A seasonal ARIMA model to forecast monthly potato yield in Sri Lanka

MSH Perera[1#], AWSP Karunarathne[2], SS Hewage[3], LBU Sandamali[4] and NV Chandrasekara[5]

Department of Statistics & Computer Science,
Faculty of Science,
University of Kelaniya,
Kelaniya 11600,
Sri Lanka
[#]<hirunisperera94@gmail.com>

Abstract –*The potato is an extensively cultivated tuber crop in the world. In Sri Lanka, cultivation has been established in four regions: Nuwara Eliya, Badulla, Jaffna and Puttalam. The objective of this paper is to analyze and forecast the monthly potato production in Sri Lanka. The monthly data from December 2005 to November 2017 were considered for this study and it consisted of 144 observations including 10 missing values. The missing values were estimated using missing values imputation techniques. Since the time series plot of potato yield shows a clear seasonal pattern, it was decided to fit a Seasonal Autoregressive Integrated Moving Average (SARIMA) model. The best fitted model was SARIMA* $(0,0,1)\ (1,1,1)_4$ *which resulted in the minimum Akaike's Information Criterion (AIC). It depicted the best performance out of all the suspected models. The forecasting accuracy of the above model was measured with Mean Absolute Error (MAE) which was 1.67. Therefore, it can be concluded that the SARIMAmodel is accurate in predicting the monthly potato yield in Sri Lanka. The model would be importantto all the stake holders of potato cultivation in the country.*

Keywords: SARIMA, AIC, ACF, PACF

I. INTRODUCTION

Potato is a major food crop grown in Sri Lanka. It is popularly known as 'the king of vegetables'.Even in globe, the potato production has grown steadily. Potatoes are used for variety of purposes such as food, medicine, beauty care and etc.The potato production in Sri Lanka is mainly targetedat the local food market, even though, weather, poor storage conditions, varying availability of good seed and diseases are the major problems limiting potato production in SriLanka. (Sathiamoorthy *et al.*, 2008) If we look into the origination of the potatoes, it was originated in Andes highlands in Peru and Europe and it was introduced to Sri Lanka in 1850. At present potato is extensively cultivated in the district of Nuwara Eliya in two major seasons., 'Yala' (Feb-July) and 'Maha.' (Aug-Dec).Jaffna and Puttalam are the other two districts where the potatoes grow in lesser.'Yala' and 'Maha' seasons are the peak periods of potato planting.Since potato production demands a heavy investment and it is an important crop in Sri Lanka with higher consumer preference, it is topical to study on the yield of the potato crop to enhancethe potato production in Sri Lanka.

Table 1.Potato production in Sri Lanka (2005-2017)

Cultivation Year	National Total
2005/2006	78827
2006/2007	75263
2007/2008	74398
2008/2009	60848
2009/2010	51308
2010/2011	64359
2011/2012	75352
2012/2013	68778
2013/2014	81661
2014/2015	70377
2015/2016	80458
2016/2017	52998

Table 1 illustrates the potato production in Sri Lanka during the years 2005 to 2017. The maximum and the minimum potato production can be observedin the years of 2013/2014 and 2009/2010 respectively.

The objective of this paper is to review, analyse and forecast the monthly potato production in Sri Lanka and to observe the behaviour of the potato yield with respect to the relevant months.

Studying the monthly potato yield in Sri Lanka will be important for formulations and implementations of the potato cultivation. Further, the prediction of potato crop yield prior to the harvest period can be very useful in pre-harvest and marketing decision making. GJ Scott et al(Scott and Suarez, 2011) presents a study on growth rates for potato in India and their implications for industry implying the importance of the potato production.

Many literatures can be found on different areas using Seasonal ARIMA models.A Seasonal ARIMA model with 12-month period was fitted to forecast the number of dengue cases in Campinas in Brazil.(Martinez, Silva and Fabbro, 2011)Another study proposed a hybrid methodology that exploits the strength of the SARIMA

model and the support Vector Machines (SVM) model in forecasting seasonal time series. The seasonal time series data of Taiwan's machinery industry production values were used to examine the forecasting accuracy.(Chen and Wang, 2007)Also a new hybrid model for short-term power forecasting of a grid-connected photovoltaic plant is introduced with two well-known methods: the Seasonal ARIMA and the SVMs.(Bouzerdoum, Mellit and Massi Pavan, 2013)The current study is performed after reviewing these literatures.

II. METHEDOLOGY AND EXPERIMENTAL DESIGN

This section describes theories and techniques used in this study.

A.Data Collection

The data required for this study were gathered from the Department ofCensus and Statistics.The data set contained the monthly potato yield in Sri Lanka from December 2005 to November 2017.

B.Missing Values Estimation

The most common data pre-processing technique is data cleaning. It includes interpretation of missing values, smoothing noisy data, identifying and removing outliers and resolving inconsistencies. Missing values imputation technique plays a prominent role among them. In this study the data set contained 144 observations and there were 10 missing values. Here 4 missing values techniques namely mean imputation, linear interpolation, log-linear interpolation and exponential smoothing were used to impute the missing values.

C.Stationarity

A stationary time series can be identified as a time series whose statistical properties, mainly mean and variance are constant over the time. Graphical methods and statistical tests are commonly used to identify whether a time series is stationary or not. The most accurate way is to use statistical tests and in this study Kwiatkowski-Phillips-Schmidt-Shin (KPSS), Augmented Dickey Fuller (ADF) and Phillips Perron (PP) tests were used to check the stationary of the time series.

D.Time Series Forecasting Methods

A time series is a collection of observations where they are indexed in time order. There can be numerous time series where the sequence observations were varied as weekly, monthly, quarterly or annually etc.The ultimate goal of time series analysis is to derive the future behaviour of the time series considering the past behaviour. In this study a univariate time series approach has been used in forecasting.

E.Univariate Time Series Approach: SARIMA Model

Seasonality in a time series is a regular pattern of changes that repeats over constant (S) time period, where S defines the number of time periods until the pattern repeats again. In a seasonal ARIMA model, seasonal AR and MA terms predictusing data values and errors at times with lags that are multiples of S. Seasonality usually causes the series to be non-stationary because the average values at some particular times within the seasonal span (months, for example) may be different than the average values at other times.

Seasonal Autoregressive Integrated Moving Average (SARIMA) is an extension of ARIMA that explicitly supports univariate time series data with a seasonal component. Configuring a SARIMA requires selecting hyper parameters for both the trend and seasonal elements of the series.

Seasonal ARIMA model can be denoted as:

$$ARIMA(p,d,q)(P,D,Q \quad (1)$$

Where,

p: trend auto regression order
d: trend difference order
q: trend moving average order
P: seasonal autoregressive order
D: seasonal difference order
Q: seasonal moving average order
s: the number of time steps for a single seasonal period.

F.Forecasting Accuracy

To know whether how well the model has been performed it is needed to consider the difference between actual and the forecasted values. It is essential to minimize the difference between actual and the forecasted values because the model performance relies on that. That is the smaller the difference, the better the model is. In this study Mean Absolute Error (MAE) has been used to assure the forecasting accuracy.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i -}{y_i} \quad (2)$$

Where,

= the actual value
= the fitted value
= number of observations

III. RESULTS

The results obtained from this study are illustrated in this section

Figure 1 implies the variation of the original potato yield data with time.

A.Missing values estimation

Table 2implies the four missing value imputation techniques, theMean Absolute Percentage Error (MAPE) and Root Mean Square Error(RMSE) values obtained.

Table 2. Missing values imputation techniques

Imputation Method	**MAPE**	**RMSE**
Mean imputation	1·12517	0·96558
Linear Interpolation	**0·52130**	**0.85213**
Log-Linear Interpolation	1.02365	1·24194
Exponential Smoothig	2·25502	3·25361

As indicated in Table 2, Linear interpolation method was identified as the best missing value imputation method with minimum MAPE and RMSE values0.52130 and 0.85213 respectively.Therefore, the study was proceeded after imputing the missing values with linear interpolation method.

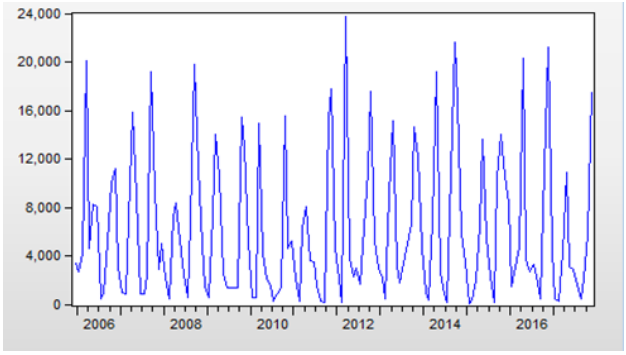


Figure 2. Variation of potato yield with time inlinearinterpolation method

Figure 2 describes the variation of monthly potato yield in Sri Lanka where the missing values were estimated using linear interpolation method. It can be clearly observed that there is a seasonal pattern and the data fluctuates around a horizontal line with constant variability. Hence, the data set seems to be stationary.

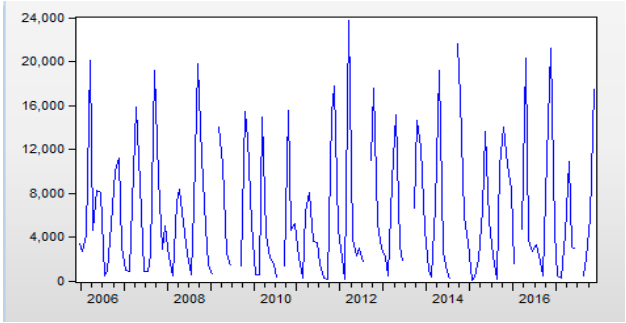
Further to check the stationarity of the data the KPSS, ADF and PP tests were followed.

B.Stationarity Tests Results

For ADF and PP tests the hypothesis is given below,

H_0: *the series is not stationary*

H_1: *the series is stationary*



For KPSS test the hypothesis is given below,

H_0: *the series is stationary*

H_1: *the series is not stationary*

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-11.25201	0.0000
Test critical values:	1% level	-3.477144	
	5% level	-2.881978	
	10% level	-2.577747	

Figure 3.ADF test result

According to the ADF test portrayed in Figure 3, p value 0.000 is less than 0.05. Thus null hypothesis is rejected. Therefore, the series is stationary at 5% level of significance.

		Adj. t-Stat	Prob.*
Phillips-Perron test statistic		-12.25589	0.0000
Test critical values:	1% level	-3.483312	
	5% level	-2.884665	
	10% level	-2.579180	

Figure 4. PP test result

The PP test results are displayed in Figure 4. Since the p value 0.000 is less than 0.05 the null hypothesis is rejected. So, the series is stationary at 5% level of significance.

		LM-Stat.
Kwiatkowski-Phillips-Schmidt-Shin test statistic		0.120544
Asymptotic critical values*:	1% level	0.739000
	5% level	0.463000
	10% level	0.347000

Figure 5.KPSS test result

Figure 5, indicates the test statistic value 0.1205 is less than the critical value 0.463. It doesn't lie in the rejection

region. Since the test statistic is less than the critical value the null hypothesis is not rejected. Therefore, the series is stationary at 5% level of significance.

All three tests confirm that the series is stationary at 5% level of significance.

C.Univariate Time Series Approach – SARIMA

For the present study, it was decided to fit a Seasonal ARIMA model since the time series plot indicated a seasonal pattern.

1) Correlogram:

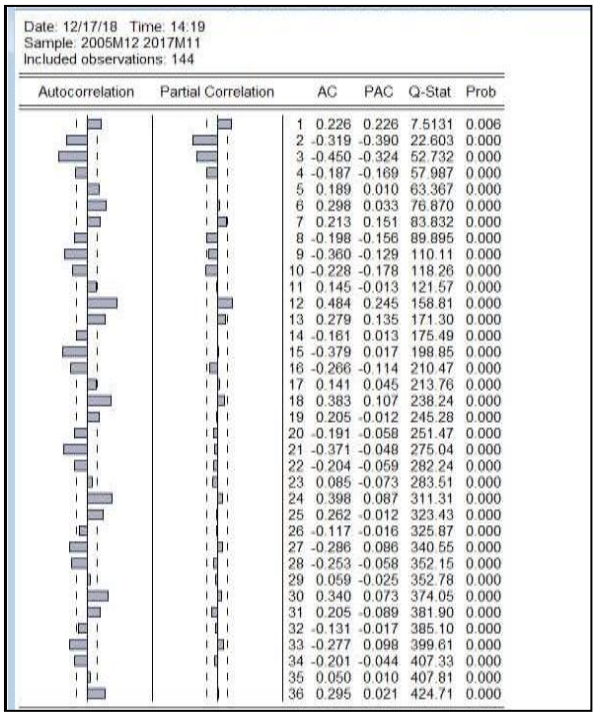


Figure 6.Correlogram

According to Figure 6 it can be clearly identified that there is a seasonal pattern in the Auto correlation function (ACF) plot. The significant cuts off lags varies 4, 8, 12 likewise and using this result it was identified that there is a seasonality with period of 4 months in the data.

2)*Model Fitting:*

A Seasonal ARIMA model was developed to model the monthly potato yield data.

Table 3exhibits some of the fitted models with their AIC values. The best model was considered as SARIMA (0,0,1) $(1, 1, 1)_4$ which has the minimum AIC.

Table 3.Fitted SARIMA models

	AIC
SARIMA (1,0,0) (1,1,1) 4	3.4206
SARIMA (2,0,0) (1,1,1) 4	3.3850
SARIMA (0,0,1) (1,1,1) 4	**3.3684**
SARIMA (0,0,2) (1,1,1) 4	3.3733
SARIMA (0,0,3) (1,1,1) 4	3.3888
SARIMA (0,0,4) (1,1,1) 4	3.6428
SARIMA (1,0,1) (1,1,1) 4	3.8478
SARIMA (2,0,1) (1,1,1) 4	3.3787
SARIMA (3,0,1) (1,1,1) 4	3.5543
SARIMA (1,0,2) (1,1,1) 4	3.6300
SARIMA (2,0,2) (1,1,1) 4	3.3885
SARIMA (3,0,2) (1,1,1) 4	3.5155
SARIMA (1,0,3) (1,1,1) 4	3.9323
SARIMA (2,0,3) (1,1,1) 4	3.7776
SARIMA (3,0,3) (1,1,1) 4	3.5901
SARIMA (4,0,1) (1,1,1) 4	3.4193
SARIMA (4,0,2) (1,1,1) 4	3.3877
SARIMA (4,0,3) (1,1,1) 4	3.8650

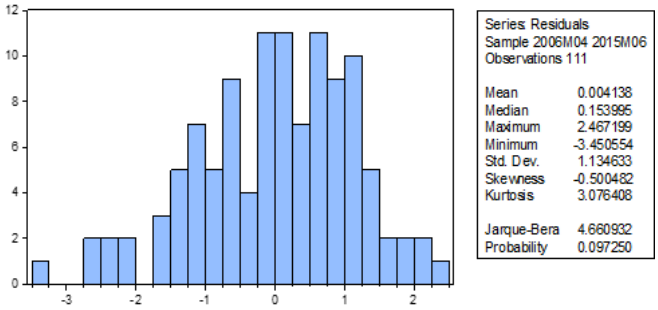
D.Model Adequacy Checking

Model adequacy checking is used to check the validity of a model.In this study normality test and the heteroscedasticity test were drawn in order to check whether the model is adequate.

1) Normality Test:

The hypothesis for the normality test is given below,

H_1:*errors are normally distributed*



H_1:*errors are not normally distributed*

Figure 7 indicates the probability value 0.097250 is greater than 0.05. Therefore, the errors are normally distributed at 5% level of significance.

2) Heteroscedasticity:

H_1:*Not presence of ARCH effect*

Figure 7. Histogram for residuals

Heteroskedasticity Test: White

F-statistic	0.747110	Prob. F(9,44)	0.6642
Obs*R-squared	7.158263	Prob. Chi-Square(9)	0.6206
Scaled explained SS	7.227193	Prob. Chi-Square(9)	0.6135

H_1:*Presence of ARCH effect*

Figure 8. Heteroscedasticity test result

According to Figure 8 the p value 0.6642 is greater than 0.05 indicates that do not reject the null hypothesis. Therefore, there is no ARCH effect at 5% level of significance.

As the fitted model SARIMA $(0,0,1)\ (1,1,1)_4$ is adequate it was used to forecast out of sample data to asses forecasting accuracy.

E.Forecasting

The monthly potato yield can be forecast as shown in the above Figure 9. Forecasting accuracy of the SARIMA $(0,0,1)\ (1,1,1)_4$ was measured with RMSE and MAE values 1.9608 and 1.6704 respectively.

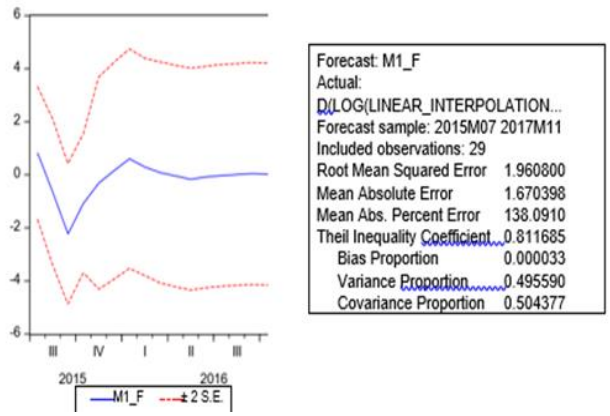
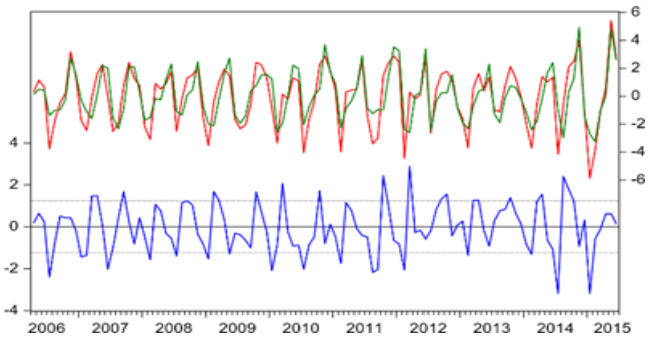


Figure 9. Forecasting with SARIMA $(0,0,1)\ (1,1,1)_4$



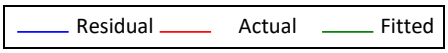


Figure10.Actual vs fitted graph with residual plot

Figure 10 illustrates the plot of predicted potato yield values against the actual potato yield values to demonstrate the correlation of accuracy. It is clear that the performance of the SARIMA $(0,0,1)\ (1,1,1)_4$model selected is quite impressive and the actual and predicted values seems to be related to each other.

IV.DISSCUSSION AND CONCLUSION

In this study a Seasonal ARIMA model was fitted to forecast the monthly potato production in Sri Lanka. Since it was observed a seasonal pattern in the data set a Seasonal ARIMA model was developed.

The data set was consisted with 144 data points including 10 missing values. In order to estimate the missing values, four missing values imputation techniques were used.

The best missing value imputation method was linear interpolation method. Several models were fitted and the best model was identified using the minimum AIC value. Here the SARIMA $(0,0,1)\ (1,1,1)_4$was the best model with minimum AIC 3.3684. And then 20% of the data was used to forecast the potato yield and the root mean absolute error of the forecasted model was 1.67.

Also, a model adequacy checking was carried out to inspect the validity of this model and it was concluded that the model was adequate because the residuals were normally distributed with a constant variance.

In this study, only the potato yield was considered to forecast the future production. But a more accurate model for potato can be fitted considering the other factors such as metrological factors. Therefore, this study can be further improved byconsidering all those factors in future.

REFERENCES

Bouzerdoum, M., Mellit, A. and Massi Pavan, A. (2013) 'A hybrid model (SARIMA-SVM) for short-term power forecasting of a small-scale grid-connected photovoltaic plant', *Solar Energy*. Elsevier Ltd, 98(PC), pp. 226–235. doi: 10.1016/j.solener.2013.10.002.

Chen, K. Y. and Wang, C. H. (2007) 'A hybrid SARIMA and support vector machines in forecasting the production values of the machinery industry in Taiwan', *Expert Systems with Applications*, 32(1), pp. 254–264. doi: 10.1016/j.eswa.2005.11.027.

Martinez, E. Z., Silva, E. A. S. da and Fabbro, A. L. D. (2011) 'A SARIMA forecasting model to predict the number of cases of dengue in Campinas, State of São Paulo, Brazil.', *Revista da Sociedade Brasileira de Medicina Tropical*, 44(4), pp. 436–40. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21860888.

Sathiamoorthy, K. *et al.* (2008) 'Potato production in Sri Lanka', *American Potato Journal*, 62(10), pp. 555–564. doi: 10.1007/bf02854402.

Scott, G. J. and Suarez, V. (2011) 'Growth rates for potato in India and their implications for industry', *Potato Journal*, 38(2), pp. 100–112.