Time Series Data Analysis on Rice Production in Sri Lanka

Munasinghe, B.S.N.G.[*a] & Peiris, M.D.P.[b]

[a,b] Department of Human Resource Management, Faculty of Commerce and Management Studies, University of Kelaniya, Sri Lanka

[a] bsngm211@kln.ac.lk, [b]mdp@kln.ac.lk

Abstract

Rice is Sri Lanka's staple meal, eaten almost every day by a significant proportion of its people. Thus, rice production is generally centered on population, price, producers, related industries, and government authorities. There are two main rice production seasons: the Maha season (from September to March) and the Yala season (from May to August). In Sri Lanka, rice is said to have a regal past. More than just a staple dish for this island nation, its significance extends far beyond that. Rice symbolizes the nation and is vital to its history, customs, and even politics. Building a set of models to determine the long-term pattern and forecast future developments in paddy production for leading years is also a secondary objective. Further, rice is a major export crop of Sri Lanka. Therefore, by understanding the pattern and predicting the amount of rice production in the future, it is possible to be prepared in case the amount of rice production decreases. The study used secondary data from the Department of Census and Statistics, Sri Lanka, from 1951 to 2019. Two models were built mainly considering Yala and Maha seasonal rice production data and yearly rice production data, respectively. Autoregressive Integrated Moving Average (ARIMA) (5, 1, 0) was the most suitable model for seasonal data as it has the lowest Akaike information criterion (AIC) and Bayesian Information Criteria (BIC) values with 19.923 Mean Absolute Percentage Error (MAPE) value. Consequently, the 3MA (3- Moving Average) model was developed, which provided a Mean Absolute Error (MAE) of 200.556 and MAPE of 9.13. Considering both results, the most suitable model to forecast rice production in Sri Lanka was determined to be the 3MA model. In light of the findings, the researchers infer that 3MA is the best technique for predicting rice output in the future. Buyers and sellers will be able to use the results to plan for rice production in the coming years, as well as identify periods of low output and investigate their causes.

Keywords: AIC (Akaike Information Criterion), ARIMA (Autoregressive Integrated Moving Average), BIC (Bayesian Information Criterion) , Yala, Maha

[*] *bsngm211@kln.ac.lk*

01. Introduction

Half of the world's population relies on rice as their primary source of nutrition. The global output of paddy rice in 2017 was 769.7 million tons. Seventeen countries in Asia and the Pacific, as well as nine in North and South America and eight in Africa, rely on rice as their primary source of nutritional energy (Munasingha et al., 2021). Moreover, a fifth of the calories consumed by humans around the world come from rice, making it the most important grain utilized to satisfy human nutrition and caloric consumption. Every Sri Lankan relies heavily on rice as a staple diet, as it is the primary agricultural output of the country, which brings 8.7 percent of the country's GDP. All the districts cultivate paddy in their respective wetlands. In addition to being grown on 34% of the country's total land area (0.77 million hectares), it is also the most valuable crop grown. Approximately 560,000 hectares (ha) are cultivated during the Maha season while 310,000 hectares (ha) during the Yala season making the total average yearly cultivating area to be 870,000 ha. According to estimates, nearly 1.8 million farm households grow paddy across the island. Today, Sri Lanka produces 2.7 million tons of rice annually, which is sufficient to meet 95% of the country's needs (Senanayake et al., 2016).

02. Research Problem

Both demand and production of rice are increasing worldwide. Sri Lanka's rice demand is expected to be grown by 1.1% annually, so rice production needs to expand by 2.9% annually to keep a supply in line with the demand. Growing more rice is the only way to meet the demand. Countries and places with low-cost labor and abundant rainfall are ideal requirements for rice cultivation. Thus, Sri Lanka is the best place in the world to grow rice. However, in recent years, fewer farmers have shown an interest in cultivating paddy crops (Jayasooriya et al., 2022). Approximately 20 million people in Sri Lanka rely on rice. As a result, the rice industry is vital to Sri Lanka's economy. It is estimated that rice production accounts for about 30 percent of the agricultural GDP. The forecasting of rice production is crucial since rice is a staple grain in Sri Lanka and a key crop sold to foreign countries. Predictions made by previous studies are lacking in a high degree of accuracy. Therefore, the purpose of this work is to develop a more precise model to predict future rice production. In order to predict future supply and demand, this study will first examine historical data on rice production. The researchers anticipate that the farmers and government officials who are responsible for regulating rice output would benefit from this study.

This study aims to develop a suitable time series model to identify the variation in rice production in Sri Lanka. Finally, an appropriate model is selected to predict rice production in Sri Lanka. Several advantages can be obtained by modeling rice production and forecasting based on it. Buyers and sellers can get an idea about rice production in relevant years. Authorities can get an idea of the times when production decreases and find out the reasons for it. In this study, an accurate model is developed differently from the models that previous research has developed to forecast Sri Lanka's rice production.

03. Literature Review

In order to foretell the island nation's future rice output, researchers in Sri Lanka (Sivapathasundaram et al., 2015) analysed it in 2012 using time series techniques with secondary data from the Sri Lankan Department of Census and Statistics from 1952 to 2010. Distinguishing the order, data 1 helped to fix the average non-stationarity. The models that were most suited for the data were the Autoregressive Integrated Moving Average (ARIMA) (2, 1, 0), Autoregressive Integrated Moving Average (2, 1, 1), and Autoregressive Integrated Moving Average (2, 1, 2). The ARIMA (2, 1, 0) model had the lowest AIC and BIC values, making it the best option. As calculated by ARIMA (2,1,0), the MAPE was 10.5%. It was estimated that 4.07 million, 4.12 million, and 4.22 million tons of paddy would be harvested in 2011, 2012, and 2013, respectively.

Previous researchers worked on trend analysis and forecasting for paddy production in Sri Lanka. Here, they produced ARIMA (2,1,1) model for the Yala season and ARIMA (2,1,0) model for the Maha season (Munasingha & Napagoda, 2021). Paddy production forecasting in Sri Lanka was studied by who used the ARIMA model. Data on rice harvests from 1952 to 2010 were analyzed. A final ARIMA (2,1,0) model with 10.3 MAPE was proposed. (Sivapathasundaram, V. & Bogahawatte, C., 2015). In 2016, they examined whether the structure of the paddy / rice market in Sri Lanka is competitive and efficient, particularly by undertaking two tracer surveys. It was shown through these polls that the profit margins for most participants in the paddy/rice value chains of both the Nadu and Samba types are not excessive when compared to the average bank lending rate of 15%. In addition, it was revealed that there is no concrete proof that rice mills and wholesalers are exploiting either rice producers or customers through oligopoly methods such as "cornering of the market." Empirical findings indicate that typical Sri Lankan rice farmers have room to increase output efficiency by as much as 30%. Access to resources, age, migration, income sources, and agricultural education are revealed to be significant factors influencing production efficiency. In addition, we discover that the families who use only their own conserved seeds are less productive than those who also buy seeds from marketplaces. This research also shows that using a wide range of varieties decreases productivity (Senanayake, S.M.P. & Premaratne, S.P., 2016).

The goal of Jayasooriya's research in 2022 is to determine how paddy output will be affected by climate change and how farmers would choose to adapt. In rural Sri Lanka, 1410 farmers who grow paddy were surveyed to collect cross-sectional data. Using climate-smart methods, socioeconomic data, and individual farmer traits, this study examines farmers' decisions to adapt to climate change. It makes predictions about how much drought and flood affect farming communities that are able to adapt, and how much they are impacted by unable communities. In order to rule out selection and endogeneity biases, this research employs an endogenous switching regression model to compare adapters and non-adapters on metrics like output, volatility, downside risk exposure, and kurtosis. As seen in the findings, adaptors are affected by drought and flood, which has implications for all four production outputs. The non-adaptors were vulnerable to environmental stress experiences, but the adaptors considerably reduced their volatility and risk exposure. Empirical research shows that factors such as climate, economy, and credit have significant effects on adaptation. An individual's income, the amount of credit available to them, the overall amount of credit they have used, and the function of extension services are all crucial factors in the success of paddy production adaption.

Overseas, the paddy harvest in Telangana State, India, was predicted in 2016 using a time series analysis conducted by the Department of Statistics at Osmania University in Hyderabad. Data on annual paddy production was gathered from the Directorate of

Economics and Statistics in Hyderabad, Telangana State, for the research project. Data on annual paddy production was gathered from the Directorate of Economics and Statistics in Hyderabad, Telangana State, for the research project. From 1974-75 to 2013-14, more than 5 million tons of paddy has been produced annually. They found that the ARIMA (4,1,0) model worked best for estimating paddy harvests. Model adequacy was evaluated using the Ljung-Box test, and the model passed with flying colors. The chosen model has a low normalized BIC value of -0.185, a MAPE of 18.285, and an absolute error of 0.548. They constructed a model to predict future rice output, which indicates that production is expected to reach 6.81 million tons in 2015–16 and 7.71 million tons in 2020–21. The annual rate of paddy harvests is rising. (Sharm et al., 2016).

Research on using the ARIMA model to forecast paddy production in Andhra Pradesh was conducted in 2010 by the Department of Statistics, P. G. College of Science, and Osmania University in Hyderabad. Annual rice production data of Andhra Pradesh from 1955-56 to 2007-08 were used for this analysis. To analyze the data, we generated autocorrelation and partial autocorrelation functions and then used a Box-Jenkins autoregressive integrated moving average model. Conventional statistical methods were used to examine the model's accuracy. The autoregressive integrated moving average model was used to predict rice output three years in advance. They've uncovered all the ARIMA models that may work, and they've settled on the one with the lowest Akaike and Schwarz Bayesian information criteria (2, 2, 0) (Raghavender M., 2010).

A study has been conducted to obtain an estimate of black gram production in Bangladesh (Rahman et al., 2016). They obtained information from the Bangladesh Bureau of Statistics for 47 years from 1967-1968 to 2013-2014. The Phillips-Perron unit root test, the autocorrelation and partial autocorrelation functions are computed to determine whether the data is stationary or not and to generate preliminary estimates of the autoregressive and moving-average structures underlying the data. In this context, the researchers fit a Box-Jenkins autoregressive integrated moving average model. Akaike Information Criterion and Bayesian Information Criterion are used to find the optimal forecasting model, while the Ljung-Box Q test statistic is used to assess the model adequacy and the Jarque-Bera test is used to check the normality of the models. The ARIMA (0,1,0) model has been found to be a reliable one for making predictions. Comparisons with the percentage of error from the actual numbers and the mean absolute percent error verified the performance (MAPE).

In this section, we explored the significance of rice production as well as its influence on the gross domestic product of Sri Lanka. Additionally, certain models for estimating the amount of rice that will be produced in Sri Lanka were explored here. There was only one type of model that the researchers could find, and that was the ARIMA model. The percentage of accurate predictions made by the same model was not particularly high. As a result of this, a 3MA model in addition to an ARIMA model will be developed for this study, and the degree of accuracy shared by the two models will be compared.

04. Methodology

In this study, the researchers will examine the theoretical framework of time series prediction. First and foremost, the researchers will have a conversation about the behaviors of time series and then create the ARIMA model and examine the model diagnostics.

Subsequently, the 3MA model should be developed. Finally, the accuracy of the two models will be compared.

Dataset

The dataset contains yearly rice production (metric tons) for YALA and MAHA seasons between 1951-2019. However, some season data is not available for YALA or MAHA seasons. The dataset was downloaded from the Department of Census and Statistics.

Variable

The dataset contains yearly YALA and MAHA seasons rice production data from 1951-2019. So, variable of this study is yearly rice production (Metric ton) values.

Univariate Time series analysis

Measurements of a single variable collected at regular intervals constitute a univariate time series. Such models can be split into two categories: Ordinary regression models with time indices as x-variables, and time-series-based models (such as AR, MA, ARMA, and ARIMA). Several features must be taken into account in this study, including stationarity, trend, seasonality, constant variance, etc.

Stationary series

If there is no change in the sample means, sample variance, or sample autocorrelation during a given time period, then the time series is termed stationary. The four basic components that are expected to affect a time series are the trend, the cyclical component, the seasonal component, and the irregular component.

Differencing

The differenced series is the change between consecutive observations in the original series, and can be written as,

$$x_t = Y_t - Y_{t-1} = \nabla Y_t$$

Where ∇ is the (backward shift) difference operator? Another way to write the differencing operation is in terms of a backshift operator$B$, defined as $BY_t = Y_{t-1}$, SO

$$\mathrm{X}_t = Y_t - Y_{t-1} = \nabla Y_t = (1\text{-}B)\ Y_t$$

with $\nabla = (1-B)$, differencing can be performed successively, if necessary, until the trend is removed; for example, the difference is

$$\mathrm{X}_t = \nabla^2 Y_t = \nabla(\nabla Y_t) = (1-B)^2 Y_t$$

In general,

$$\mathrm{B}^d Y_t = Y_{t-d}$$
$$\nabla^d = (1-B)^d$$

Transformations

One of the primary steps in time series modeling is transforming data. Furthermore, the variance of most time series changes with time, making them non-stationary. When looking to normalize data and minimize heteroscedasticity, the log transformation is superior to alternative transformations. The corresponding logarithms and mathematical definitions of these factors are as follows.

$$Y_t = log10(Y_t)$$

For negative data, you can add a suitable constant to make all the data positive before applying the transformation. This constant can be then subtracted from the model to obtain predicted (i.e., the fitted) values and forecasts for future points.

Unit Root Test for Stationary

To identify where the time series is exactly stationary or non-stationary, unit root tests are used such as the Augmented Dickey-Fuller test (1979) and Kwiatkowski-Phillips-Schmidt-Shin test (1992) and Phillips–Perron tests. Statistical software provides these tests.

Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test

KPSS test is used to test the null hypothesis of the series is stationary around the mean or trend

H_0: series is stationary.
H_1: Series is not stationary.
The test derives based on the model.

$$y_t = \beta_0 D_t + \mu_t + \mu_t$$
$$\mu_t = \mu_t + 1 - \varepsilon_t$$

Where D_t is the deterministic component, μ_t is I (0) and $\varepsilon_t \sim$ W N (0, σ^2)
Null hypothesis is H_0: $\sigma^2 = 0$, which implies that μ_t is a constant

Augmented Dickey Fuller (ADF) test

The ADF test tests the null hypothesis that a time series $y_t = c + \delta t + \emptyset y_{t-1} + \beta_1 \Delta y_{t-1} + \cdots + \beta_p \Delta y_{t-p} + \varepsilon_t$

possess a unit root where Δ is the differencing operator such that $\Delta y_t = y_t - y_{t-1}$

H_0: $\emptyset=1$ (The series possess unit root / series is not stationary)
H_1: $\emptyset<1$ (The series does not possess a unit root / series is stationary)
Phillips- Perron (PP) Unit Root Tests
The test regression for the PP tests is

$$\Delta y_t = \beta_0 D_t + \pi y_t - 1 + \mu_t$$

Where μ_t is I (0) and may be heteroskedastic. The PP tests correct for any serial correlation and heteroscedasticity in the errors μ_t of the test regression.
The PP test tests the null hypothesis of $\pi=0$, in other words
H_0: The series possess a unit root (series is not stationary)

H_1: The series do not possess a unit root (series is stationary)

Autocorrelation Function (ACF)

Autocorrelation is the correlation between two values y_t and y_t+k in the same variable at times t and t+k assuming that the observations are equip-spaced.

$$\rho_k = \frac{E[(y_t - \mu)(y_{t+k} - \mu)]}{\sqrt{E[(y_t - \mu)^2]E[(y_{t+k} - \mu)^2]}} = \frac{Cov(y_t, y_{t+k})}{Var(y_t)}$$

This is known as the autocorrelation coefficient at lag autocorrelation function (ACF) is the collection of ρ values at k=0, 1, 2. The order of the MA(q) process is identified from the cutoff lag in the ACF plot.

Partial Autocorrelation Function (PACF)

Partial autocorrelation between y_t and $y_t + k$ is the autocorrelation between y_t and $y_t + k$ after removing the effect of other lags 1,2...., k-1. The partial autocorrelation of the kth order is defined as

$$\emptyset_{kk} = Corr(X_t - P(X_t|X_{t+1}, \dots, X_{t+k-1}), X_{t+k} - P(X_{t+k}|X_{t+1}, \dots, X_{t+k-1}))$$

Where P(W|Z) is the best linear projection of on. The order of the AR (p) process is identified from the cutoff lag in the PACF plot.

Autoregressive Integrated Moving Average (ARIMA) process

Let Z_t be a discrete purely random process with mean zero and variance

{Xt}is said to be an autoregressive process of order p, AR (p) if

$$X_t = \delta + \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \cdots + \alpha_p X_{t-p} + Z_t$$

{Xt} is said to be a moving average process of order q, MA (q) if

$$X_t = \mu + Z_t - \beta_1 Z_{t-1} + \beta_2 Z_{t-2} + \cdots + \beta_q Z_{t-q}$$

The process formed by combining AR (p) and MA (q) processes is called mixed autoregressive-moving average process ARMA (p, q)

$$X_t = \delta + \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \cdots + \alpha_p X_{t-p} + Z_t - \beta_1 Z_{t-1} + \beta_2 Z_{t-2} + \cdots + \beta_q Z_{t-q}$$

To fit the ARMA model, it is θ necessary to become stationary. For that different techniques are used. When the series is differenced by order of d and the ARMA process formed, then it is called Autoregressive Integrated Moving average process ARIMA (p, d, q)

$$\varphi(B)(1 - B)^d X_t = \delta + \theta(B) Z_t$$

Where φ(B) is the AR polynomial and θ (B) is the MA polynomial is the backshift operator.

Akaike– information criterion for model selection

The difficulty of model selection arises because there are typically multiple models for characterizing a population from a given set of observations in practice. In this case, the time series modeling approach is crucial in choosing the right model from the highlighted data. Both the AIC and BIC are often used as criteria for selecting the best models. It has been found through reviews that AIC is the preferred criterion for the selection in the majority of studies.

The Akaike– information criterion is computed as

$$\mathrm{AIC} = -\frac{2\mathrm{l}}{\mathrm{n}} + \frac{2\mathrm{k}}{\mathrm{n}}$$

Where l is the log likelihood.

Diagnostic checking

Heteroscedasticity

One key assumption of regression is that the variance of the errors is constant across observations. When the variances of errors are constant, then homoscedasticity is present. Then, the violation of constant error variance heteroscedasticity is present (Dudewicz, E.J, 1980). It is more helpful to determine that residuals have constant variance. Generally, the hypothesis of these tests is as follows,

H_0: There is no heteroscedasticity between residuals

H_1: There is heteroscedasticity between residuals.

Jarque and Bera test

To check whether the residuals are normally distributed or not, the Jarque-Bera test can be used. The test statistic of the Jarque-Bera test is

Test statistics $=\frac{\mathrm{n}}{6}\{(\mathrm{Skewness})^2 + \frac{(\mathrm{kurtosis}-3)^2}{4}\}$

Corresponding hypotheses are

H_0: Residuals are normally distributed.

H_1: Residuals are not normally distributed.

Furthermore, the researchers can use the Quantile-Quantile plot to check whether the residuals follow a normal distribution.

ISBN: 978-624-5553-43-3

Ljung-Box Chi-Square test

Another measurement to check the randomness of residuals is the Ljung-Box Chi-Square test and this must show a p-value greater than 0.05.

Generally, the hypothesis is as follows.

H_0: There is no autocorrelation between residuals
H_1: There is autocorrelation between residuals.

The test statistic of the Ljung-Box test is

$$Q = n(n+2)\sum_{i=1}^{m}\frac{\gamma_j{}^2}{n-j}$$

Where n is the sample size, γ_i^2 is the autocorrelation at lag j, and m is the number of lags being tested at the corresponding significance level.

The hypothesis can be rejected if

$$Q > x^2{}_{1-\alpha,m-p}$$

Here x^2 is the chi-squared distribution and p is the number of estimated parameters (No, T. and Lee, T. (2020)

Moving Average Models

According to the linear dependence of the current value on the present and previous error terms, a moving average process (or the moving average model) can be defined. Similar to white noise, it is assumed that the error terms are uncorrelated and regularly distributed.

For example, if the order of your moving average model is q, the researchers would write it as MA (q). The present value is expressed in this model as a linear combination of the series mean mu, the current error term epsilon, and the historical error terms epsilon and mu (epsilon). An efficiency represented by theta can be used to measure how many prior mistakes have affected the current value. In mathematical terms, we can describe a typical moving-average process as follows:

$$y_t = \mu + \ \varepsilon_t + \vartheta_1\varepsilon_{t-1} + \vartheta_2\varepsilon_{t-2} + \cdots + \vartheta_q\varepsilon_{t-q}$$

$$\varepsilon_t - white\ noice$$

The amount of error terms from the past that affect the current value is controlled by the order q of the moving average model. If it's the first order, when denoting an MA (1) process, the model would look something like this:

$$y_t = \mu + \ \varepsilon_t + \vartheta_1\varepsilon_{t-1}$$

Model Forecast Accuracy Criteria

There are several ways to determine the performances of forecasting models. The researchers denoted the actual value and forecasted value in period t as $Y\ t\ and\ \hat{Y}_t$ respectively.

Mean Absolute Percentage Error (MAPE)

For each interval, the researchers divide the absolute error by the observable values to get the Mean Absolute Percentage Error (MAPE). Then the researchers take an average of those constant percentages. This method comes in handy whenever the size or size of a prediction variable matters greatly when assessing the precision of a forecast. The Mean Absolute Prediction Error (MAPE) measures how far off the estimates are from reality **(**Chai, T. and Draxler, R.R. (2014**).**

$$\mathrm{MAPE} = \frac{\sum_{t=n+1}^{n+h} \left| \frac{\hat{Y}_t - Y_t}{Y_t} \right|}{h} \times 100$$

Root Mean Square Error (RMSE)

The RMSE is a quadratic scoring rule which measures the average magnitude squared error. The calculating technique of the RMSE was the difference between forecast and actual values which were squared and then get the average over the sample. RMSE is frequently used to compare the forecasts. So, the errors are squared before they are averaged.

The root mean squared error calculated by

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{t=n+1}^{n+h} (\hat{Y}_t - Y_t)^2}{h}}$$

One problem associated with the use of the RMSE or similar measures is the fact that the forecast error variance varies across time. It can vary because of the nonlinearities in the model and because of the variations in exogenous variables (if included in the model). Fair (1986) states that rigorous statistical interpretation cannot be put on the RMSE because they do not estimate any parameter in the mode (Shekhar. 2008).

Mean Absolute Error (MAE)

The Mean absolute error MAE is also dependent on the scale of the dependent variable but it

is less sensitive to large deviations than the usual squared loss.

$$\mathrm{MAE} = \frac{\sum_{t=n+1}^{n+h} \left| \hat{Y}_t - Y_t \right|}{h}$$

The main objective of this study is to build an appropriate time series model to identify the variation in rice production in Sri Lanka and forecast the production of rice in Sri Lanka using the fitted time series model. This study is conducted by using the secondary data of the Department of Census and Statistics of Sri Lanka from 1951 to 2020. This fits a suitable Time series model by using the statistical software EViews. A suitable time series model will be used to fit the data set which is a complement to the trend regression approach and forecasting of the concerned variable to the near future. The annual value of the concerned variable is used in this study. So, the study will be done in four stages such as Identification process, Estimation, Diagnostic testing, and Forecasting.

05. Results

Figure 1 shows the time series plot from 1952 to 2006 period (model fitting dataset). This data is used as the training set (model development dataset). The rest of the data (2007-2019) was used as test data (model evaluation dataset) which seems to be non-stationary with an upward nonlinear trend component and a non-constant variance. There seem to be some points in the data set which can be identified as outliers.

Time series analysis for rice production in Yala and Maha season (1951-2006)

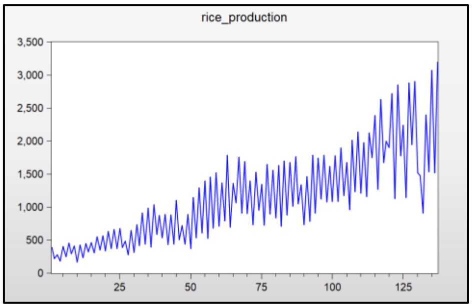


Figure 1:Time series plot of seasonal rice production from 1952-2006

Checking the stationarity of the series of yearly rice production

A time series is considered stationary if there is no change in the sample means, sample variances, or sample autocorrelations over a certain time interval.

Table 2:Stationarity Result before differencing

Test Name	Value	Stationarity
ADF	p-value = 0.8952	p-value>0.05 hence series not-stationary
PP	p-value = 0	p-value<0.05 hence series stationary
KPSS	LM-Stats = 1.1750	LM-Stats >0.46 hence series not-stationary

The results of the above three tests show that according to the ADF and KPSS tests series is not stationary, while the PP test shows that the series is stationary at the 5% level of significance. Therefore, the difference in the price series was considered, and checked the stationarity.

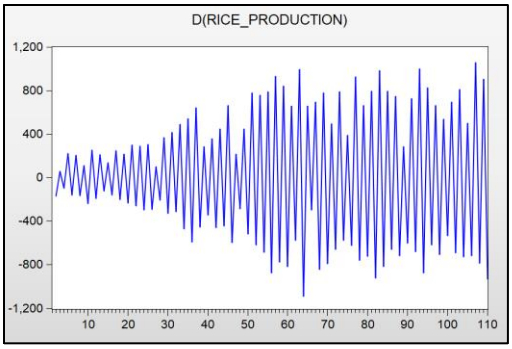


Figure 2 Time series plot after getting the first difference

Figure 2 shows the first difference series of rice production. This series seems to be stationary and has no upward trend component and constant variation.

Checking the stationarity of the d (rice production) (first difference) series

Table 2: Stationarity Result after differencing

Test Name	Value	Stationarity
ADF	p-value = 0	p-value<0.05 hence series stationary
PP	p-value = 0	p-value<0.05 hence series stationary
KPSS	LM-Stats = 0.2432	LM-Stats <0.46 hence series not-stationary

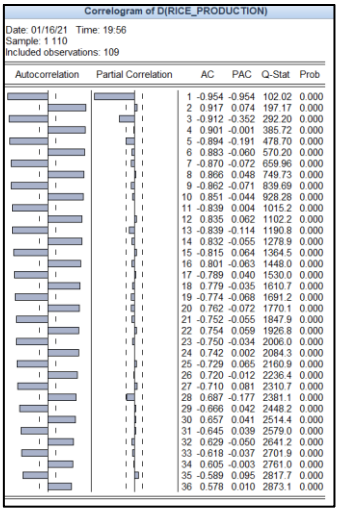
The results of the above three tests show that the d (rice production) series is stationary at 5% level of significance.

Checking ACF and PACF to estimate the model.

A correlogram is a graph which shows the degree of correlation between two sets of data.

Figure 3 shows the ACF and PACF plots of the first difference series. In the ACF plot, all the cutoff lags are significant which means that it becomes zero abruptly, in the PACF plot lag 1, lag 3, and lag 5 are significant. ("Cuts off" means that it becomes zero abruptly, and "tails off" means that it decays to zero asymptotically (usually exponentially).

ISBN: 978-624-5553-43-3



Correlogram of D(RICE_PRODUCTION)

Date: 01/16/21 Time: 19:56
Sample: 1 110
Included observations: 109

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.954	-0.954	102.02	0.000
		2	0.917	0.074	197.17	0.000
		3	-0.912	-0.352	292.20	0.000
		4	0.901	-0.001	385.72	0.000
		5	-0.894	-0.191	478.70	0.000
		6	0.883	-0.060	570.20	0.000
		7	-0.870	-0.072	659.96	0.000
		8	0.866	0.048	749.73	0.000
		9	-0.862	-0.071	839.69	0.000
		10	0.851	-0.044	928.28	0.000
		11	-0.839	0.004	1015.2	0.000
		12	0.835	0.062	1102.2	0.000
		13	-0.839	-0.114	1190.8	0.000
		14	0.832	-0.055	1278.9	0.000
		15	-0.815	0.064	1364.5	0.000
		16	0.801	-0.063	1448.0	0.000
		17	-0.789	0.040	1530.0	0.000
		18	0.779	-0.035	1610.7	0.000
		19	-0.774	-0.068	1691.2	0.000
		20	0.762	-0.072	1770.1	0.000
		21	-0.752	-0.055	1847.9	0.000
		22	0.754	0.059	1926.8	0.000
		23	-0.750	-0.034	2006.0	0.000
		24	0.742	0.002	2084.3	0.000
		25	-0.729	0.065	2160.9	0.000
		26	0.720	-0.012	2236.4	0.000
		27	-0.710	0.081	2310.7	0.000
		28	0.687	-0.177	2381.1	0.000
		29	-0.666	0.042	2448.2	0.000
		30	0.657	0.041	2514.4	0.000
		31	-0.645	0.039	2579.0	0.000
		32	0.629	-0.050	2641.2	0.000
		33	-0.618	-0.037	2701.9	0.000
		34	0.605	-0.003	2761.0	0.000
		35	-0.589	0.095	2817.7	0.000
		36	0.578	0.010	2873.1	0.000

Figure 3: Correlogram of the First differenced series

As the series is stationery and univariate, an ARIMA model was fitted initially. Several possible model combinations were tested and the model with minimum AIC value was selected as the best model among them and their results are shown in Table 3.

Table 3: Candidate ARIMA models for the seasonal rice production series

Model	AIC	Significance
ARIMA (1,1,0)	13.02950	Significant
ARIMA (3,1,0)	12.83558	Significant
ARIMA (5,1,0)	12.80118	Significant

ARIMA (5,1,0) was the minimum AIC and BIC value among the candidate models identified and the model was fitted to the time series plot of rice production.

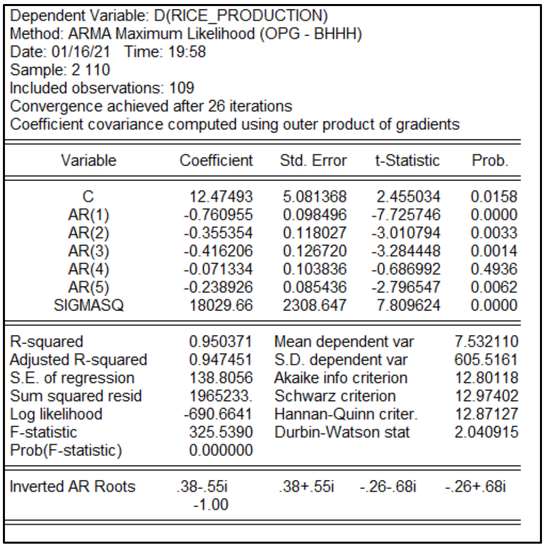
ARIMA (5,1.0)

Figure 4 shows the developed ARIMA (5,1,0) model details and accuracy.

Fitted Model

According to Figure, ARIMA (5,1,0) model can be written as in Equation 1.

$$X_t = X_{t-1} - 12.47493 - 0.76095(X_{t-1} - X_{t-2}) - 0.35535(X_{t-2} - X_{t-3}) - 0.41621 (X_{t-3} - X_{t-4}) - 0.07133(X_{t-4} - X_{t-5}) - 0.23893(X_{t-5} - X_{t-6})$$



```
Dependent Variable: D(RICE_PRODUCTION)
Method: ARMA Maximum Likelihood (OPG - BHHH)
Date: 01/16/21   Time: 19:58
Sample: 2 110
Included observations: 109
Convergence achieved after 26 iterations
Coefficient covariance computed using outer product of gradients

        Variable      Coefficient    Std. Error    t-Statistic    Prob.

           C           12.47493      5.081368      2.455034     0.0158
         AR(1)        -0.760955      0.098496     -7.725746     0.0000
         AR(2)        -0.355354      0.118027     -3.010794     0.0033
         AR(3)        -0.416206      0.126720     -3.284448     0.0014
         AR(4)        -0.071334      0.103836     -0.686992     0.4936
         AR(5)        -0.238926      0.085436     -2.796547     0.0062
        SIGMASQ        18029.66      2308.647      7.809624     0.0000

R-squared              0.950371    Mean dependent var       7.532110
Adjusted R-squared     0.947451    S.D. dependent var       605.5161
S.E. of regression     138.8056    Akaike info criterion    12.80118
Sum squared resid      1965233.    Schwarz criterion        12.97402
Log likelihood        -690.6641    Hannan-Quinn criter.     12.87127
F-statistic            325.5390    Durbin-Watson stat       2.040915
Prob(F-statistic)      0.000000

Inverted AR Roots      .38-.55i      .38+.55i     -.26-.68i     -.26+.68i
                        -1.00
```

Figure 4: ARIMA (5,1,0) model details

Model Diagnostics.

Normality Checking.

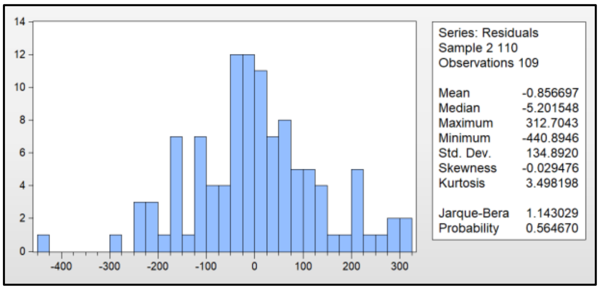
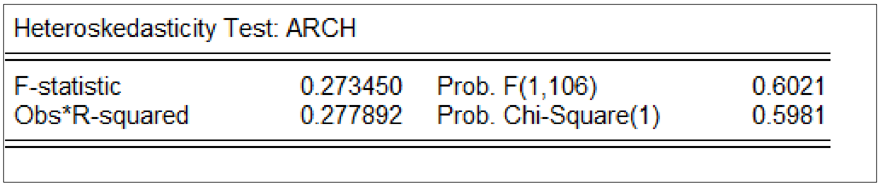


Figure 5: Jarque-Bera test results of residuals

The Jarque-Bera test is a goodness-of-fit test that checks to see if the skewness and kurtosis of a set of data are consistent with that of a normal distribution. Any non-zero value for the Jarque-Bera test's statistic indicates that the sample data does not follow a normal distribution.

According to the Figure, the p-value 0.564670 is greater than 0.05. Therefore, the assumption of the normality of the residuals is not violated at the 5% level of significance.

Heteroscedasticity



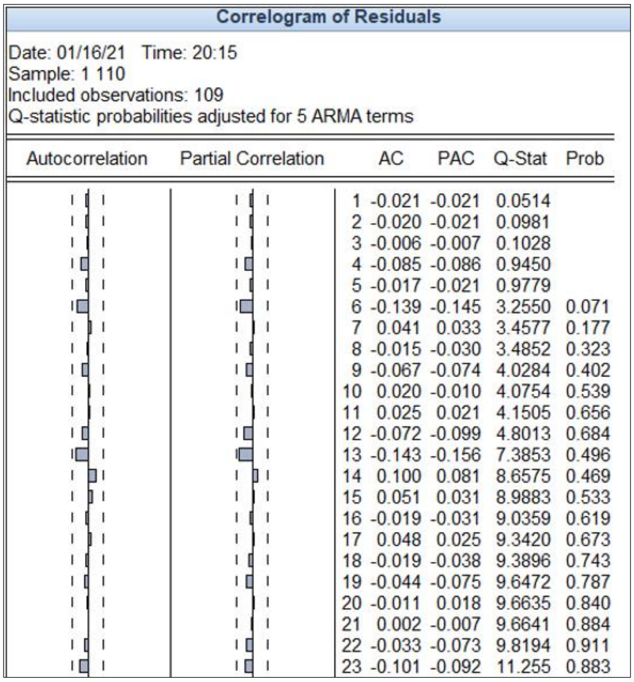
Heteroskedasticity Test: ARCH

F-statistic	0.273450	Prob. F(1,106)	0.6021
Obs*R-squared	0.277892	Prob. Chi-Square(1)	0.5981

Figure 6: Heteroscedasticity Test Results

The above shows the result for heteroscedasticity. P-value is 0.6021 which is greater than 0.05. It concludes that the residuals do not have an ARCH effect at a 5% level of significance.

Independence of the residuals



Correlogram of Residuals

Date: 01/16/21 Time: 20:15
Sample: 1 110
Included observations: 109
Q-statistic probabilities adjusted for 5 ARMA terms

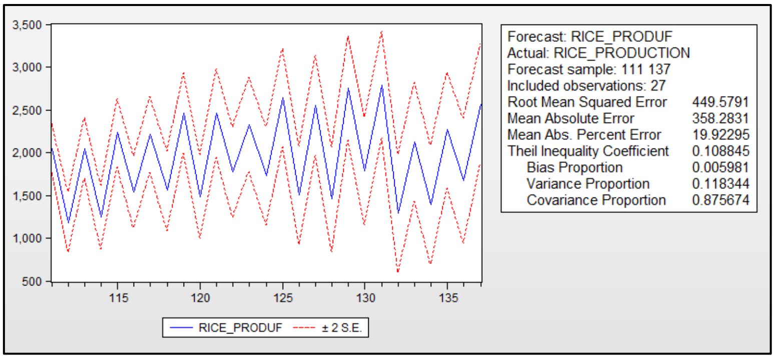
	AC	PAC	Q-Stat	Prob
1	-0.021	-0.021	0.0514	
2	-0.020	-0.021	0.0981	
3	-0.006	-0.007	0.1028	
4	-0.085	-0.086	0.9450	
5	-0.017	-0.021	0.9779	
6	-0.139	-0.145	3.2550	0.071
7	0.041	0.033	3.4577	0.177
8	-0.015	-0.030	3.4852	0.323
9	-0.067	-0.074	4.0284	0.402
10	0.020	-0.010	4.0754	0.539
11	0.025	0.021	4.1505	0.656
12	-0.072	-0.099	4.8013	0.684
13	-0.143	-0.156	7.3853	0.496
14	0.100	0.081	8.6575	0.469
15	0.051	0.031	8.9883	0.533
16	-0.019	-0.031	9.0359	0.619
17	0.048	0.025	9.3420	0.673
18	-0.019	-0.038	9.3896	0.743
19	-0.044	-0.075	9.6472	0.787
20	-0.011	0.018	9.6635	0.840
21	0.002	-0.007	9.6641	0.884
22	-0.033	-0.073	9.8194	0.911
23	-0.101	-0.092	11.255	0.883

Figure 7: Correlogram of Squared Residuals

The correlogram of squared residuals in Figure 7 indicates that autocorrelations and partial autocorrelations p-values are greater than 0.05, Therefore; it can be concluded that residuals are independent at a 5% level of significance.

Forecasting the Rice Production by using ARIMA (5,1,0) model.

Forecasted Series



```
Forecast: RICE_PRODUF
Actual: RICE_PRODUCTION
Forecast sample: 111 137
Included observations: 27
Root Mean Squared Error      449.5791
Mean Absolute Error          358.2831
Mean Abs. Percent Error      19.92295
Theil Inequality Coefficient 0.108845
     Bias Proportion         0.005981
     Variance Proportion     0.118344
     Covariance Proportion   0.875674
```

Figure 8: Forecasted series Results.

Figure 8, gives the root mean squared error as 449.579 and the mean absolute error as 358.283. It indicates that the fitted model is good enough because of the small error.

Actual vs Fitted residuals Results.

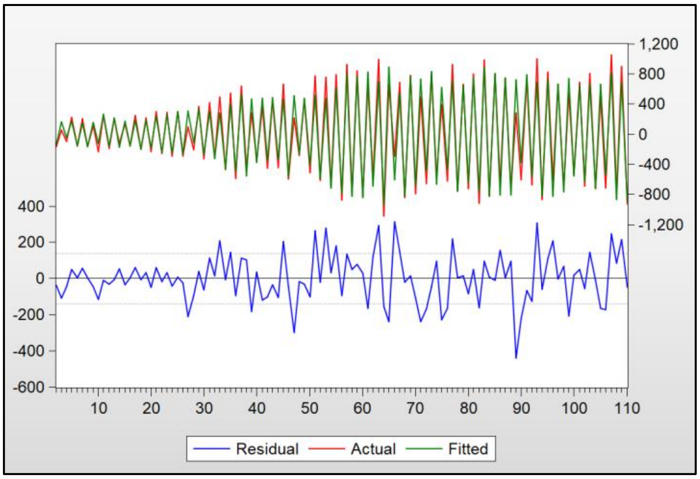


Figure 9: Actual vs Fitted residuals Results.

Figure 9 shows the actual value and forecast values for the fitted model. Here, the green color line shows the forecast values and the red color line shows the actual values. The blue color line shows the residual values between the forecast value and the actual value.

Actual Vs Fitted value graph (Forecasted)

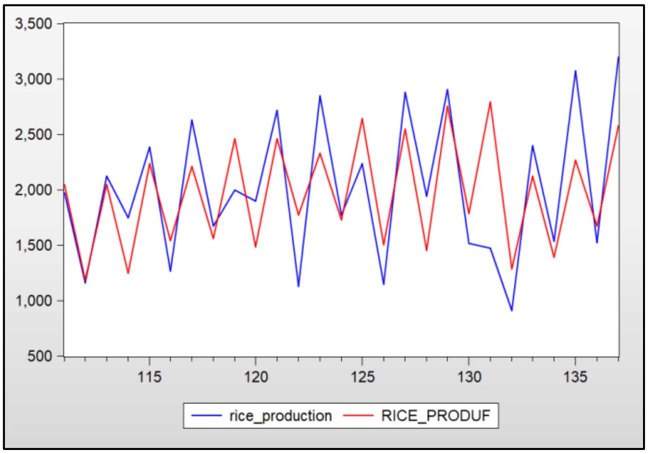


Figure 10: Actual vs Fitted value graph.

Moving Average Model

Long-term trends can be predicted with the help of the moving average, a statistical tool. This method entails repositioning a specific range of numbers while averaging a collection of numbers within that range.

Here 3-Moving-Average model was developed. The following table shows the performance measurement of the developed model.

Table 4:3MA model results

Measurement	Values
Root mean squared error	303.136
Mean absolute error	200.556
Mean absolute percentage error	9.13

According to the result, it is clear that the 3MA model root mean squared error is 303.136, mean absolute error is 200.556 and mean absolute percentage error is 9.13. Considering those results, it seems that 3MA model is more accurate than the ARIMA model

05. Discussion

In this study, the researchers build a suitable time series model for rice production in Sri Lanka and the secondary data was taken from the Department of Census and Statistics of Sri Lanka from 1951 to 2019. Fit two suitable time series models were developed in this data set. The time series analysis for yearly rice production in Yala and Maha seasons fitted model is ARIMA (5,1,0). Then the model was developed with 3MA. The 3MA model is the most accurate model than the ARIMA model in this research and previous research.

The 3MA model is more accurate than the two ARIMA models that Munasingha and Napagoda produced, the ARIMA (2,1,1) model for the Yala season and the ARIMA (2,1,0) model for the Maha season. Further,, in 2015, Sivapathasundaram, V. & Bogahawatta C. proposed the ARIMA (2,1,0) model with a MAPE of 10.3. It was also not as accurate as the 3MA model because the current model MAPE is 9.13.

The researchers could achieve their main objectives through this study. It is to build an appropriate time series model to identify the variation of the rice production in Sri Lanka with high accuracy and forecast production of rice in Sri Lanka using the fitted time series model. These two objectives are satisfied in the 3MA model. Some points in the actual and forecasted rice production series were violated because of climatic factors such as rainfall temperature, flood destruction, insect pest, inadequate infrastructure, and mechanization. In addition to climate-related problems, there are systemic dysfunctions in the industry, such as a lack of efficient storage to avoid water damage, logistical challenges, and Socioeconomic factors.

This research does not consider another climatic factor or variable to forecast rice production. Because of that, the researchers hope to conduct a future study to consider one or more other variables that influence rice production.

06. Conclusion

The primary objective of this research is to develop a statistical model for predicting Sri Lanka's annual rice production by analyzing historical data and observing seasonal patterns. Two main categories of models were developed to achieve the objective. They are the ARIMA model and the Moving Average model. Both the Yala and Maha seasons, as well as the total annual rice output, have their own forecasts. According to the results, it is clear that the Mean Squared Error, Mean Absolute Error and Mean Absolute Percentage Error of the Autoregressive Integrated Moving Average model are higher than the 3-Moving-Average model. In light of the findings, it can be concluded that 3MA is the best technique for predicting rice output in the future among the two models. Buyers and sellers will be able to use the results to plan for rice production in the coming years, as well as identify periods of low output and investigate their causes.

References

Munasingha, M.A. and Napagoda, N.A. (2021) “Trend analysis and forecasting for paddy production in Sri Lanka,” *Applied Economics & Business*, 5(2), p. 1. Available at: https://doi.org/10.4038/aeb.v5i2.33.

Senanayake, S.M.P. and Premaratne, S.P. (2016) “An analysis of the paddy/rice value chains in Sri Lanka,” *Asia-Pacific Journal of Rural Development*, 26(1), pp. 105–126. Available at: https://doi.org/10.1177/1018529120160104.

Suresh, K. *et al.* (2021) “How productive are rice farmers in Sri Lanka? the impact of resource accessibility, seed sources and varietal diversification,” *Heliyon*, 7(6). Available at: https://doi.org/10.1016/j.heliyon.2021.e07398.

Sivapathasundaram, V. and Bogahawatte, C. (2015) “Forecasting of paddy production in Sri Lanka: A time series analysis using Arima Model,” *Tropical Agricultural Research*, 24(1), p. 21. Available at: https://doi.org/10.4038/tar.v24i1.7986.

Jayasooriya, S.P. (2022) "Climate change adaptation and production risks among Paddy Farmers in Sri Lanka," *Sri Lanka Journal of Economic Research*, 10(1), p. 91. Available at: https://doi.org/10.4038/sljer.v10i1.176.

Rahman, N., Hasan, M., Hossain, M., Baten, M., Hosen, S., Ali, M., & Kabir, M. (2016). Forecasting Aus Rice Area and Production in Bangladesh using Box-Jenkins Approach. Bangladesh Rice Journal, 20(1), 1–10. https://doi.org/10.3329/brj.v20i1.30623

Rahman, N. M. F., & Baten, M. A. (2016). Forecasting area and production of black gram pulse in Bangladesh using arima models. Pakistan Journal of Agricultural Sciences, 53(4), 759–765. https://doi.org/10.21162/PAKJAS/16.1892

Sharma, M. R., & Raju, G. (2016). Paddy Production in Telangana State : Current and Future Trends. (March), 436–438.

Sivapathasundaram, V., & Bogahawatte, C. (2015). Forecasting of Paddy Production in Sri Lanka: a time series analysis using ARIMA Model. Tropical Agricultural Research, 24(1), 21. https://doi.org/10.4038/tar.v24i1.7986

Chai, T. and Draxler, R.R. (2014) "Root mean square error (RMSE) or mean absolute error (mae)? – arguments against avoiding RMSE in the literature," Geoscientific Model Development, 7(3), pp. 1247–1250. Available at: https://doi.org/10.5194/gmd-7-1247-2014.

Shekhar, S. and Xiong, H. (2008) "Root-mean-square error," Encyclopedia of GIS, pp. 979–979. Available at: https://doi.org/10.1007/978-0-387-35973-1_1142.

No, T. and Lee, T. (2020) "Wild bootstrap ljung-box test for residual autocorrelation in vector autoregressive models," Journal of the Korean Data And Information Science Society, 31(3), pp. 477–485. Available at: https://doi.org/10.7465/jkdi.2020.31.3.477.

Dudewicz, E.J. (1980) "Heteroscedasticity." Available at: https://doi.org/10.21236/ada087373.