Development of Forecasting Model for Paddy Cultivation in Sri Lanka

Karunarathna VLAD[1]
Ms. Fransisco GS[2]

ABSTRACT

The paddy field cultivation is one of the main employments in the rural areas of Sri Lanka. There are two seasons of paddy cultivation called "Yala" and "Maha". Rice is the staple food of the nation, and evidently paddy cultivation has taken prominence over all other food crops from ancient times in Sri Lanka. Paddy is the largest employer of seasonal labor (more than 10% of total population) in the domestic agricultural sector in Sri Lanka. Annual paddy production data were collected by Department of census through croup cutting survey. In this research a time series analysis is done to obtain the best fitted model for total paddy production in Sri Lanka. Also a regression equation was obtained to identify the factors which are affecting to the total paddy production in Sri Lanka. The analysis is totally based on secondary data which are obtained by the Department of census. Number of plus, miners and neutral variables were identified which effect the paddy production and these variables are controlled to increased paddy production in Sri Lanka. Considering the time series analysis of total paddy production in Sri Lanka the best model obtained was ARIMA (1, 1, 0). By using the time series model paddy production in Sri Lanka can be forecasted for the next five years.

KEY WORDS: Fitted model, Paddy production, Regression equation

INTRODUCTION

The Department of Census and Statistics (DCS) is the National Statistical Office in Sri Lanka which functions as the central government agency responsible for the collection, compilation, analysis and dissemination of reliable and timely statistical data relating to population and housing, agriculture, industries, trade and services, national accounts, prices and other social and economic activities of the country for the purpose of planning, formulation and implementation of development programs. In order to achieve this, the following subjects and functions have been assigned to the DCS but the scopes of these activities have increased very much over time.

Production of adequate quantities of food, especially rice to meet the requirements of the expanding population is one of the challenges faced by Sri Lanka. The country's present population is around 20 million and the rate of population growth is 25:2 per 1000. With the present population growth, farmers of Sri Lanka face a formidable challenge to produce more and more rice each year to make the country self - sufficient in rice. To measure the progress of paddy cultivation in Sri Lanka and for future development of the infrastructure facilities required to improve the cultivation of paddy, to increase the production, it is necessary to have accurate information on acreages, average yield and production of paddy.

Such information and the forecasts for the future years will no doubt be useful to policy makers, planners and research workers. To check whether the country has reached self-sufficiency, if not, to decide the amount of rice to be imported to meet the requirements, accurate and timely information on production we required.

[1]Graduate, Department of Mathematical sciences, Faculty of Applied Sciences, Wayamba University of Sri Lanka.

2Senior Lecturer, Department Mathematical sciences, Faculty of Applied Sciences, Wayamba University of Sri Lanka.

RESEARCH OBJECTIVE

The main objective of this research is to identify a forecast model and forecast the total paddy production of Sri Lanka for the next five years. Also, this research will analyze the total paddy production using regression analysis with the available variables such as asweddumized area, sown area, harvested area, etc.

LITERATURE REVIEW

The Correlation transformer is used to determine the extent to which changes in the value of an attribute are associated with changes in another attribute. The data for a correlation analysis consists of two input columns. Each column contains values for one of the attributes of interest. (Draper and Smith, 1998). The Correlation transformer can calculate various measures of association between the two input columns. The data in the input columns also can be treated as a sample obtained from a larger population, and the Correlation transformer can be used to test whether the attributes are correlated in the population. In this context, the null hypothesis asserts that the two attributes are not correlated, and the alternative hypothesis asserts that the attributes are correlated. The Correlation transformer calculates any of the following correlation-related statistics on one or more pairs of columns. (Lindley, 1987)

In regression analysis, it is also of interest to characterize the variation of the dependent variable around the regression function, which can be described by a probability distribution. (Fox , 1997, Julian)

Time series analysis comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. Time series forecasting is the use of a model to forecast future events based on known past events to predict data points before they are measured. (Audi , 1996)

In study of forecasting paddy yield using season time series model 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 of rice yield. However, it should be updated from time to time with incorporation of current data. (Raghavender, 2009)

METHODOLOGY

This research develops a forecasting model for paddy cultivation in Sri Lanka. The method used for forecasting is time series analysis. Also this research will develop a regression equation for the total paddy production in Sri Lanka.

The research design will be "Correlation research design" because the goal of this research is to assess the relationship among the paddy production and variables such as asweddumized area, sown area, harvested area, etc. There are both advantages and disadvantages of this research design. It allows testing of expected relationships among variables for prediction but here inferences cannot be drawn about casual relationship among variables.

DATA COLLECTION AND ANALYSIS

This research will use the secondary data from the department of census and statistics which collected under following methods.

Paddy Statistics

Statistics on paddy cultivation is prepared seasonally by the department and includes district wise estimates of extents, average yields and production. The asweddumized, cultivated and harvested extents are obtained by a complete enumeration of all the paddy parcels in the country. A crop cutting survey adopting a stratified multi-stage random sampling design is conducted seasonally to estimate the average yield by mode of

irrigation, at districts level. A representative sample of size approximately 5000 paddy parcels is used for this survey. The paddy production is computed as the product of the average yield and net harvested extent.

Source of Data

The aspects of paddy cultivation discussed here are certain farming practices, acreage sown and harvested, yield, and production. Data for the analysis of these aspects are drawn from two main sources, namely the complete enumeration of parcels and the bi-annual crop estimation survey, both conducted by the Department of Census and Statistics. The complete enumeration is the source of data on acreage aswedumized, sown, and harvested, while the crop estimation survey provides data on yield and production of paddy as well as on farming practices.

Complete Enumeration of Parcels

The department of census and statistics collect data on sown and harvested paddy acreage on a complete enumeration basis. The unite of enumeration is the "parcel", on what is known as the p1 form, details are collected such as the name of cultivators, the extent asweddumized, extent harvested, the extent double cropped and extent sown for each season in each cultivation year. Data on paddy acreage obtained in form p1is classified by mode of irrigation major, minor and rainfed. Data obtained at village level are consolidated up to Assistant Government Agent's (AGA) Division, and then up to district and the whole island for each season.

Crop estimation survey

'In view of the importance of the paddy cultivation in the agricultural economy of the Island, a survey was conducted by the department of Census and Statistics in 1951 to collect information which is essential for formulation, implementation and monitoring of agricultural development projects. Since then this survey has been conducted bi-annually, throughout the country and it provides paddy statistics on the basic of results of crop cutting experiments numbering about 5000 in Maha seasons and 4000 in Yala seasons. This survey is the main source of quantitative information on paddy cultivation in Sri Lanka.

The sampling design adopted for the survey is a stratified multistage random design with districts and AGA Divisions being administrative strata. The AGA Divisions are sub-stratified by mode of irrigation. In each stratum the number of villages for the survey is determined in proportion to the area, with due consideration given to practical issues such as required accuracy, availability of officers etc. As a whole, about 2700 villages are selected for the survey in Maha season and 2000 villages for Yala season. In each selected village, two parcels are selected for the survey. Information on farming preparation of land etc. is also collected from the cultivators whose fields are selected for the survey. In addition, data on farming practices are collected from four other parcels in each selected village.

The time series analysis is used for forecasting process. This process consists with four main steps. They are; Identification, Estimation, Diagnostic checking and Forecasting.

The regression analysis is used to obtain a regression equation for total paddy production in Sri Lanka.

The input data is plotted for examine transformation, differencing and seasonality. Then autocorrelation of the basic series was calculated. After deciding whether differencing is need or not the ACF (Autocorrelation Function) and PACF (Partial Autocorrelation Function) were calculated for the differenced series. By examined the ACF and PACF a model was identified. Then parameter estimation was performed to calculate parameter values and goodness of fit statistics. The adequacy of the model was checked to determine

whether modifications are necessary. After performing necessary modifications forecasting is perform.

For regression analysis, first obtained the scatter plot of paddy production vs. other independent variables to view the relationship graphically. Then by using the correlation analysis the relationship between paddy productions vs. other variables was checked. Then by using step wise regression analysis fitted data set (ignore the correlated variables) was obtained. Finally using the fitted data set the regression equation was obtained.

Introduction about the Variables

The variables are based on data, collected by the department of census and statistics using the crop cutting survey.

Production : Annual total paddy production in Sri Lanka. (In thousand metric tons)
Sown area : Annual total sown area of paddy cultivation in Sri Lanka (In Hectares)
Harvested area : Annual total Harvested area of paddy cultivation in Sri Lanka (In Hectares)
Average : The annual average value of paddy production in kilograms per one hectare
Asuweddumized extent : Annual total area prepared for paddy cultivation in Sri Lanka (In Hectares)

RESULTS AND DISCUSSION

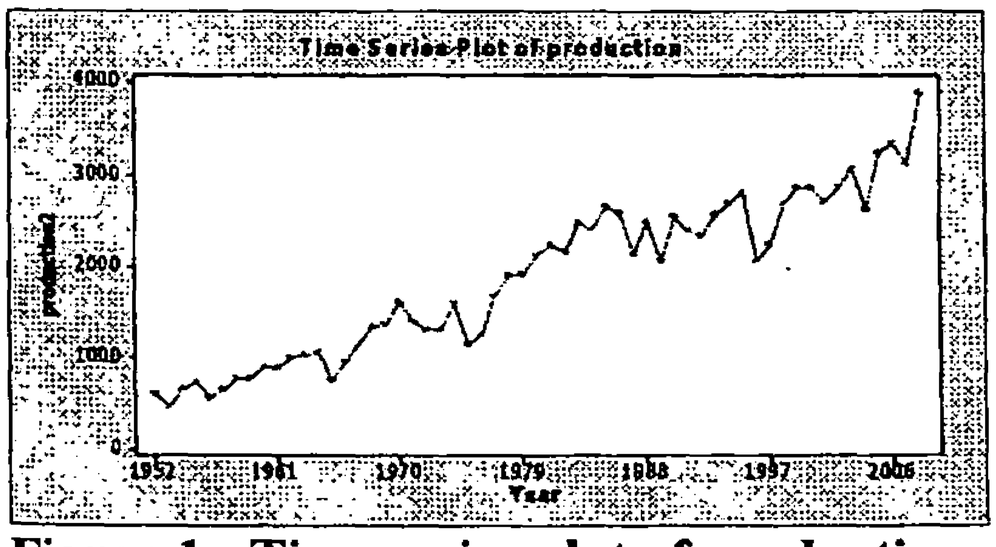

Figure 1 - Time series plot of production

According to the figure the value of total annual production has a positive increase. So there exists a trend. There isn't any repeating pattern. Therefore there is no seasonality. So by using the time series plot we can predict that in this forecasting model do not include seasonality part.

The paddy production in Sri Lanka has been fluctuating with sharp drops in the years 1953, 1965, 1975, 1982 and 1984. The increase in production since 1956 has fairly rapid until 1964. In 1965 there is a decrease in total paddy production in Sri Lanka. The highest value production recorded in 2008 and lowers value recorded in 1953. In the time duration between 71-72 and 88-89, there is a decrease in paddy production due to the political background in the country.

By using Autocorrelation Function for 1st order differenced data and Partial Autocorrelation Function for 1st order differenced data, ARIMA model for the total paddy production was fitted, the parameters consider as p=1, q=1 and d=1. The fitted model was ARIMA (1, 1, 1).

Testing the significance of the parameters for ARIMA (1, 1, 1)
For test the hypothesis
H 0: Parameter value is zero
H 1: Parameter value is not zero

Table 1 - Final Estimates of Parameters

Type	Coef	SE Coef	T	P
AR 1	0.3693	0.1549	2.38	0.021
MA 1	0.9569	0.0820	1.67	0.000
Constant	32.161	2.136	5.06	0.000

For a good model the p value must be less than the significant level α. Therefore the p value must be less than 0.05 hence, constant and all parameters are significant at significant level $\alpha = 0.05$.

To check the Randomness of the residuals in the model the Box-Pierce Chi-Square statistic for the total paddy production was conducted.
Randomness of the residuals for ARIMA (1, 1, 1)
For test the hypothesis
H 0: residuals are uncorrelated.
H 1: residuals are correlated.

P-value <significant level, reject H0

Table 2 - Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag	12	24	36	48
Chi-Square	8.6	18.7	27.9	32.0
DF	09	21	33	45
P-Value	0.476	0.604	0.717	0.928

By looking at the Box-Pierce Chi-Square statistic, all p-values are greater than significant level of α (0.05) therefore, we fail to reject null hypothesis. This means the residuals are uncorrelated.

By considering the ACF and PACF for the residual of model ARIMA (1,1,1) more than 95 % of the data of both ACF and PACF plots are within the significant region. All values in ACF and PACF are approximately equal to zero. This implies that residuals were random.

Therefore the ARIMA (1,1,1) model can be accepted as a good model.

Some alternative models were founded using the above procedure. Those are ARIMA (1, 1, 0), ARIMA (0, 1, 1) and ARIMA (2, 1, 0).

Considering the Normal probability plots for the residuals and Histogram of the residual we can decide the model ARIMA (1, 1, 0) found as the best model because the best align normal probability plot and the best Bell shape of the Histograms of residual found at the ARIMA (1, 1, 0) consider to the other models.

According to the scatter plot of production vs. independent variables, there is small variability between productions vs. variables. As there exist a strong positive linear relationship between all variables. The data plot lies very close to the regression line in total average. Therefore the variability is very small in total average. There are no outliers in four variables. When the variables are increases, total paddy production connectedly increases; it is common to four independent variables. By using Matrix Plot of production vs. independent variables, presence of a good positive relation between the productions vs. independent variables can be stated.

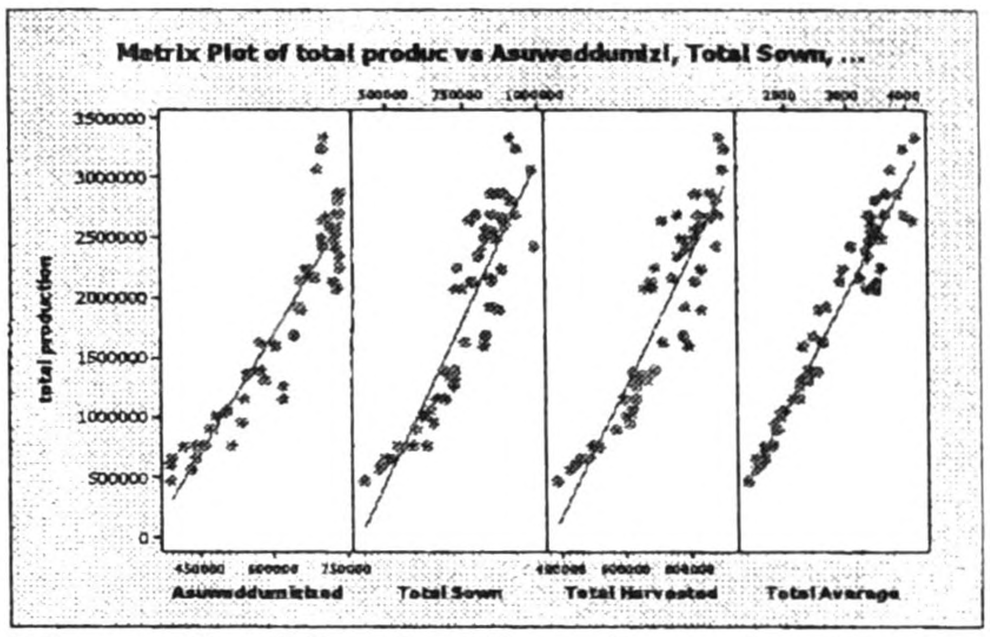

Figure 2 - Matrix Plot of production vs. independent variables

Correlation analysis was used to identify the relationship between the variables in calculated approach. By using a Hypothesis testing check whether there exist a relationship or not. Null Hypothesis has taken as the correlation between two variables is equal to zero.

Table 3 - Correlation between Production and variables

Production Vs. Variable	Pearson correlation	Calculated T value	Table T value	Relationship
Asuweddu	0.926	17.85	2.11	Exist
Sown	0.911	16.08	2.11	Exist
Harvested	0.923	17.46	2.11	Exist
Average	0.973	18.40	2.11	Exist

Table 4 - Step wise regression table

Step	1	2	3
Constant	-896076	-1534483	-1305354
Total Average	974	669	792
T-Value	30.59	22.96	19.71
P-Value	0.000	0.000	0.000
Total Harvested		2.11	2.43
T-Value		12.54	14.42
P-Value		0.000	0.000
Asuweddumizized			-1.28
T-Value			-3.98
P-Value			0.000
S	193448	97382	85909
R-Sq	94.64	98.67	98.98
R-Sq(adj)	94.54	98.62	98.92
Mallows C-p	215.2	17.2	3.5

By the regression analysis we can get the regression equation as,

Production = - 1305354 - 1.28 Asuweddumizized + 2.43 Harvested+ 792 Average

R^2 is a statistic that will give some information about the goodness of fit of a model. In regression, the R^2 coefficient of determination is a statistical measure of how well the regression line approximates the real data points. In the step wise procedure we get an R^2 value of 94.64 in the first step, an R^2 value of 98.67 in the second step and impressive R^2 value of 98.98 in the third step. An R^2 value of 98.98 indicates that the regression line all most fits the data.

CONCLUSION

According to the findings of this research which are based on the data collected by department of census and statistics a number of plus, miners and neutral variables can be identified which effect the paddy production. The plus variables being "average production" and "harvested area", miners being "Asweddumized area" and sown area is being neutral variable. Trying to control these variables in the correct way can lead to increased paddy production in Sri Lanka.

The main variable being the average production, it is vital that the administration should encourage the farmers to produce more paddies by giving them new improved seeds, fertilizer, technical advice, and other benefits, and also tries to reduce wastage.

Considering the time series analysis of total paddy production in Sri Lanka the best model can be obtained as ARIMA (1, 1, 0). By using the time series model paddy production in Sri Lanka can be forcasted for next five years. The forecasted values are shown in following Table

Table 5 - Forecast values for the total paddy production in Sri Lanka

Period	Forecast	Lower	Upper
2009	3662.33	3166.53	4158.12
2010	3823.23	3242.60	4403.86
2011	3838.25	3145.53	4530.97
2012	3910.24	3134.90	4685.58
2013	3959.98	3105.22	4814.74

If correctly used this research, government can contribute in achieving the final goal of increasing paddy production of Sri Lanka.

REFERENCES

Audi, R., Ed. (1996). "curve fitting problem," The Cambridge Dictionary of Philosophy. Cambridge, Cambridge University Press. pp.172–173.

Draper, N.R. and Smith, H. (1998).Applied Regression Analysis Wiley Series in Probability and Statistics

Fox, J. (1997). Applied Regression Analysis, Linear Models and Related Methods. Sage

Lindley, D.V. (1987). "Regression and correlation analysis," New Palgrave: A Dictionary of Economics, v. 4, pp. 120–23.

Julian C. Stanley, "II. Analysis of Variance," pp. 541–554.