MJFAS
Malaysian Journal of Fundamental and Applied Sciences

Mohd Shafri *et al.* | Malaysian Journal of Fundamental and Applied Sciences, Vol. 21 (2025) 1750-1763

RESEARCH ARTICLE

Enhancing Paddy Production in Malaysia: A Comparative Analysis of Multiple Regression with External Factors

Ahmad Syakir Mohd Shafri, Siti Rohani Mohd Nor*, Siti Mariam Norrulashikin

Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia

Abstract Rice is a basic food that is consumed by almost half of the world's population, especially in Malaysia. Unfortunately, paddy productivity has recently decreased dramatically, which has hampered local rice supply and forced Malaysia to depend on rice from neighbouring nations. Therefore, the modelling of paddy production is important because it provides an insight that ensure a sufficient supply of a locally produced paddy. In this study, several multiple regression models were used to predict paddy production model in Malaysia from 1980 to 2022. The multiple regression models are then compared, and the best model is determined. The regression model used in this study are Multiple Linear Regression (MLR), Multivariate Adaptive Regression Spline (MARS) and Support Vector Regression (SVR). The developed models will be evaluated using RMSE, MAPE and Ljung-Box Test. These are excellent tools for gauging the precision of the fitted model, thus can be used to evaluate the model. Based on the study, the MARS model is better at modelling Malaysian paddy production from 1980 to 2022 since it has the smallest number of measurement errors. Therefore, MARS model is the best regression model to be used to model paddy production in Malaysia compared to MLR and SVR model. Since all the models have autocorrelation, a more effective approach and model can be presented to overcome the autocorrelation issue in the future.

Keywords: Forecasting, MARS, MLR, paddy production, SVR.

Introduction

The main food source eaten by about half of the world's population is rice. Early in the 1960s, small-scale farming gave rise to Malaysia's production of rice and paddy, which eventually took off and became the country's most important food crop. Throughout the years, the Malaysian rice production system has encountered many difficulties, including severe weather patterns, insufficient soil fertility and nutrient management, a lack of awareness and knowledge among farmers, opposition to genetically modified planting materials, and insufficient application of technology, thus making the nation's output and consumption of rice fall between 67 to 70% when expressed as a straightforward measure of self-sufficiency [7].

Malaysia was only able to produce 1.52 million metric tonnes (MT) of rice, even though 2.90 million MT of rice were used in 2021 [6]. Malaysia's current rice output cannot keep up with the nation's expanding domestic demand given that the country only produces approximately 72% of the rice it needs to feed itself in 2019. Malaysia imports most of its rice from Pakistan, Vietnam, and Thailand. When compared to Thailand, Vietnam, Indonesia, and the Philippines, Malaysia's level of rice self-sufficiency remained the lowest [16]. 2020 will see a worsening of the problem due to the pandemic's impact on our rice crop [2]. The paddy and rice industry has attracted the attention of policymakers despite contributing very little to the country's GDP because of its extensive connections to socioeconomics, culture, and food security. As a result, more researchers created highly advanced technological techniques to increase domestic rice production. For this reason, budgeting and short-term planning for Malaysian rice production are

***For correspondence:**
sitirohani@utm.my
Received: 10 Nov. 2024
Accepted: 03 Feb. 2025

made possible for researchers, politicians, and rice farmers through the modelling and forecasting of rice production.

The process of developing a model must be considered to forecast paddy yield. In order to fully understand how these components affect the statistics, it is vital to evaluate the contributing factors that may influence paddy production, such as size of planted area, the production of rice, average temperature, and precipitation. To predict the dependent variable, which in this case, the paddy production, these elements will be regarded as independent variables. Following the determination of each variable and the collection of data, the data must be pre-processed to identify the nature of the data and select the optimal technique for the paddy production data and each variable that is being used.
Samsudin et. al. (2008) evaluates the modelling capabilities of the three different techniques, Multiple Linear Regression (MLR), Autoregressive Integrated Moving Average (ARIMA), and Artificial Neural Network (ANN) in relation to a set of data on rice yields [16]. Despite that, the approach only considers the internal aspects of paddy production and ignores any other external factors that can have an impact on it, making the model less accurate to represent the paddy production. In 2014, Alam et. al. (2014) uses Integrated Agriculture Development Area (IADA) microdata from Northwest Selangor, Malaysia from 1992 to 2007 to examine the effects of temperature and rainfall on the paddy sector using the time series linear and log linear OLS regression models [3]. Although it is simpler to interpret and consider temperature and rainfall as variables for the model, the OLS regression model is not appropriate for paddy production data with multicollinearity and nonlinearity problems. Meanwhile in 2016, Abiola ***et al***. (2016) looked at the MADA, Malaysia's paddy rice production's allocative efficiency and resource utilization [1]. In the study, descriptive statistics, gross margin analysis, independent samples F-tests, Ordinary Least Square analyses, and Cobb-Douglas production function analysis which combines the traditional neoclassical test of economic and technical efficiencies were used. However, like Samsudin *et al*., the model is less accurate in representing paddy production since the approach only considers the internal features of paddy production and excludes any other external factors that might have an impact on it. In addition, to improve the model's accuracy in representing the paddy production data, multiple regression models should be taken into consideration rather than only one regression model.
The previously mentioned literature analysis makes it abundantly evident that, due to the nature of paddy production data, a model for paddy production must account for both multicollinearity and nonlinearity and include external contributing elements rather than paddy production alone. In this study, Multiple Linear Regression (MLR), Multivariate Adaptive Regression Spline (MARS) and Support Vector Regression (SVR) were used to develop paddy production regression model in Malaysia from 1980 to 2022. These regression models were selected because of their capability to handle nonlinearity and multicollinearity, as well as their ability to fit many variables that contribute to paddy production. The contributing factors such as size of the planted area, rice production, average temperature and average precipitation were included into the forecasting model as independent variables. Their performances were then compared in modelling analysis to determine the best regression model.

Methodology

Data Scope

For this research, the paddy production data is obtained by making data requests through email from the Department of Agriculture and Food Industry and the climate data from Climate Change Knowledge Portal (CCKP). The paddy production data of Malaysia from 1980 until 2022 is used to develop the model. Paddy production, y will be the dependent variable, while size of planted area, x_1, rice production, x_2, average temperature, x_3, and precipitation, x_4 will be the independent variables

Correlation Test

Prior to developing the paddy production forecasting model, the nature of the data obtained must be identified. To determine a linear correlation, the data will first go through the Pearson Correlation Test. Pearson Correlation coefficient, r formula is given as follows [21].

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \quad (1)$$

By referring to equation (1), there is a correlation between the variables if the Pearson Correlation coefficient, r is not equal to zero. The correlation coefficient does not differ significantly from zero, indicating that there is no linear relationship, according to the null hypothesis, or H_0. Since the correlation coefficient differs considerably from zero and the relationship is linear, reject H_0 if the p-value is less than 0.05.

Multicollinearity Test

Multicollinearity increases variance and results in a consistent yet erratic coefficient of a variable. Regression analysis's Variance Inflation Factor (VIF) is a metric for multicollinearity. In a multivariate regression model, multicollinearity occurs when there is a correlation between several independent variables. As a result, the VIF test can be used to calculate the extent to which multicollinearity has inflated the variance of a regression coefficient. VIF formula is given by

$$VIF_i = \frac{1}{1-R_i^2} \tag{2}$$

where R_i^2 is the unadjusted coefficient of determination for regressing the ith independent variable on the remaining ones [20]. If the VIF value of the test obtained from equation (2) equals 1, it can be concluded that there is no correlation between the variables. The variables are moderately correlated if the VIF value falls between 1 and 5, and highly correlated if it exceeds 5.

Nonlinearity Test

The data will lastly undergo a nonlinearity test. To ensure that the proposed model can handle nonlinearity in the data, it is crucial to perform a nonlinearity test. The Ramsey Regression Specification Error Test (RESET) will be used to assess nonlinearity. Suppose there is a linear regression model

$$y_i = \lambda_1 + \lambda_2 x_i + u_i \tag{3}$$

where λ_i is the coefficient of the independent variables, x_i and y_i will be the dependent variables. This model will be the null hypothesis, H_0 which could be interpret as linearity [14]. Regression can be applied to equation (3) and augment of y_i and u_i, which are $\hat{y}_i$ and $\hat{u}_i$ can be produced. If $\hat{y}_i$ and $\hat{u}_i$ are plotted and produced a pattern, then it is suspected that the regression model is not suitable, indicating nonlinearity, thus rejecting the null hypothesis, H_0.

Multiple Linear Regression (MLR)

A statistical method called multiple linear regression (MLR), or just multiple regression, makes use of many explanatory variables to forecast the value of a response variable. Modelling the linear relationship between the response (dependent) variables and the explanatory (independent) variables is the aim of MLR. Below is the general formula and calculation for MLR

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon \tag{4}$$

where y_i is the dependent variable, x_i is the independent variables, β_0 is the y-intercept, β_p is the slope coefficients for each explanatory variable and ϵ is the model's error term. In this case, x_1 will be size of planted area, x_2 is rice production, x_3 is average temperature and x_4 is precipitation.

Multivariate Adaptive Regression Splines (MARS)

A non-parametric regression method known as multivariate adaptive regression splines (MARS) is like an extension of linear models in that it automatically accounts for nonlinearities and variable interactions. MARS builds models in the form of

$$\hat{f}(x) = c_0 + \sum_{i=1}^{k} c_i B_i(x) \tag{5}$$

A weighted sum of basis functions $B_i(x)$ makes up the model and a constant coefficient is denoted by each $c_i(x)$. MARS model is generated from basis function defined with set C, where $C = \{(x_j - t)_+, (t - x_j)_+ | t \in \{x_{1j}, x_{2j}, \ldots, x_{nj}\}, j \in \{1, \ldots, p\}\}$. Each element of C can be regarded as a basis function.

Support Vector Regression (SVR)

Given $\{(x_i, y_i), i = 1, \ldots, n\}$ where x_i is the vector of independent variables and y_i is the dependent variable. Support Vector Regression (SVR) is one of the support vector machine techniques that is used to obtain the optimal function $f(x)$ that reduces the empirical probability. The general formula for SVR is

$$R_e = \frac{1}{N}\sum_{i=1}^{N} L_\varepsilon(y - f(x)) \tag{6}$$

where L_ε penalizes model errors between training and estimated values.

Evaluation of the Model

The Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE), which are excellent tools for gauging the precision of the fitted model, can be used to evaluate the model.

$$MAPE = 100 * \left(\frac{\sum_{i=1}^{n} |F_i - O_i| / O_i}{n}\right) \quad (7)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (F_i - O_i)^2}{n}} \quad (8)$$

The number of variables is n, the predicted variable is called F_i, and the actual variable is called O_i. Moreover, the parameters' performance is determined by their lowest MAPE and lowest RMSE using the equations (7) and (8).

Results and Discussion

Descriptive Statistics

It is crucial to do descriptive statistics on the paddy production data prior to validating the data's nature. To help with the description and comprehension of the features of a specific data set, descriptive statistics offer succinct explanations of the sample and data measurements. To learn more about any given set of data, it is first necessary to plot the data against time. Table 1 shows the descriptive statistics table of each variable.

Table 1. Descriptive statistics table

Variable	Min	Max	Mean	Standard deviation	Skewness
Paddy production (kg/ha)	630,833	716,873	675,698.100	17,874.010	-0.243
Size of planted area (ha)	1,571,674	2,844,983	2,196,603	321,765.600	0.013
Rice production (mt)	1,010,279	1,834,831	1,424,557	216,182.900	0.018
Average temperature (°C)	25.600	26.930	26.279	0.279	0.007
Precipitation (Kmm)	2.410	3.570	2.969	0.316	0.032

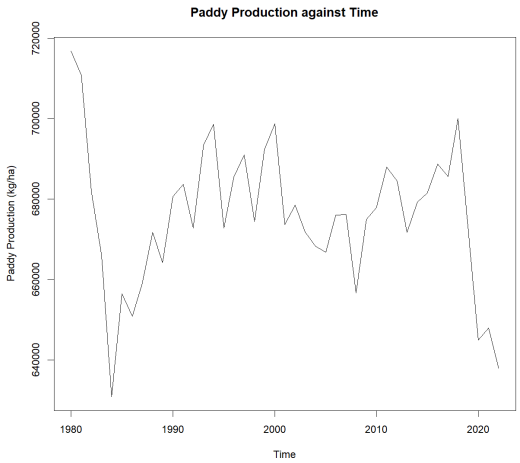

Figure 1. Paddy production (kg/ha) against time graph

As can be observed from Figure 1, when comparing the paddy production graph to the time graph, one can see that it drastically decreased from 1980 to 1995 due to the implementation of a policy that streamlined government involvement while encouraging aggressive private sector participation in agricultural development [5], when it was less than 640,000 kilogram per hectare (kg/ha), before rising again. It steadily rises and then alternates between increases and declines until the year 2000, at which point it starts to decline gradually in 2005. Up till 2019, the production of paddy increased somewhat. After 2020, there is a sharp decline until 2022. Paddy production had a mean of 675,698.100 and a standard deviation of 17,874.010, respectively. Paddy production has a minimum value of 630,833 kg/ha and a maximum value of 716,873 kg/ha. The skewness for the paddy production is -0.243.

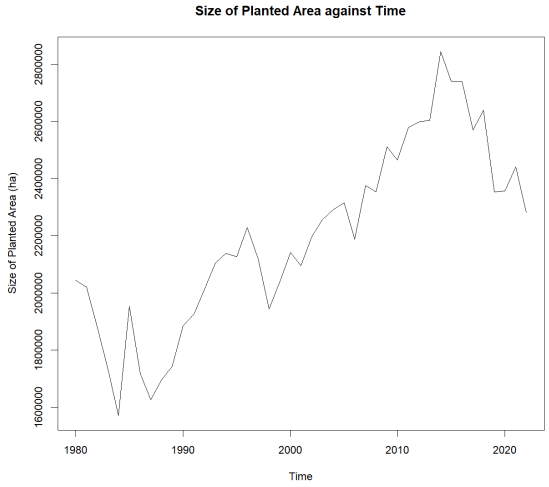

Figure 2. Size of planted area (ha) against time graph

Referring to Figure 2, from 1980 to 1987, the size of the planted area drops, increases, and then decreases again. The area's size gradually increases before slightly declining in 1998. Starting in 1999, the amount keeps growing until it reaches 2015, when the greatest area is over 2,800,000 hectares (ha). After that, the amount of planted land gradually reduces until 2022. The mean and standard deviation of planted area were 2,196,603 and 321,765.6, respectively. The minimum and maximum values of planted area are 1,571,674 ha and 2,844,983 ha, respectively. The skewness for the size of planted area is 0.013.

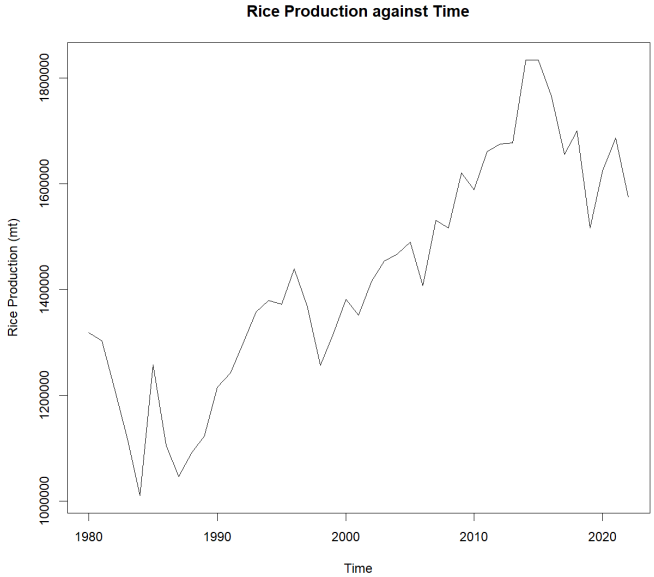

Figure 3. Rice production (mt) against time graph

When evaluating the rice production vs time graph from 1980 to 1987, it is evident from Figure 3 that the output of rice lowers, increases, and then decreases once more. The area grows throughout time, then starts to shrink a little in 1998. It started to accumulate in 1999 and reached over 1,800,000 metric tonnes (mt) of rice output in 2015. Following that, until 2022, the amount planted progressively decreased. The standard deviation of rice production was 216,182.900, and the mean was 1,424,557. The production of rice has minimum values of 1,010,279 mt and maximum values of 1,834,831 mt, respectively. The skewness for the rice production is 0.018.

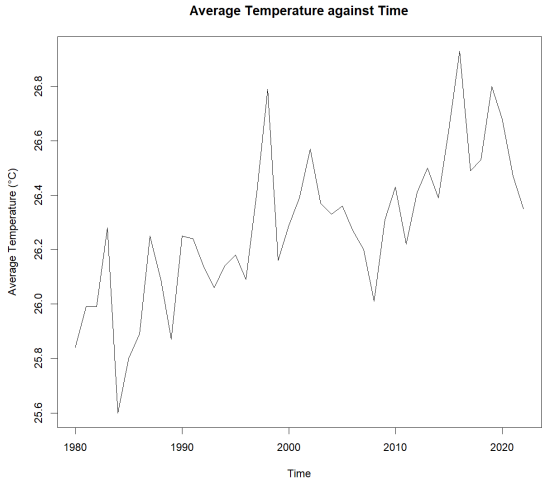

Figure 4. Average temperature (°C) against time graph

Figure 4 shows that the average annual temperature in Malaysia climbed in 1981, then remained stable for a year before rising again in 1983. The average temperature begins to fall in 1984, then steadily recovers before falling again in 1988. It rose substantially in 1990 to well over 27°C before falling again in 1999. It continues to rise until 2002, when it begins to decrease again in 2003. The average temperature began to rise around 2005 before declining again between 2017 and 2022. The mean temperature was 26.279, while the standard deviation was 0.279. The average temperature ranges from 25.6 °C to 26.93 °C. The skewness of the average temperature is 0.007.

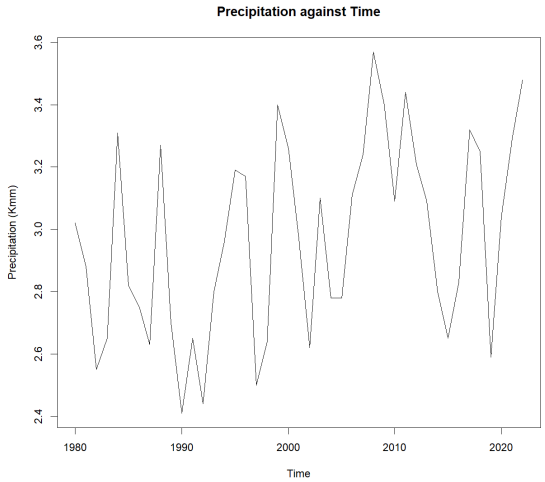

Figure 5. Precipitation (Kmm) against time graph

Finally, precipitation against time graph as can be seen in Figure 5. The graph shows that the annual precipitation data fluctuates between increases and declines on a yearly basis. The year 1990 had the least amount of precipitation (2.41 Kmm), while the year 2008 had the most (3.57 Kmm). Precipitation

had a mean of 2.969 and a standard deviation of 0.316. Minimum and maximum values of the precipitation are 2.41 Kmm and 3.57 Kmm, respectively. The skewness for the precipitation is 0.032.

Pearson Correlation Test

The Pearson correlation test is a statistical method for determining the strength and direction of a linear relationship between two continuous variables. It is based on the Pearson correlation coefficient, which measures how closely the variables are linearly related. The variables denote as x_1 for size of planted area, x_2 for rice production, x_3 for average temperature and x_4 for precipitation. The Pearson Correlation test results are displayed below.

Table 2. Pearson Correlation test results

Variables	p-value	Correlation coefficient
x_1 and x_2	0.0000	0.9912
x_1 and x_3	0.0000	0.6152
x_1 and x_4	0.0385	0.3167
x_2 and x_3	0.0000	0.6294
x_2 and x_4	0.0302	0.3308
x_3 and x_4	0.4139	-0.1279

As can be seen from Table 2, by observing the correlation coefficient, all variables have positive correlation with each other except x_3 and x_4 that has negative value of correlation coefficient. Among all the correlations, x_1 and x_2 has the strongest correlation with the value approaching 1 while x_3 and x_4 has the weakest correlation with the value approaching 0. From the p-value, Pearson Correlation test shows that all correlations are linear except x_3 and x_4 that has p-value greater than 0.05, thus rejecting H_0 that refers the correlation to be linear. Thus, x_3 and x_4 correlation is not linear.
All variables have correlation with one another, as the result shows, even though x_3 and x_4 have negative correlation. Apart from x_4, where the p-value is more than 0.05 and indicates that the relationship is not linear, the p-value for all tests indicates that the correlation is linear.

VIF Test

VIF test is a statistical approach used to identify multicollinearity in regression models. Multicollinearity occurs when two or more independent variables are closely correlated, causing distortions in regression coefficient estimation and reducing model interpretability. The VIF test results are displayed below.

Table 3. VIF test results

Variables	VIF Value
x_1	58.9932
x_2	63.1828
x_3	2.1393
x_4	1.4516

If the VIF value from the test is greater than 5, it indicates that the dataset contains strong multicollinearity. If the VIF value from the test is less than 5, it indicates that the dataset contains weak multicollinearity. If the VIF value from the test is equal to 1 the dataset contains no multicollinearity. Table 3 shows that x_1 and x_2 exhibit substantial multicollinearity, as indicated by a VIF value more than 5. x_3 and x_4 show weak multicollinearity because the VIF value obtained from the test is less than five. According to the results of the test, the proposed model must be capable of handling multicollinearity.

RESET Test

The Ramsey RESET is a statistical test that detects specification problems in regression models. It specifically evaluates whether the model is missing crucial variables, is poorly described, or has an erroneous functional form. The RESET test results are displayed below.

Table 4. RESET test results

Variables	p-value
x_1 and y	0.0022
x_2 and y	0.0006
x_3 and y	0.0189
x_4 and y	0.2565

From Table 4, x_1, x_2 and x_3 exhibit nonlinearity relationship from the RESET test. It is clear from the RESET Test results that the model for paddy production data should be able to handle nonlinear relationships. This makes sense given the nature of paddy production data, which may not always show a steady trend and produce a nonlinear model.

Fitting Regression Model

A linear model comprising all the variables, which in this case, the size of the planted area, rice production, average temperature, and precipitation, must be built to fit the paddy production data into the MLR model. Malaysia's paddy production data from 1980 to 2022 is fitted into the MLR model, and the model will be evaluated with the p-value for the Ljung-Box test and the corresponding RMSE and MAPE values.

In MARS model, the initial model fitted employed the backward pruning procedure, the default function with a degree of 1 and had no upper limit on the number of terms in the pruned model. The model is then tuned to find the optimal parameter to utilize for the MARS model, which in this case is degree of 3 and 5 as the maximum number of terms in the pruned model, by training the MARS parameter. To determine which model best represents Malaysia's paddy output from 1980 to 2022, a comparison was made using the same model with the same tuning parameters, with the exception that the pruning method was changed to the forward pruning approach. Both models for forward and backward approach were compared.

For SVR model, the default function, epsilon 0.1, was utilized in the first model fitted, which was an eps-regression type. Given that the paddy production data is nonlinear, the radial basis kernel was chosen to match the SVR. After that, the model is adjusted by training the SVR parameter to determine the ideal parameter to use for the SVR model, in this example, epsilon is equal to 0. The same model with the identical tuning settings was compared, with the difference that the SVR model employed nu-regression type, to ascertain which model better portrays Malaysia's paddy production from 1980 to 2022. The nu-regression type and the eps-regression type models were compared.

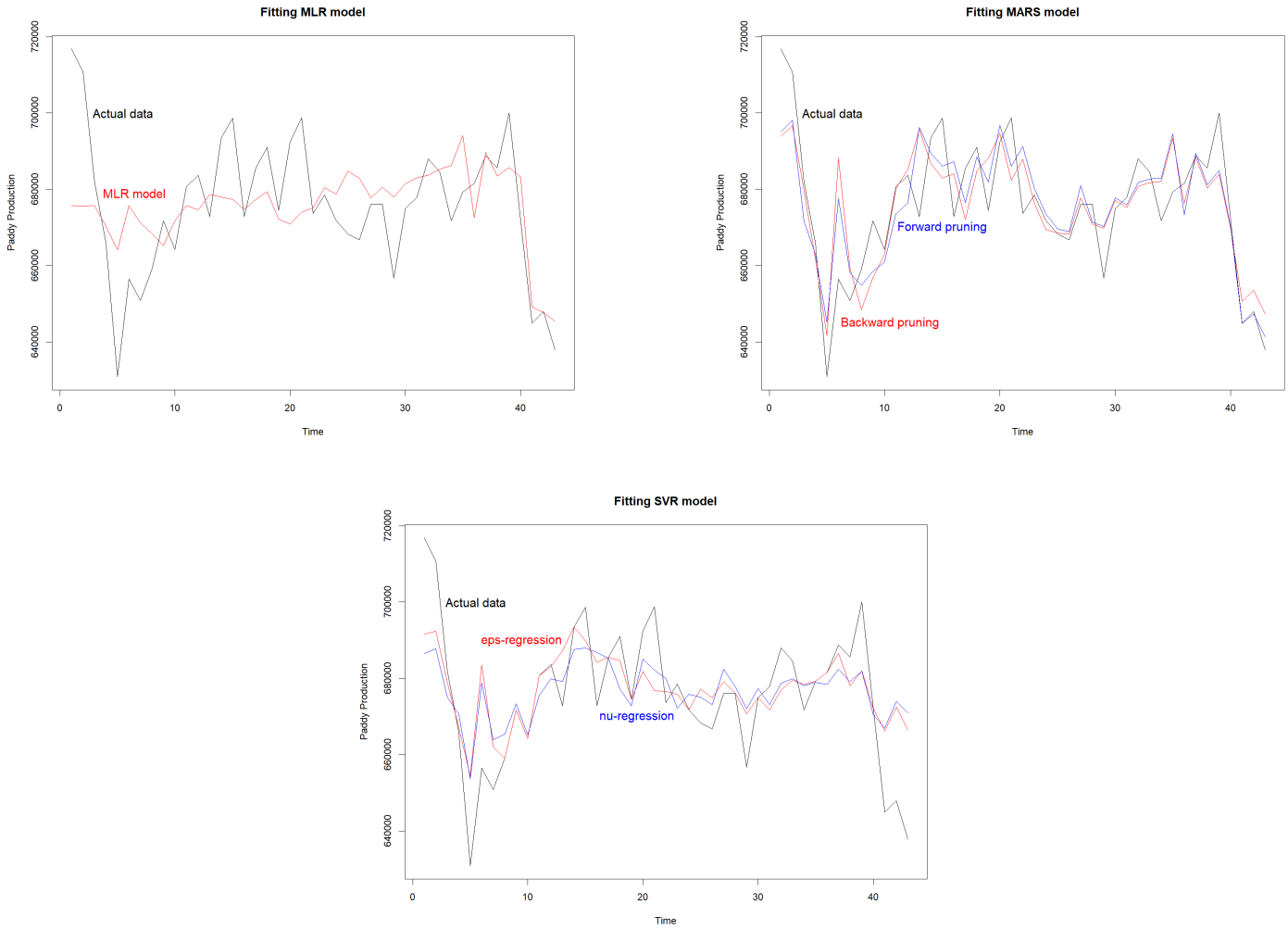

Figure 6. Fitting regression model

Table 5. Fitted MLR, SVR and MARS model evaluation

Model	Method	RMSE	MAPE	Ljung-Box test (p-value)
MLR model		14671.82	0.0164	0.0000
MARS model	Backward pruning	10975.69	0.0124	0.0006
	Forward pruning	10041.51	0.0116	0.0001
SVR model	eps-regression	12122.51	0.0123	0.0016
	nu-regression	12786.53	0.0144	0.0059

As can be referred to Table 5, according to the Ljung-Box test, the MLR model has a p-value of less than 0.05, indicating the presence of serial correlation in the model. MLR model also have significantly high RMSE and MAPE value, which is 14671.82 and 0.0163 respectively.

For MARS model, in comparison to backward pruning model, which has an RMSE value of 10975.69, forward pruning model has a smaller value of 10041.51. Additionally, forward pruning model MAPE value is lower than backward pruning model, at 0.0116 as opposed to 0.0124. As a result of this demonstration of forward pruning model superiority over backward pruning model, forward pruning model will be utilized to depict the fitted MARS model of paddy production. According to the Ljung-Box test, both models have a p-value of less than 0.05, indicating the presence of serial correlation in both.

The analysis of each SVR models reveal that the eps-regression model has a lower value of RMSE, which is 12122.51, than the nu-regression model, which has an RMSE value of 12786.53. Furthermore, the MAPE value of the eps-regression model is 0.0123, which is less than the nu-regression model's 0.0144. The fitted MARS model of paddy production will be shown using the eps-regression model as an outcome of this demonstration of its superiority over the nu-regression model. The Ljung-Box test shows that both models have a p-value of less than 0.05, which suggests that serial correlation exists in both.

Model Comparison

The optimal model among the three models that can model paddy production in Malaysia from 1980 to 2022 is found by combining and comparing the best fitted models for MLR, MARS, and SVR with their significant appropriate parameters. The fitting of Malaysia's paddy production data from 1980 to 2022 into the MLR, MARS and SVR model is displayed below, along with the p-value for the Ljung-Box test and the corresponding RMSE and MAPE values.

Table 6. Comparison of each model evaluation

Model	RMSE	MAPE	Ljung-Box test (p-value)
MLR	14831.35	0.0164	0.0000
MARS	10975.69	0.0124	0.0006
SVR	12122.51	0.0123	0.0016

Table 6 illustrates that MARS has the lowest RMSE value, which is 10975.69 when compared to MLR and SVR, which are 14831.35 and 12122.51 simultaneously. In addition, MARS's MAPE value is the lowest at 0.01243852 when compared to concurrently 0.01636378 and 0.01231306 for SVR and MLR. The Ljung-Box test indicates that all three models exhibit autocorrelation; nevertheless, the MARS model is the most suitable for use as a fitted model to accurately describe Malaysia's paddy production data from 1980 to 2022.

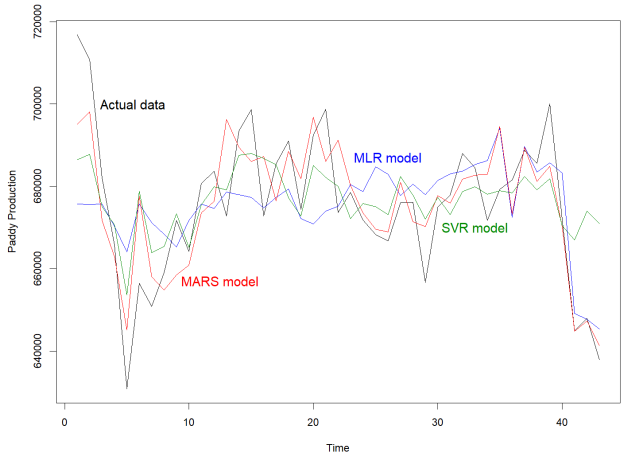

Figure 7. Comparison of each fitted model

Residuals and Outliers

To observe the residuals and outliers for each model, the residuals and the box plot for the fitted model needs to be plot so that the nature of the residual and outliers can be observed. A boxplot is a basic but effective technique for visualizing data distributions and finding outliers. Understanding and interpreting these outliers allows researchers to acquire insights into the data's unpredictability, uncover potential anomalies, and develop suitable handling techniques.

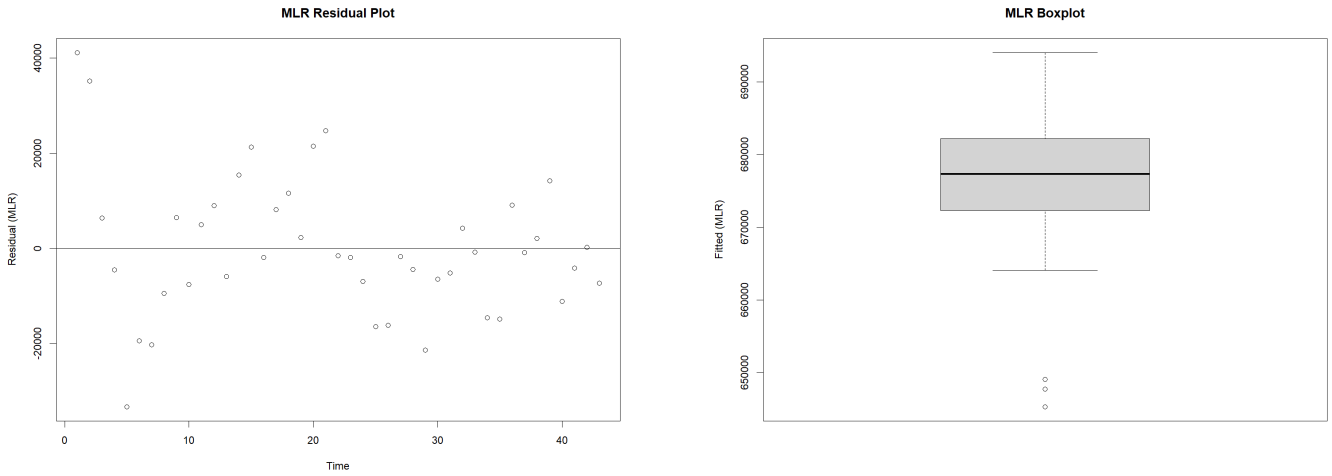

Figure 8. Residual plot and box plot for MLR model

The residual variance, also known as the variance of the error term, shows heteroscedasticity in the residual plot of the MLR model, meaning that it is not constant across observations. In addition, the residual lacks a noticeable trend and is not uniformly distributed, indicating that it is nonlinear. Additionally, symmetrically distributed, the residual tends to cluster in the middle of the plot. The minimum value for the MLR residual is -33248.17 while the maximum value is 41155.5. Observing the box plot for MLR model in Figure 8, there are 3 outliers exist in MLR model located below the first quartile of the box plot. The MLR model distribution is squeezed, which means the data has less variability. The data set for MLR model is also positively skewed. The minimum value for the outliers is 645304.4. The median for MLR model is 677307.5.

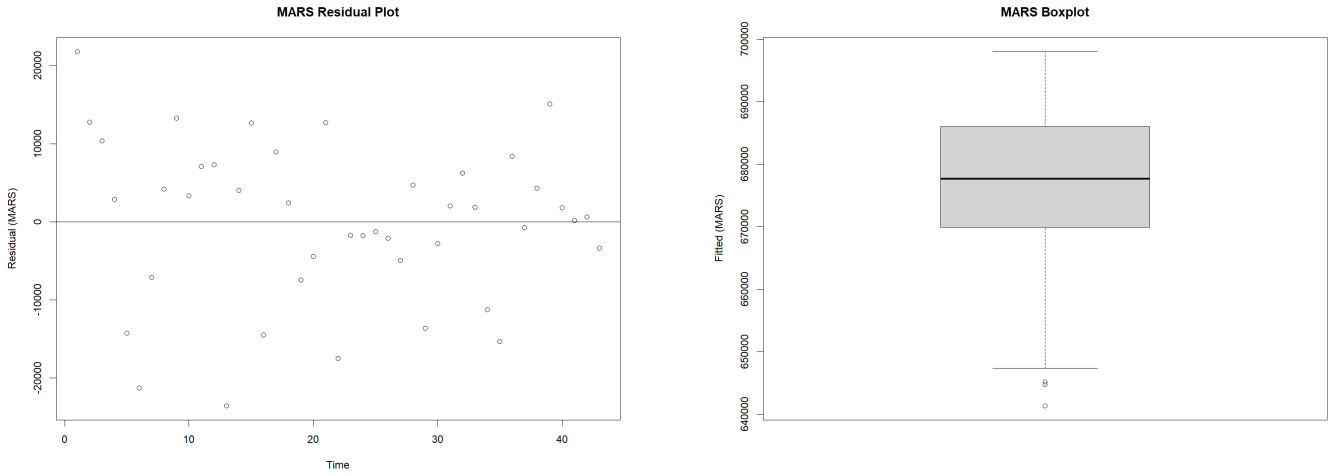

Figure 9. Residual plot and box plot for MARS model

The non-constant residual variance indicates the presence of heteroscedasticity in the MARS model's residual plot as can be seen in Figure 9. The residual is nonlinear, as it is not evenly distributed and does

not exhibit a discernible trend, much like the MLR model. The MARS model is the most accurate when compared to the SVR and MLR models since it shows the least amount of error. Furthermore, the residual tends to cluster in the middle of the plot when it is symmetrically distributed. The MARS residual has a minimum value of -23555.23 and a maximum value of 21814.03. Three outliers in the MARS model exist, as seen by the box plot, and are situated below the box plot's first quartile. MARS model is more stretched compared to MLR model, which translates to the variety of the MARS data set. Additionally, the MARS model's data set is negatively skewed. 641339.1 is the lowest value for the outliers. 677665 is the MLR model's median.

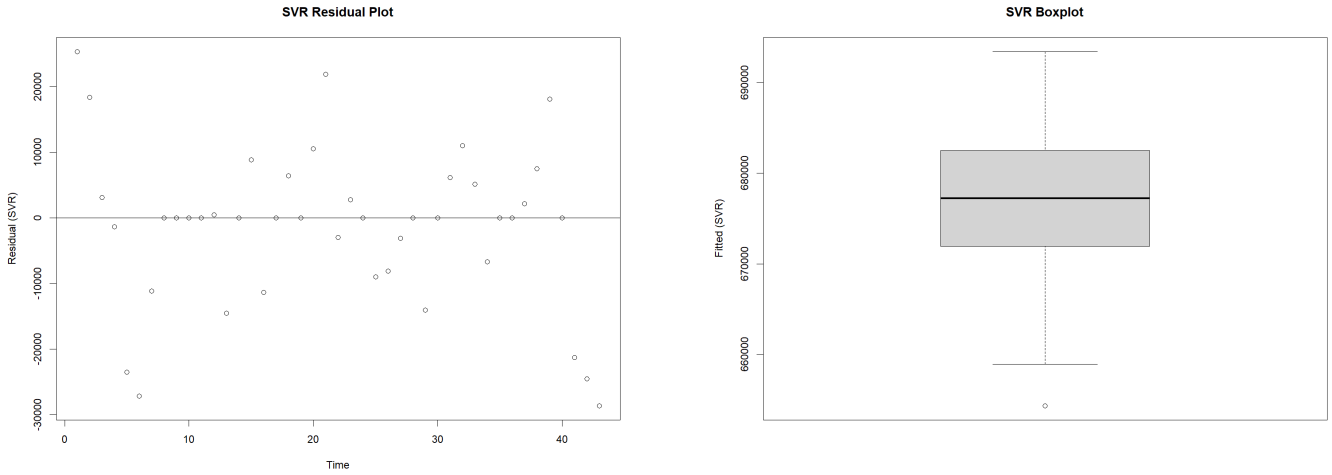

Figure 10. Residual plot and box plot for SVR model

In the residual plot of the SVR model in Figure 10, the residual variance exhibits heteroscedasticity, indicating that, like MLR and MARS, it is not constant across observations. Furthermore, the residual is not evenly distributed and does not exhibit a discernible trend, suggesting that it is nonlinear. Nonetheless, it is clustered around the 0 of the y-axis. The SVR residual ranges from a minimum of -28639.34 to a maximum of 25370.37. Unlike MLR and MARS model which have 3 outliers, only 1 outlier exist in the SVR model. Compared to other models, SVR has a more consistent error with a minimum outlier. Furthermore, the data set used in the SVR model also has a symmetrical distribution. The outlier is located below the first quartile of the box plot and has a value of 654350.2. The median for SVR model is 677239.6.

Since the MARS model is the most suitable regression model for modelling paddy production data in Malaysia, it can simulate the impact of natural disasters such as floods or droughts on paddy productivity, assisting with disaster preparation and recovery planning. If an unexpected tragedy occurs, the models can anticipate potential production shortfalls, allowing for timely measures to avoid food emergencies. Modelling paddy production provides insights that can significantly improve agricultural planning and resource management. It can help governments and stakeholders better allocate resources for agricultural growth. Modelling investigates the relationship between contributing factors that may have a significant impact on yields, supporting farmers in maximizing paddy production. Furthermore, anticipating paddy production using models can help to stabilize markets by minimizing price volatility caused by supply-demand imbalances.

This study has some potential limitations. One of them is a restricted resource of data. Data gathered for modelling purposes is insufficient because it only includes 43 annual data points. The modelling technique is based on a large dataset to ensure that the model accurately represents paddy production data. The proposed models also lack a mechanism for adequately handling outliers. Given that paddy production data is heavily influenced by climate, it tends to become volatile and diverge greatly from other observations in the dataset. MLR, SVR, and MARS models lack a way to manage outliers, hence other methods, such as parameter estimation, ought to be included to ensure that the models can handle outliers appropriately.

Conclusions and Recommendations

The study developed a regression model of Malaysian paddy production from 1980 to 2022 using Multiple Linear Regression (MLR), Multivariate Adaptive Regression Spline (MARS), and Support Vector Regression (SVR). Each model is evaluated thoroughly using RMSE, MAPE and Ljung-Box Test. Based on the evaluation, the best model to fit and accurately describe Malaysia's data on paddy output between 1980 and 2022 is the MARS model since both MLR and SVR has higher value of RMSE and MAPE and SVR model only capable of detecting 1 outlier, making it unsuitable to use as a model to predict paddy production data. MARS model also capable of handling the nonlinearity and multicollinearity in the data, making it a reliable regression model to be used in making prediction of paddy production data. Since there are outliers in the data, for future study, it is recommended that a parameter estimation method that can handle outliers properly could be implemented to the regression model to acquire a better method to model paddy production data.

Acknowledgement

The authors would like to express their appreciation to the Ministry of Higher Education Malaysia for providing funding through the Fundamental Research Grant Scheme (FRGS), Proposal No: FRGS/1/2023/STG06/UTM/02/12, with vote number R.J130000.7854.5F620 and Universiti Teknologi Malaysia for their financial support under the UTM Fundamental Research, Reference No: PY/2024/00951, with vote number Q.J130000.3854.23H62.

References

[1] Abiola, O. A., Mad, N. S., Alias, R., & Ismail, A. (2016). Resource-use and allocative efficiency of paddy rice production in Mada, Malaysia. *Journal of Economics and Sustainable Development, 7*(1).

[2] Adnan, N., & Nordin, S. M. (2020). How covid-19 affects the Malaysian paddy industry? Adoption of green fertilizer as a potential resolution. *Environment, Development and Sustainability, 23*(6), 8089–8129. https://doi.org/10.1007/s10668-020-00978-6

[3] Alam, M. M., Siwar, C., Talib, B., & Toriman, M. (2014). Impacts of climatic changes on paddy production in Malaysia: Micro study on IADA at North West Selangor. *Alam, MM, Siwar, C., Talib, B., and Mohd Ekhwan*, 251–258.

[4] Boehmke, B., & Greenwell, B. (2020a, February 1). Hands-on machine learning with R. Chapter 7 Multivariate Adaptive Regression Splines. https://bradleyboehmke.github.io/HOML/mars.html

[5] Daño, E. C., & Samonte, E. D. (2005). Public sector intervention in the rice industry in Malaysia. *Southeast Asia Regional Initiatives for Community Empowerment (SEARICE)*, Quezon City, 2548.

[6] Department of Agriculture. (2021, August 23). Principal Statistics of Paddy and Rice by All Seasons, Malaysia. Retrieved from *Department of Agriculture Official Portal*.

[7] Dorairaj, D., & Govender, N. T. (2023, April 19). Rice and paddy industry in Malaysia: Governance and policies, research trends, technology adoption and resilience. *Frontiers*. https://www.frontiersin.org/articles/10.3389/fsufs.2023.1093605/full

[8] Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics, 19*(1), 1–67.

[9] Hayes, A. (2023). Multiple linear regression (MLR) definition, formula, and example. *Investopedia*.

[10] Huang, C., Liu, D. D., & Wang, J. S. (2009). Forecast daily indices of solar activity, F10.7, using support vector regression method. *Research in Astronomy and Astrophysics, 9*(6), 694.

[11] Kabacoff, R. I. (2017). Multiple (linear) regression. *Quick-R: Multiple Regression*. https://www.statmethods.net/stats/regression.html

[12] Labjar, H., Cherif, W., Nadir, S., Digua, K., Sallek, B., & Chaair, H. (2016). Support vector machines for modelling phosphocalcic hydroxyapatite by precipitation from a calcium carbonate solution and phosphoric acid solution. *Journal of Taibah University for Science, 10*(5), 745–754.

[13] Mogaji, K. A. (2016). Geoelectrical parameter-based multivariate regression borehole yield model for predicting aquifer yield in managing groundwater resource sustainability. *Journal of Taibah University for Science, 10*(4), 584–600.

[14] Prabowo, H., Suhartono, S., & Prastyo, D. D. (2020). The performance of Ramsey test, White test, and Terasvirta test in detecting nonlinearity. *Inferensi, 3*(1), 1–12.

[15] Sagar, C. (2017, March 8). Building regression models in R using support vector regression. *KDnuggets*. https://www.kdnuggets.com/2017/03/building-regression-models-support-vector-regression.html

[16] Samsudin, R., Saad, P., & Shabri, A. (2008). A comparison of neural network, ARIMA model and multiple regression analysis in modeling rice yields. *Editorial Advisory Board*, 113.

[17] Sarena Che Omar, S. A. (2019). The status of the paddy and rice industry in Malaysia. 1–221.

[18] Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT Press.

[19] Sun, F. K., Lang, C., & Boning, D. (2021). Adjusting for autocorrelated errors in neural networks for time series. *Advances in Neural Information Processing Systems, 34*, 29806–29819.

[20] The Investopedia Team. (2024, June). Variance inflation factor (VIF). *Investopedia*.

https://www.investopedia.com/terms/v/variance-inflation-factor.asp
[21] Turney, S. (2024, February 10). Pearson correlation coefficient (R): Guide & examples. *Scribbr*. https://www.scribbr.com/statistics/pearson-correlation-coefficient/
[22] Vapnik, V. (2013). *The nature of statistical learning theory*. Springer Science & Business Media.