chili prices in Malaysia from 2018 to 2022 using a Support Vector Regression (SVR) model. The red line represents the observed chili prices, while the blue line indicates the fitted values determined by the SVR model. The graph shows that the SVR model captures the general trend and seasonal fluctuations in chili prices over the years. Despite some deviations, especially during periods of sharp price spikes and dips, the fitted values generally follow the observed prices, indicating that the SVR model is effective in predicting the general trend of chili prices.

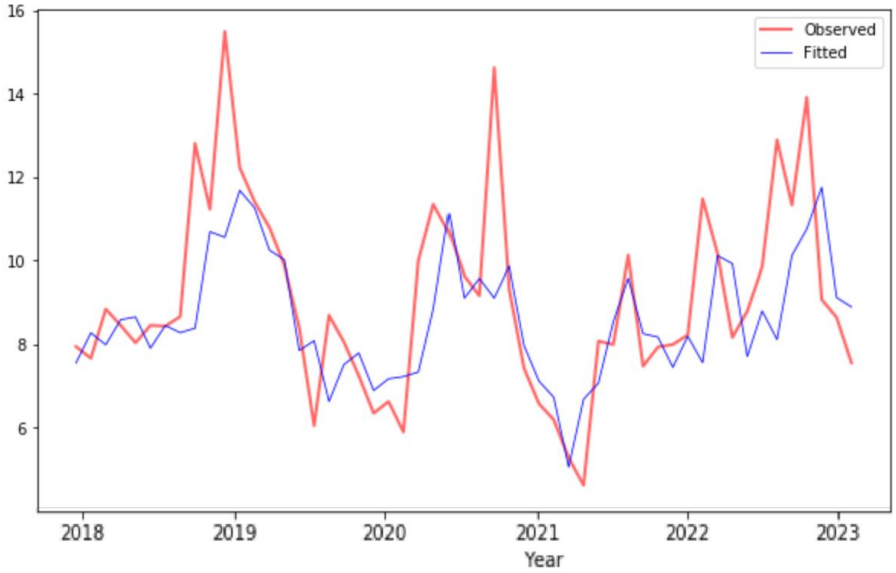

Figure 4. Observed vs Fitted Chili Price with SVR model

A closer examination of the figure reveals certain key insights. In times of significant price volatility, such as in 2019 and 2022, the observed chili prices show pronounced peaks and troughs. The

SVR model succeeds in approximating these extreme fluctuations, albeit with a certain delay and less accuracy. This suggests that while the SVR model is robust in capturing the general trend and seasonal patterns, it may have its limitations in accurately predicting sudden, sharp price changes. These deviations highlight potential areas where the model can be improved, such as the inclusion of additional external factors or the use of hybrid models to improve forecast accuracy in volatile periods.

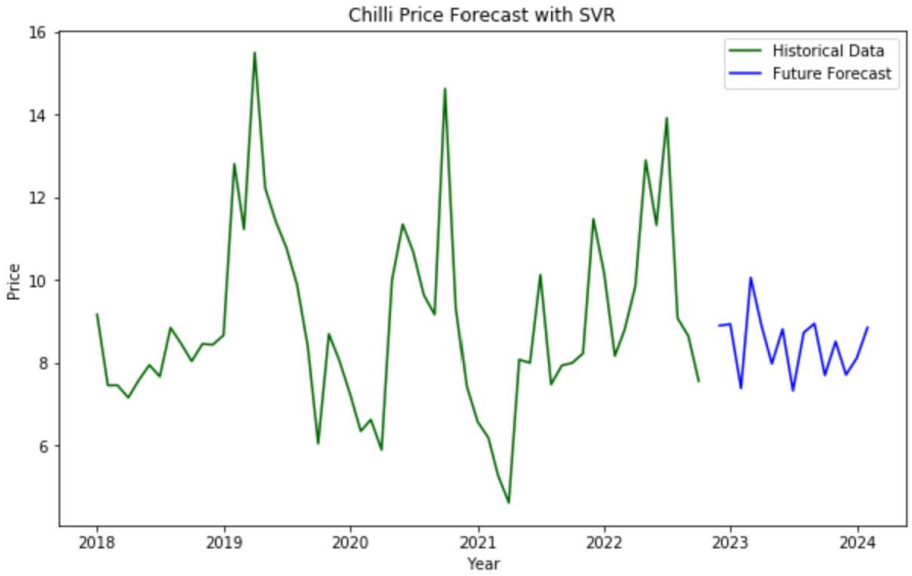

Figure 5. Chili Price Forecast with SVR

Figure 5 presents the forecast of chili prices in Malaysia until 2024 using the SVR model. The green line represents the historical data from 2018 to 2022, while the blue line shows the forecast prices from 2023 to 2024. The historical data shows significant volatility, with several sharp peaks and troughs indicating fluctuations in chili prices over the years. The forecast data indicates that chili prices will continue to fluctuate, although the amplitude of these changes appears to be less pronounced compared to the historical data.

One notable insight from the forecast is the relative stability of the predicted prices, which could indicate that the SVR model expects less extreme price movements in the near future. This could be due to several factors, including possible improvements in supply chain stability, agricultural practises or external economic conditions affecting the chili market. In addition, the forecast suggests that prices will still fluctuate but within a smaller range, indicating a possible trend towards market stabilisation. This information could be valuable for players in the chili industry, such as farmers, traders and policy makers, to make informed decisions on production and pricing strategies.

4.5 Model Comparison

Table 3 presents the accuracy measures of the ARIMAX and SVR models and compares their performance using the Root Mean Square Error (RMSE), the Mean Absolute Error (MAE) and the Mean Absolute Percentage Error (MAPE). The results show that the ARIMAX model has an RMSE of 1.6332, a MAE of 1.2263 and a MAPE of 13.7654. On the other hand, the SVR model has an RMSE of 1.6757, an MAE of 1.1052 and a MAPE of 12.1057. Although the SVR model has a slightly higher RMSE value, it demonstrates lower MAE and MAPE values compared to the ARIMAX model. This indicates that although the SVR model may have marginally higher overall error, it tends to produce more consistent and accurate predictions in terms of absolute and percentage errors.

All tested models were compared and analyzed. Standard ML models such as Support Vector Regression (SVR) exhibit good model fitting for small-scale data. However, there is a less significant difference between the performances of the top three models, as shown in Table 3. Due to limited training data for machine learning algorithms, it is recommended to employ ARIMAX in forecasting chili prices.

Table 3. Error measurements using MLR, ARIMAX and SVR

Measures	MLR	ARIMAX	SVR
RMSE	1.9244	1.6332	1.6757
MAE	1.4192	1.2263	1.1052
MAPE	16.2779	13.7654	12.1057
Rank	3	2	1

Referring to Table 3, ARIMAX demonstrates superior statistical modeling compared to Multiple Linear Regression (MLR). The ARIMAX models can capture the patterns and follow the trend effectively by incorporating significant exogenous variables.

Research comparing SVR (Support Vector Regression) and ARIMAX (Autoregressive Integrated Moving Average with Exogenous Variables) highlights their distinct performance characteristics and suitability for different types of data. SVR, a machine learning model, often outperforms ARIMAX in capturing complex non-linear patterns, particularly in datasets with volatility or limited size. For example, a study on forecasting Indonesia's Consumer Price Index (CPI) demonstrated that SVR achieved lower RMSE and MAPE values compared to ARIMAX. This advantage is attributed to SVR's ability to handle multi-dimensional data (Ghofur et al., 2022).

Conversely, ARIMAX, which is grounded in statistical time series modeling, excels when the relationships between variables are well-understood and mostly linear. It effectively incorporates exogenous variables for predictive insights but struggles with high-frequency non-linear fluctuations. Studies, such as those comparing emissions forecasts in Thailand, have noted ARIMAX's limitations under volatile conditions, where ML models like SVR or even ANN (Artificial Neural Networks) performed better (Janhuaton et al., 2024). While ARIMAX is suitable for traditional, linear trends, SVR's flexibility makes it preferable for complex, non-linear datasets, especially those influenced by dynamic, external factors.

5. Conclusion

The lower MAE and MAPE values of SVR suggest its advanced predictive capabilities than traditional models such as ARIMAX and MLR. These measures emphasise the effectiveness of the SVR model in capturing the underlying patterns and dynamics of asset returns, making it a more reliable forecasting tool. ARIMAX demonstrates superior statistical modeling compared to Multiple Linear Regression (MLR). The ARIMAX models effectively capture and predict patterns by incorporating significant exogenous variables. The trade-off between accuracy and interpretability is an important consideration for both approaches (statistical and machine learning modeling). While SVR models lack interpretability but excel in accuracy, the statistical approach provides good interpretability insight. Hence, SVR models require a model diagnostic approach to include interpretability insight. In conclusion, forecasting chili prices in Malaysia is essential for ensuring food security, supporting the agricultural sector, promoting economic stability, and facilitating informed decision-making across the entire chili supply chain.

6. Acknowledgements

The authors would like to express their gratitude to the Ministry of Agriculture and Food Security, Malaysia (KPKM) and the Malaysian Agriculture Research and Development Institute (MARDI) for providing the resources and facilities required for this study (grant number: KRE-243). The authors also extend their thanks to the MMISG 2023 contributors for their valuable suggestions and technical support, which significantly contributed to this research.

7. References

Ghofur, A., Al, F., Dewi, Y. S., & Anggraeni, D. (2022). Comparison of Support Vector Regression and Autoregressive Integrated Moving Average with Exogenous Variable on Indonesia Consumer Price Index. *SAR Journal*, 5(3): 144-148. https://doi.org/10.18421/SAR53-05

Ahmad, A. A., Jamaluddin, J. A., Yusof, F., Safari, S., & Mohd Yusof, R. (2018). Maximizing the Benefit of Domestic and Export Markets Scenario: Predicting Models for Durian Production. *Economic and Technology Management Review*, 13.

Alfred, R., Leikson, C., Boniface, B., Tanakinjal, G. H., Kamu, A., Kogid, M., & Andrias, R. M. (2022). Modelling and Forecasting Fresh Agro-food Commodity Consumption Per Capita in Malaysia using Machine Learning. *Mobile Information Systems*, 2022(1), 6106557.

Arndt, C., Diao, X., Dorosh, P., Pauw, K., & Thurlow, J. (2023). The Ukraine War and Rising Commodity Prices: Implications for Developing Countries. *Global Food Security*, 36: 100680.

Basnayake, B. R. P. M., Kaushalya, K. D., Wickaramarathne, R. H. M., Kushan, M. A. K., & Chandrasekara, N. C. (2022). An Approach for Prediction of Weekly Prices of Green Chili in Sri Lanka: Application of Artificial Neural Network Techniques. *Journal of Agricultural Sciences–Sri Lanka*, 17(2).

Ben-Hur, A., Ong, C. S., Sonnenburg, S., Schölkopf, B., & Rätsch, G. (2008). Support Vector Machines and Kernels for Computational Biology. *PLoS Computational Biology*, 4(10): e1000173.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A Training Algorithm For Optimal Margin Classifiers. *In Proceedings of The Fifth Annual Workshop on Computational Learning Theory* (pp. 144-152). ACM.

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A. J., & Vapnik, V. (1997). Support Vector Regression Machines. In M. C. Mozer, M. I. Jordan, & T. Petsche (Eds.*), Advances in Neural Information Processing Systems* (pp. 155-161). MIT Press.

Hamjah, M. A. (2014). Temperature and Rainfall Effects on Spice Crops Production and Forecasting the Production in Bangladesh: An Application of Box-Jenkins ARIMAX Model. *Mathematical Theory and Modelling*, 4(10): 149-159.

Janhuaton, T., Ratanavaraha, V., & Jomnonkwao, S. (2024). Forecasting Thailand's Transportation CO2 Emissions: A Comparison Among Artificial Intelligent Models. *Forecasting*, 6(2): 462-484. https://doi.org/10.3390/forecast6020026

Lestari, E. P., Prajanti, S. D. W., Wibawanto, W., & Adzim, F. (2022). ARCH-GARCH Analysis: An Approach to Determine the Price Volatility of Red Chili. AGRARIS*: Journal of Agribusiness and Rural Development Research*, 8(1): 90-105.

Nguyen, T. T., Nguyen, D. D., Nguyen, S. D., Prakash, I., Van Tran, P., & Pham, B. T. (2022). Forecasting Construction Price Index using Artificial Intelligence Models: Support Vector Machines and Radial Basis Function Neural Network. *Journal of Science and Transport Technology*, 9-19.

Noh, H., Son, G., Kim, D., & Park, Y. S. (2024). H-ADCP-Based Real-Time Sediment Load Monitoring System using Support Vector Regression Calibrated by Global Optimization Technique and Its Applications. *Advances in Water Resources*, 185: 104636.

Orru, G., Pettersson-Yeo, W., Marquand, A. F., Sartori, G., & Mechelli, A. (2012). Using Support Vector Machine to Identify Imaging Biomarkers of Neurological and Psychiatric Disease: A Critical Review. *Neuroscience and Biobehavioral Reviews*, 36(4): 1140-1152.

Pratiwi, L. F. L., Rosyid, A., & Hasyim, A. (2020). Forecasting of Chili Prices in the Special Region of Yogyakarta, Indonesia based on Harga Pangan applications (ARIMA approach). *In: Proceedings The 4th International Conference on Green Agro-Industry*.

Putriasari, N., Nugroho, S., Rachmawati, R., Agwil, W., & Sitohang, Y. O. (2022). Forecasting A Weekly Red Chili Price in Bengkulu City using Autoregressive Integrated Moving Average (ARIMA) and Singular Spectrum Analysis (SSA) Methods. JSDS: *Journal of Statistics and Data Science*, 1(1). https://ejournal.unib.ac.id/index.php/jsds/index

Rachmaniah, M., Suroso, A. I., Syukur, M., & Hermadi, I. (2022). Supply and Demand Model for Chili Enterprise System using a Simultaneous Equations System. *Economies*, 10(12): 312. https://doi.org/10.3390/economies10120312

Shahizan, S., Balasubramaniam, K., Bakar, N. A., & Masrom, M. (2023). Price Forecasting Analysis of Red Chili using Seasonal Regression and Quadratic Trend Models. *International Journal of Advanced Research in Technology and Innovation*, 5(3): 21-31. http://myjms.mohe.gov.my/index.php/ijarti

Sukiyono, K., & Janah, M. (2019). Forecasting Model Selection of Curly Red Chili Price at Retail Level. *Indonesian Journal of Agricultural Research*, 2(1): 1-12.

Vapnik, V. N., & Chervonenkis, A. Y. (1964). A Class of Algorithms for Pattern Recognition Learning. *Avtomat. i Telemekh*, 25(6): 937-945.

Vapnik, V. N., & Lerner, A. Y. (1963). Recognition of Patterns with Help of Generalized Portraits. *Avtomat. i Telemekh*, 24(6): 774-780.