DOI: http://doi.org/10.4038/aeb.v5i2.33

Trend Analysis and Forecasting for Paddy Production in Sri Lanka

Munasingha, M.A.P. and Napagoda, N.A.D.N.

Department of Mathematical Sciences, Faculty of Applied Sciences,
Wayamba University of Sri Lanka, Kuliyapitiya, 60200, Sri Lanka

**Corresponding Author:*
Email: poornimamunasingha@gmail.com

ABSTRACT

Sri Lanka is mainly an agricultural country and about 40 per cent of its working population is engaged in agriculture island-wide. Rice is cultivated during two seasons; *Maha* season (October-March) usually accounts for about 65% of annual production with the remaining 35% coming from the *Yala* season (April-September). The objectives of this study are to investigate the present trend of paddy production and to develop the most appropriate time series models for paddy production in *Yala* and *Maha* seasons separately. The paddy production data were obtained from the Department of Census and Statistics in Sri Lanka from 1952 to 2020. Shapiro-Wilks test was applied to check the normality of the dataset. According to the results, the Mann-Kendall trend test, and Cox – Stuart trend test, were used to detect the presence of trends in the data. It was confirmed that there was an increasing tendency in both seasonal models but the slope of paddy production in the *Yala* season was less compared to the paddy production in the *Maha* season. In this study, Auto-Regressive Integrated Moving Average (ARIMA) method was applied to forecast based on the historical data. The well-fitted ARIMA model for the paddy production of *Yala* season was ARIMA (2,1,1) and paddy production of *Maha* season was ARIMA (2,1,0). The performances of these models were mainly validated with the Akaike Information Criterion (AIC), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE) values. Finally, the best model for the *Yala* and *Maha* seasons was applied separately to predict the values of the variable over the next three years. As the result of forecasting of paddy production is a requirement for planning purposes and the import policy of rice, should be based on this kind of research.

KEYWORDS: ARIMA model, Cox-Stuart trend test, Mann-Kendall trend test

Introduction

Rice was the single most important crop occupying 34 per cent (0.77 /million ha) of the total cultivated area in Sri Lanka. On average 560,000 hectares were cultivated during *Maha* and 310,000 hectares during *Yala* making the average annual extent sown with rice to about 870,000 hectares. Sri Lanka currently produced 2.7 million of rough rice annually and satisfied around 95 per cent of the domestic requirement (oxford business group, 2021).

The focus of this research was identifying the current trends in paddy production and developing separate time series models suitable for paddy production in the *Yala* and *Maha* seasons.

The forecasting of paddy production is a necessity for planning purposes, and the import policy of rice should be based on research forecasting.

Literature Review

The following is a summary of some of the research that has been carried out on paddy cultivation around the world.

Paidipati and Banik (2020) discussed the ARIMA and LSTM-NN models and analyzed trends for rice cultivation in India. The forecasting of rice cultivation was developed from the years 2006 to 2018. The Mann-Kendall test and the Cox-Stuart test were used to identify trends in this research. In this overall study, the LSTM-NN models were more flexible and it was helped to develop more accurate models for predicting the future values than ARIMA models. A disadvantage of the LSTM-NN models cannot be used for every research. It is more effective to determine the method to use depending on the size of the sample (Boulmaiz et al., 2020).

Raghavender (2010) analyzed yearly rice production data for the period of 1956 to 2008 using the ARIMA model. The developed model for rice production was found to be ARIMA (2, 2, 0). This research was limited to finding a predictive model for total paddy production in Andhra Pradesh.

Sivapathasundaram and Bogahawatte (2012) have developed a time series model to detect the long-term trend and prediction for future changes of total paddy production in Sri Lanka for the three leading years. This analysis was utilized the secondary data of the Department of Census and Statistics of Sri Lanka from 1952 to 2010. ARIMA (2, 1, 0) was the selected model in this study with the lowest AIC and BIC values. This research was limited to finding a predictive model for total paddy production in Sri Lanka from 1952- 2010.

Vanitha (2012) developed a time series model to predict the long-term trends and overall changes in the overall paddy production in the Batticaloa District for the three leading years. For this analysis, the secondary data published by the Department of Census and the Annual Report of the Central Bank of Sri Lanka from 1952 to 2009 were used. The developed model for paddy production in the Batticaloa district was found to be ARIMA (2, 1, 0). The ARIMA model was used to match the dataset and to predict the relevant variable in the near future. This research was limited to finding a forecast model for the total paddy production in the Batticaloa District during the period 1952-2009.

The main importance of this research was to study the behaviour of paddy production before developing models that were predicted in production and to gain a better understanding of the current trends in production. The models to be predicted should be constantly updated, including the latest data. Although much research has developed on the total production in Sri Lanka, separate time series models for the *Yala* and *Maha* seasons have not developed. This research was created separately time series models for the *Yala* and *Maha* seasons and emphasized their importance.

Methodology

Data Collection

The paddy production data for *Yala* and *Maha* seasons were obtained separately from the Department of Census and Statistics in Sri Lanka (Department of Census and Statistics, 2020) from 1952 to 2020.

Data Analysis

Analysis of Trend

The Shapiro-Wilk test (Shapiro and Wilk, 1965) was used to test the normality of this research dataset. Subsequently, two non-parametric tendency tests (Mann-Kendall trend test, and Cox-Stuart trend test) were used to identify the tendencies of the data.

The Mann - Kendall test is based on the following test statistic,

$$\boldsymbol{S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} sgn\left(x_j - x_i\right)} \qquad [1]$$

Where;

x_j and x_i : Sequential values
n : Length of the dataset

$$sgn(\theta) = \begin{cases} +1 \text{ if } \theta > 0 \\ 0 \text{ if } \theta = 1 \\ -1 \text{ if } \theta < 1 \end{cases}$$

The parameters of the Mann-Kendall test were used to detect the trend i.e., the presence of upward or downward trend and the magnitude of the trend. There were two parameters, which were useful in obtaining the strength of the trend and the magnitude of the slope. The variance of the Mann – Kendall test statistic (S) in case of ties and no ties, can be obtained as,

$$Var(S) = \begin{cases} \dfrac{\boldsymbol{n(n-1)(2n+5)-\sum_{i=1}^{n} t_i(i)(i-1)(2i+5)}}{\boldsymbol{18}}, \text{ if there are ties} \\ \dfrac{\boldsymbol{n(n-1)(2n+5)}}{\boldsymbol{18}}, \text{ if there are no ties} \end{cases} \qquad [2]$$

Where,

t_i : Number of ties of extent *i.*

For large *N*, the test statistic is,

$$Z_S = \begin{cases} \dfrac{\boldsymbol{S-1}}{[\boldsymbol{var(S)}]^{\mathbf{0.5}}} & \text{for } \boldsymbol{S > 0} \\ 0 & \text{for } S = 0 \\ \dfrac{S+1}{[var(S)]^{0.5}} & \text{for } S < 0 \end{cases} \qquad [3]$$

Where, Z_s follows the standard normal distribution. If there was a trend present, the magnitude of the trend can be obtained with the support of Sen's slope ($\boldsymbol{\beta}$). This Sen's slope was very much associated with the Mann-Kendall test. It was a non-parametric estimate of the slope. This can be obtained as,

$$\boldsymbol{\beta = Median\left(\frac{x_j - x_i}{j-i}\right), \forall\, j > i} \qquad [4]$$

Where, x_j and x_i are the two data values for the time points *j* and *i* (*j*>*i*). The Median of these *N* values of β_i is known to be as Sen's slope estimator, such that

$$Q_i = \begin{cases} \boldsymbol{\beta_{(N+1)/2}} & \text{when } N \text{ is odd} \\ \frac{1}{2}\left(\beta_{N/2} + \beta_{(N+2)/2}\right) & \text{when } N \text{ is even} \end{cases} \qquad [5]$$

The value of *Q* indicates the characteristics of the trend, *i.e.*, if *Q* is positive, it indicates an upward trend and if *Q* is negative, it indicates a downward trend.

The Cox – Stuart trend test is applicable to detect the presence of trends dependent on time, considering the observations are independent. If $X_1, X_2, \dots, X_n$ be the n observations, let

$$c = \begin{cases} \boldsymbol{n/2} & \text{if n is even} \\ (n+1)/2 & \text{if n is odd} \end{cases} \qquad [7]$$

The data values are then paired as $X_1, X_{1+c}, X_2, X_{2+c}, \dots, X_{n-c}, X_n$. Then, the Cox-Stuart test is simply a sign test on these paired data.

Model Building and Fitting

The ARIMA model was used to fit the data set and forecast the concerned variable to the near future (Box et al., 1994). Since annual values of two seasons were used separately, a univariate non-seasonal ARIMA (p,d,q) was used.

ARIMA model was expressed as,

$$\boldsymbol{\emptyset(B)W_t = \theta(B)Z_t} \quad [8]$$
$$\boldsymbol{W_t = \nabla^d X_t = (1-B)^d X_t}$$
$$\boldsymbol{(1-\alpha_1 B-\alpha_2 B^2-\cdots-\alpha_p B^p)\times(1-B)^d X_t = Z_t(1+\beta_1 B+\beta_2 B^2+\cdots+\beta_q B^q})$$

Where,

t : Indexes time
W_t : Value of the time series in period *t*
B : Backshift operator
Z_t : Discrete purely random variable process with mean 0 and variance σ_Z^2
$\theta(B), \emptyset(B)$: Polynomial of order *q* and *p* respectively
X_t : Moving average process of order *q*
d : Number of differences
p : Order of Autoregressive process
q : Order of moving average process

Modelling was accepted by four stages: Identification process, Estimation, Diagnostic testing, and Forecasting. The parameters p and q of the ARIMA model were obtained with the help of significant spikes in autocorrelation and partial autocorrelation functions. Finally, an appropriate Box-Jenkins ARIMA model was fitted.

Evaluation of Models

The model evaluation was mainly performed with the help of Akaike Information Criteria (AIC), Mean Absolute Percentage Error (MAPE), and Root Mean Square Error (RMSE). Using those methods, the best model was selected.

Mean Absolute Percentage Error (MAPE)

$$\boldsymbol{MAPE = \frac{\sum\left|\frac{(y_t-\hat{y}_t)}{y_t}\right|}{n} \times 100, (y_t \neq 0)} \quad [9]$$

Where,

y

y_t : Actual value at time *t*
$\hat{y}_t$: Fitted value
n : Number of observations

$$\boldsymbol{Accuracy = 100 - MAPE} \quad [10]$$

Root Mean Square Error (RMSE)

$$\boldsymbol{RMSE = \sqrt{\frac{\sum_{t=1}^{n}(y_t-\hat{y}_t)^2}{n}}} \quad [11]$$

Where,

y_t : Actual value at time *t*

$\hat{y}_t$: Fitted value

n : Number of observations

Akaike Information Criterion (AIC)

$$\boldsymbol{AIC = -\frac{2\,l}{n} + \frac{2\,k}{n}} \qquad [12]$$

Where;

l : Log-likelihood

n : Number of observations

k : No. of estimated parameters in the model

The performance of the parameters depends upon the lowest AIC value, RMSE value, and the highest Accuracy percentage.

Forecasting

After selecting the most accurate model, paddy production related to the *Yala* and *Maha* seasons was forecasted separately for the three leading years.

Results and Discussion

Analysis of Trends

The trend analysis for the paddy production related to the *Yala* and *Maha* seasons in Sri Lanka was examined using tendency tests (Mann-Kendall trend test, and the Cox-Stuart trend test). The normality of the data was checked before applying the trend tests (Table 1).

Table 1: Shapiro-Wilk Normality Test Results of Paddy Production in Sri Lanka (1952 to 2020)

	***Yala* Season**	***Maha* Season**
p-value	0.00	0.02
w- value	0.93	0.96

It was confirmed that these time-series data of paddy production in the *Yala* and *Maha* seasons were not normally distributed. So, the two non–parametric tests, the Mann–Kendall test, and Cox–Stuart test were used to confirm the exact presence of a trend (Table 2).From these two tests, it was confirmed that there was a tendency for paddy production in the *Yala* and *Maha* seasons, as the tau value of the Mann-Kendall test was close to 1 and the p-value of the Cox-Stuart test was 1. The *p-value* was greater than 0.05, indicating current trends in the data.

Table 2: Trend Analysis with Mann – Kendall and Cox – Stuart Test for the Time-Series Data of the Paddy Production in Sri Lanka (1952 to 2020)

Parameters	Mann–Kendall Trend (tau value)	Sen's Slope (Q)	Cox–Stuart Trend (p-value)
Paddy production in *Yala* season	0.83	20.00	1
Paddy production in *Maha* season	0.84	33.79	1

The magnitude of the slope was obtained by the *Q-value* of Sen's slope. *Q* was positive, it indicated an upward trend. According to the *Q-values*, the slope of paddy production in the *Yala* season was less compared to the paddy production in the *Maha* season (Figure 1).

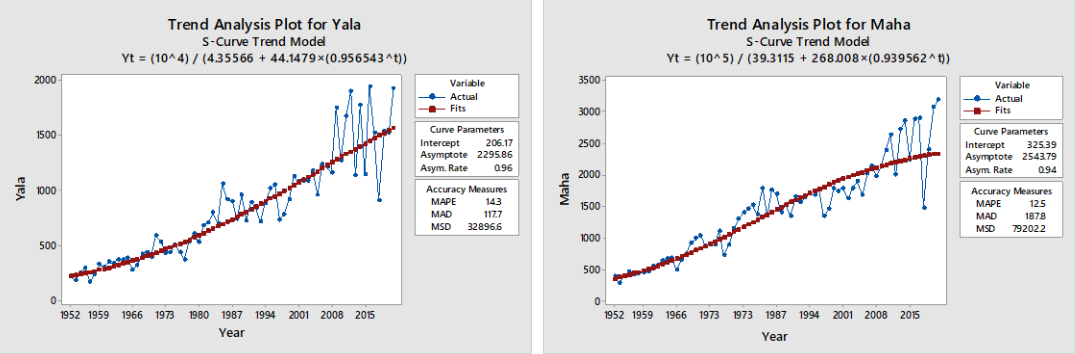

Figure 1: Trend Analysis Plot for Paddy Production in *Yala* and *Maha* Season

Fitting Models with ARIMA

The ARIMA models were developed based on the autoregressive (*p*), moving average (*q*), and the order of differencing (*d*), for making the data stationary. The values of p and q were obtained with help of the significant spikes in the ACF and PACF plots (Figure 2). Using those results tentative models were created. These models were checked using the Box-Jenkins methodology.

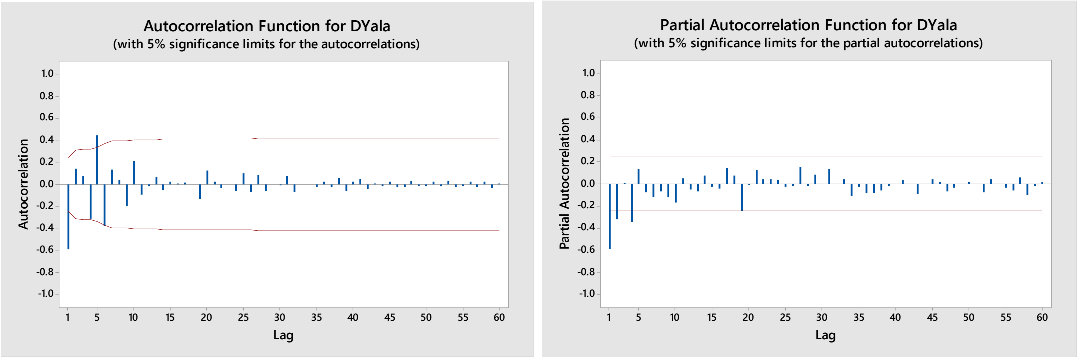

Figure 2: ACF and PACF for Paddy Production in *Yala* Season

In the *Yala* season, the tentative model was ARIMA (2,1,1). Based on this model, four sub-tentative models can be obtained as follows:

ARIMA (2,1,0), ARIMA (1,1,1), ARIMA (0,1,1), ARIMA (1,1,0)

All models were developed for paddy production in the *Yala* season and finally concluded that ARIMA (2,1,1), ARIMA (2,1,0), and ARIMA (1,1,0) models were adequate for paddy production in *Yala* season. Using MAPE, RMSE, and AIC techniques, accuracy can be checked. According to the results, ARIMA (2,1,1) was the best model to forecast the paddy production obtained from the *Yala* season. This model had a better accuracy percentage.

ACF and PACF for Paddy Production in *Maha* Season are shown in figure 3. During the *Maha* season, the tentative model was ARIMA (2,1,3). On the basis of this tentative model, ten sub-tentative were obtained as follows:

ARIMA (2,1,2), ARIMA (2,1,1), ARIMA (2,1,0), ARIMA (1,1,3), ARIMA (0,1,3), ARIMA (1,1,2), ARIMA (0,1,2), ARIMA (1,1,1), ARIMA (0,1,1), ARIMA (1,1,0).

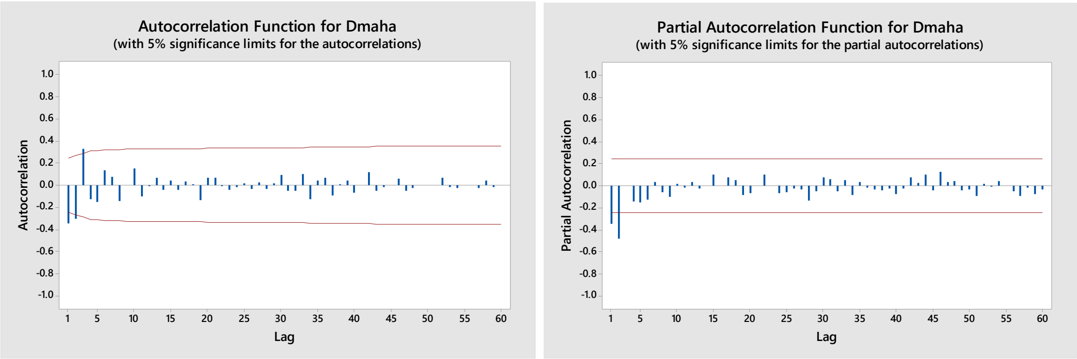

Figure 3: ACF and PACF for Paddy Production in *Maha* Season

All models were developed for paddy production in *Maha* season and finally concluded that ARIMA (2,1,0), ARIMA (0,1,3), ARIMA (1,1,0), and ARIMA (0,1,1) models were adequate for paddy production in *Maha* season. Using MAPE, RMSE, and AIC techniques, accuracy can be checked. According to the results, ARIMA (2,1,0) was the best model to forecast the paddy production obtained from the *Maha* season. This model had a better accuracy percentage. For the *Yala* and *Maha* seasons, Table 3 has listed the accuracy-test values of the best model selected separately from the tentative models.

Table 3: Accuracy Measures for Best Models

Parameters	Model	RMSE	MAPE	Accuracy	AIC
Paddy production in *Yala* season	ARIMA (2,1,1)	308.96	15.60	84.40	53.80
Paddy production in *Maha* season	ARIMA (2,1,0)	495.30	15.80	84.20	57.90

Based on the best models selected for the *Yala* and *Maha* seasons, Table 4 and Table 5 provide the forecast values for the next three years.

Table 4: Forecasted Values for Three Lending Years (*Yala* Season)

Year	Estimated Production (000 Mt)	95% Limits	
		Lower	Upper
2021	1668.60	1270.78	2066.42
2022	1759.06	1338.22	2179.91
2023	1801.68	1317.32	2286.05

Table 5: Forecasted Values for Three Lending Years (*Maha* Season)

Year	Estimated Production (000 Mt)	95% Limits	
		Lower	Upper
2021	2853.03	2333.25	3372.80
2022	3053.91	2482.00	3625.83
2023	3208.27	2624.80	3791.74

Comparison between the Total Production Model and Seasonal Models

The developed model for total paddy production in Sri Lanka was found to be ARIMA (2, 1, 0) (Sivapathasundaram and Bogahawatte, 2012). Our research has created two separate models for the *Yala* and *Maha* seasons. ARIMA (2, 1, 1) was the most suitable model for the *Yala* season, and ARIMA (2, 1, 0) was the most suitable model for the *Maha* season. Compared to the model of the total paddy production, the model for the *Maha* season was similar to the total production model and received a different model for the *Yala* season.

Since two different models are available, production can be measured separately for the two seasons as needed. This is a quicker and easier method than finding the total product and measuring the product for each season. Because the models are designed based solely on the data for each season, the accuracy of those products is higher than the values obtained after finding the total product.

Conclusions

The trend analysis of the paddy production data was showed an increasing production trend for both *Yala* and *Maha* seasons and the trend of paddy production in *Maha* season was higher than *Yala* season. In the *Yala* season, ARIMA (2, 1, 1) was the most suitable model as this model had a better accuracy percentage. During the *Maha* season, ARIMA (2, 1, 0) was the most suitable model as this model had a better accuracy percentage.

The principal objective of developing an ARIMA model for a variable was to generate post-sample period forecasts for that variable. The validity of the forecasted values can be checked when the data for the lead periods become available. The model can be used by researchers for forecasting paddy production in Sri Lanka seasonally.

It is more convenient and accurate as production can be predicted for each season using two separate models. However, it should be updated from time to time with the incorporation of current data.

As further developments in this research, it can be extended to compare the major cultivable states in Sri Lanka. This makes it possible to estimate the paddy cultivation values seasonally for each province.

References

Boulmaiz, T., Guermoui, M., & Boutaghane, H. (2020). Impact of training data size on the LSTM performances for rainfall–runoff modeling. *Modeling Earth Systems and Environment, 6*, 2153-2164.

Box. G.E.P., Jenkins. G.M., and Reinsel, G.C. (1994). Time series analysis. *Forecasting and control.* (3rd ed.). Prentice Hall Press, Englewood cliffs, New Jersey.

Department of Census and Statistics (2020). Paddy extent sown and harvested, average yield and production. Retrieved from http://www.statistics.gov.lk/Agriculture/StaticalInformation/rubpaddy on December, 2020.

Oxford Business Group (2021). *Sri Lanka tackles challenges to rice production to end reliance on imports.* Retrieved from https://oxfordbusinessgroup.com/analysis/ self-sufficiency-goals-riceindustry-must-overcome-severalchallenges-increaseproduction-andend on June, 2021.

Paidipati, K. K., & Banik, A. (2020). Forecasting of Rice Cultivation in India–A Comparative Analysis with ARIMA and LSTM-NN Models. *EAI Endorsed Transactions on Scalable Information Systems, 7*(24).

Raghavender, M. (2010). Forecasting paddy production in Andhra Pradesh with ARIMA model. *Int. J. of Agric. and Stat. Sci, 6*(1), 251-258.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika, 52*(3/4), 591-611.

Sivapathasundaram, V., & Bogahawatte, C. (2012). Forecasting of paddy production in Sri Lanka: a time series analysis using ARIMA model. *Tropical Agricultural Research, 24* (1), 21-30.

Vanitha, S. (2012). Paddy production pattern and future forecasting of Batticaloa district: a time series analysis. *Proceeding of 2nd International Symposium, Postgraduate Institute of Agriculture, University of Peradeniya. Sri Lanka.* 197 -199.