Environment and Ecology 40 (3B) : 1428—1437, July—September 2022
ISSN 0970-0420

On Selection of Best Model from Spline Regression and ARIMA Models for Forecasting Rabi Food Grain Production of Odisha

Abhiram Dash, Debasis Bhattacharya, Digvijay Singh Dhakre, Madhuchhanda Kishan

Received 14 April 2022, Accepted 10 May 2022, Published on 11 Aujust 2022

ABSTRACT

The present paper makes a comparative study of spline regression and ARIMA modelfor forecasting purpose. Various suitable spline regression models and ARIMA models are applied to forecast the production of *rabi* food grains grown in Odisha. The data set for the year from 1970-71 to 2019-20 has been used in the study which has been divided into two parts named as training set data (for the period 1970-71 to 2015-16) used to build the modeland testing set data (for the period 2016-17 to 2019-20) which are being held up for cross-validation of the selected model. The spline regression and ARIMA models found appropriate have been tried on the observed data on production of *rabi* food grains for the training set data. Firstly, the suitable models from the two groups (i.e. spline regression and ARIMA) are selected by conducting model diagnostics test and comparing the model fit statistics. The selected model from each group is then cross validated by using the testing set data. After successful cross-validation, the selected best fit model among the two groups of models has been used for forecasting production of rabi food grains for the year from 2020-21 to 2024-25. The model selection process suggests the use of logarithmic spline, power spline and ARIMA(1,1,0) without constant models for the forecasting purpose. The result of cross-validation of these three selected models ensures the logarithmic spline model for forecasting production of *rabi* food grains of Odisha. The forecast values give a very good news since they show that the production of *rabi* food grains of Odisha will be increasing in the future years.

Keywords ARIMA model, Forecasting, Model diagnostics test, Model fit statistics, Spline regression.

INTRODUCTION

Forecasting future values in case of time series data can be done by using suitable ARIMA model. In most cases ARIMA modelis used for future forecasting. But the main drawback of ARIMA model is that it can give reliable forecast for a short future period. The reason is that the uncertainty increases in the movement of the series for the period for which the prediction has to be made is quite far in future time.

To get a forecast for a relatively longer period, a modification of regression techniques can be more

Abhiram Dash[1]*, Debasis Bhattacharya[2], Digvijay Singh Dhakre[3], Madhuchhanda Kishan[4]

[1]Assistant Professor, Department of Agril. Statistics, College of Agriculture, OUAT 751003, Bhubaneswar, India
[2]Professor, Department of Agril. Statistics Institute of Agriculture, Visva-Bharati, Santiniketan, India
[3]Assistant Professor, Department of Agril. Statistics Institute of Agriculture, Visva-Bharati, Santiniketan, India
[4]Assistant Professor, College of Agriculture (OUAT), Bhawanipatna, India
Email: abhiramouat@gmail.com
*Corresponding author

useful. Spline regression is such a model where different curves are fitted to different section of the dataset without losing the continuity of the curve.

The spline regression technique is applied in this study as the data set related to the variables considered in the study have been considered over a long period of time.The data set considered in this study subjected to vary for different time periods and abrupt jump (s) can be seen from one segment of time period to the other. Spline regression technique is used to capture these abrupt jumps in the value of the variables without losing the continuity of the model. Thus, two different groups of models are fitted and are compared among themselves with respect to important statistical criteria and the best one among them is used for forecasting purposes. Hence, the main objective of this study is to compare the effectiveness between ARIMA model and spline regression model as a forecasting model.

The list of important food grains produced in Odisha includes rice, maize, ragi and wheat, which come under cereals and red gram, green gram, black gram, and cowpea, which come under pulses. The state of Odisha ranks 12th position with respect to the production of food grains taken as average over the years 2012-13 to 2019-20, at all India level (Odisha Agricultural Statistics 2020).

The forecasting of production of food grains is important in formulating strategies to frame the agricultural planning of the state. Two different forecasting methods mentioned above using suitable models are compared to select the best forecasting model. The method yielding the best model which could provide the efficient forecast for a particular variable is used for obtaining the forecast value.

MATERIALS AND METHODS

The study relates to forecasting of production of food grains in the state of Odisha in *rabi* season for the year 2020-21 to 2024-25. The data on production of *rabi* food grains in Odisha are collected for the period 1970-71 to 2019-20 from Odisha Agricultural Statistics, 2020 published by the Directorate of Agriculture and Food Production, Odisha.

The models are fitted by following two approaches in broad sense, which are spline regression model approach and ARIMA model approach. The data set for the year from 1970-71 to 2015-16 is considered to be training set data used to build the model. The data for the years from 2016-17 to 2019-20 are kept for cross validation purpose of the selected model (s) under both the approaches and are thus considered to be testing set data. Under each approach, suitable forecasting models have been fitted to the training set data. The best model among the fitted models under each approach is selected for comparison. The study of scatter plot of the data on production of *rabi* food grains of Odisha helps us to get an idea about models that could possibly fit well to the data. Different models found to be suitable are linear, compound, logarithmic,power model and have been tried.

Spline regression approach involves in fitting the selected models over the training set data after splitting the entire period into different time segments based on the scatter plot of the data. Spline regression technique involves in joining two or more separate regression lines at a point known as spline knot (s) while their slopes are allowed to be different at that point. Thus, the data for the whole period (i.e., 1970-71 to 2019-20) is splitted with the help of scatter plot of the data and a suitable model is fitted in each occasion.

The best forecasting model is selected from both spline regression and ARIMA approach and are then cross validated by using the data for the period from 2016-17 to 2019-20. The model yielding the lowest MAPE during cross – validation is used for forecasting of the variable for the period from 2020-21 to 2024-26.

Prior to fitting of the models, the data are checked for presence of outliers. The Inter-Quartile Range of the data series, denoted as IQR is used for checking of outliers.

Hence, IQR = Q_3 - Q_1, where Q_1 and Q_3 are the first and third quartiles, respectively.

The observations which are less than $Q_1 - 3\times IQR$ or more than $Q_3 + 3\times IQR$ are referred to as extreme outliers (Bhattacharya and Roychowdhory 2010). The outlier observations found by following the above mentioned procedure are eliminated from the data set before analysis.

Abrief description of different spline regression models used in the study are given below. In all the models $X_t$is the value of the variable at time t, β_0 and β_1 are the parameters of the model used in the study and ε_t is the random error component operating with $X_t$at time t.

The spline models are fitted using spline regression technique with two knots placed at time period, k_1 and k_2 in the following manner:

Linear spline model:

$$X_t=\beta_0+\beta_1.\,t.\,I_{(1\le t\le k1)}+\{\beta_1.\,t+A_1\,(t-k_1)\}.\,I_{(k1+1\le t\le k2)}+\{\beta_1.\,t\ +A_1\,t+A_2\,(t-k_2)\}.I_{(k2+1\le t\le n)}+\varepsilon_t$$

Logarithmic spline model:

$$X_t=\ \beta_0+\beta_1.\,\ln(t)\,.\,I_{(1\le t\le k1)}+\{\ \beta_1\,.\,\ln(t)\ +A_1.\,\ln(t-k_1)\}\ .\ I_{(k1+1\le t\le k2)}+\{\ \beta_1\,.\,\ln(t)\ +A_1\,.\,\ln(t)\ +A_2.\,\ln(t-k_2)\}\ .\ I_{(k2+1\le t\le n)}+\varepsilon_t$$

Compound spline model:

$$X_t=\beta_0\,.\,\beta_1 t\,.I_{(1\le t\le k1)}.\ \{\beta_1 t.\,A_1\,(t-k_1)\}.\,I_{(k1+1\le t\le k2)}.\ \{\ \beta_1 t.\ A_1 t\,.A_2\,(t-k_2)\ \}.\ I_{(k2+1\le t\le n)}\exp(\varepsilon_t)$$

The compound spline model can be transformed to linear form by a natural log transformation and written as,

$$\ln(X_t)=\ \ln\,\beta_0+t\,.\,\ln(\beta_1).I_{(1\le t\le k1)}+\{t\,.\,\ln(\beta_1)+(t-k_1).\ \ln(A_1)\}I_{(k1+1\le t\le k2)}+\{t\,.\,\ln(\beta_1)+t\,.\,\ln(A_1)+(t-k_2).\,\ln(A_2)\}\ I_{(k2+1\le t\le n)}+\varepsilon_t$$

Power spline model:

$$X_t=\beta_0.t^{\,\beta 1}.I_{(1\le t\le k1)}\{t^{\,\beta 1}.(t-k_1)^{\,A1}\}.I_{(k1+1\le t\le k2)}.\{\ t^{\,\beta 1}\ .\ \ (t\text{-}k_2)^{\,A2}\}.I_{(k2+1\le t\le n)}\exp(\varepsilon_t)$$

The power spline model is transformed to linear form by natural log transformation as,

$$\ln(X_t)=\ \ln\,\beta_0+\beta_1.\ \ln(t).\ I_{(1\le t\le k\_1)}+\ \{\beta_1.\ln(t)+A_1\ \ln(t-k_1)\}.\ I_{(k1+1\le t\le k2)}+\ \{\beta_1.\ln(t)+A_1.\ln(t)+A_2\ \ln(t-k_2)\}.\ I_{(k2+1\le t\le n)}+\varepsilon_t,$$

where, $I_{(p)}$ is the indicator function which is 1, if P holds and 0, otherwise.

Ordinary Least Squares technique is used to estimate the parameters of the model considering a check for all the model assumptions to be satisfied by the selected model.

The model fit statistics, viz., R^2, adjusted R^2, Root Mean Square Error (RMSE),Mean absolute Percent Error (MAPE) andcorrected Akaike's Information Criteria(AICc) have been used as the model selection criteria.

The overall significance of the model is tested by using Snedecor's F test.

The significance of the estimated parametric coefficients are tested by using t-test.

The following statistical tests are considered for testing different assumptions made on errors in the model:

Durbin-Watson test for testing independence of residuals.

Shapiro-Wilk's test for testing normality of residuals.

Breusch-Pagan test for testing homoscedasticity of the errors

Durbin-Watson (D-W)test: This test considers the first order autocorrelation among the residuals. (Montgomery *et al.* 2001).

In D-W test the null hypothesis is H_0: The errors are independent and the alternative hypothesis is H_1: The errors are not independent.

The Durbin-Watson test statistic (D-W statistic) is defined as

$$d=\frac{\sum_{t=2}^{n}(e_t-e_{t-1})^2}{\sum_{t=1}^{n}e_t^{\,2}}$$

where, e_t and e_{t-1} are the errors obtained from the model for the time period t and (t-1) respectively

Independency of errors can be assumed if p-value of the test statistic d is greater than 0.05

Shapiro-Wilk's test: This test is used for testing normality of the residuals. The null hypothesis for this test is H_0: The errors follow normal distribution which is tested against the alternative hypothesis, H_1:

The errors do not follow normal distribution.

To carry out the test, the data pertaining to errors are arranged in ascending order so that $e_{(1)} \leq e_{(2)} \leq \ldots \leq e_{(n)}$, where $e_{(i)}$ is the ith order observations on errors.

The Shapiro-Wilk's (S-W) test statistic is given by

$$W = \frac{s^2}{b}$$

where, $s^2 = \sum_{k=1}^{m} a(k)\{e_{(n+1-k)} - e_{(k)}\}^2$,

$b = \sum_{t=1}^{n} (e_t - \bar{e})^2$ (Lee *et al*. 2014)

If n is even, then the value of m is $\frac{n}{2}$ and if n is odd, then $m = \frac{n-1}{2}$

n is the number of observations, $e_{(k)}$ is the k^{th} order statistic in the set of residuals,

e_t is the residual at time 't' and $\bar{e}$ is the mean of e_t. Values of the coefficients a(k) for different values of k and particular values of n are obtained from the table of Shapiro-Wilk.

For a given value of n, the value of p that is closest to 'W' can be obtained from Shapiro-Wilk's table. If the p value exceeds 0.05, then the null hypothesis cannot be rejected. If it lies below 0.05 but above 0.01, then the null hypothesis is rejected at 5% level. If the p value is below 0.01, then the null hypothesis is rejected at 1% level.

Breusch-Pagan test

The homoscedasticity of errors obtained from the regression model can be tested by using Breusch-Pagan test (Breusch and Pagan1979). Here the null hypothesis is taken as H_0: Errors have constant variance i.e., homoscedastic.

Alternative hypothesis, H_1: Errors have non-constant variance i.e., heteroscedastic

BP statistic follows chi-square distribution with 'k' degrees of freedom, where 'k' is the no. of parameters involved in the model

Breusch-Pagan test statistic is given by,$\chi^2 = n \times R^2$; Where, n is the no. of observations, R^2 is the coefficient of determination of the regression of squared residuals (obtained from the original regression) on the independent variable (which is time t, in the present study)

If the BP statistic has a p-value below 0.05, then the null hypothesis is rejected and heteroscedasticity is assumed to be present in the residuals and the regression model used can be considered to be inappropriate fit.

Among the fitted models having overall significance, significant parametric coefficients and satisfying the diagnostics tests, the one having highest R^2, highest adjusted R^2, lowest RMSE, MAPE andAICcis considered to be the best fit model for that dependent variable.

$R^2 = \frac{SSM}{SSE}$, where, SSM is the sum of square due to model, SSE is the sum of square due to error.

The expressions for SSM and SSE are, respectively, Adjusted R^2 is defined as Adjusted

$$R^2 = 1 - (1-R^2) \times \frac{(n-1)}{(n-p)}$$

where, p is the no. of coefficients involved in the model.

Adjusted R^2 penalizes the model for adding some independent variables which are not necessary to fit the data and thus adjusted R^2 will not necessarily increase with the increase in the number of independent variables included in the model.

$$RMSE = \left\{ \frac{\sum_{t=1}^{n} (yt - \hat{yt})^2}{(n-p)} \right\}^{1/2}$$

Mean Absolute Percent Error,MAPE $= (\sum_{i=1}^{n} \frac{pi - oi}{oi} \times 100)/n$,

where Pi and Oi are the predicted and observed values for the i^{th} year respectively, i= 1, 2,…, n.

Akaike's Information Criteria (AIC) estimates the relative amount of information lost by a given model. The less information a model loses, the higher the quality of that model.

$$AIC = \frac{1}{n}\left[\frac{RSS}{\sigma^2}\right] + 2k$$

$$\sigma^2 = \frac{SSE}{n - k}$$

where, SSE is the residual sum of squares and k is the number of parameters involved in the model, RSS is the regression sum of squares.

Corrected Akaike's Information Criteria (AICc) is a better criterion and should be used instead of AIC when sample size is small in comparison to the number of estimated parameters. Burnham and Anderson 2002 recommend the use of AICc instead of AIC when n/kis less than 40, where n is the no. of observations and k is the no. of parameters involved in the model. Since in the present study the models are fitted by using 45 no. of observations i.e., the year from 1970-71 to 2014-15 and the no. of parameters, k is at least 2, n/k is always less than 40.

So in the present study, AICc is used as model selection criteria instead of AIC.

AICc = AIC + 2k(k+1)/(n-k-1)
where n is the no. of observations and k is the no. of parameters involved in the model.
The difference between the AICc of a particular model and the best model with the lowest AICc can be used for relative assessment of the fitted model. This difference can be represented as ΔAICc .

As a thumb rule given by Raftery(1996)models having $\Delta\text{AICc} \leq 2$ are not substantially poor than the best model. Those models in which $4 \leq \Delta\text{AICc} \leq 7$ have considerably poor fit than the best fitted model, and models having $\Delta\text{AICc} > 10$ can be considered to be much poor fit than the best fitted model. But for a very high value of AICc, these comparison may seem trivial.

Fitting of ARIMA model

Prior to selection of suitable ARIMA model, the data must be made stationary. A check for stationarity of the given time series data is done by using augmented Dickey Fuller test. In this test, the null hypothesis is that a unit root is present in the data which is tested against the alternative hypothesis that there is no unit root in the data(Xiao and Philips 2014). If the original data is not stationary, then the first differences of the data are checked for stationarity. If first difference series is also found to be non-stationary, then the second difference series is checked and so on. Butusually the first difference series or maximum second difference series is found to be stationary.

By looking at the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots of the stationary data series, the orders of AR and MA terms that are needed to build the model can be tentatively identified. The lag beyond which the PACF cuts off is the indicated number of AR terms to be retained in the model. The lag beyond which the ACF cuts off is the indicated number of MA terms to be retained in the model. In the model the number of AR terms is denoted by 'p' and the number of MA terms is denoted by 'q'.

Let Y_t be the value of the time series at time t, where, t = 1,2,3,…,n

If the order of differencing, d=1, then yt = Yt – Yt-1.
If the order of differencing, d=2,
then yt= (Yt – Yt-1) – (Yt-1 – Yt-2)
= Yt – 2Yt-1 + Yt-2.
The forecasting equation of ARIMA model with 'p' number of AR terms 'q' number of MA terms and order of differencing 'd' is expressed as:

$$Y_t = \mu + \theta_1 Y_{t-1} + \theta_2 Y_{t-2} + \ldots + \theta_p Y_{t-p} - \Phi_1 \varepsilon_{t-1} - \Phi\ _2 \varepsilon_{t-2} - \ldots - \Phi_q \varepsilon_{t-q} ,$$

where, Yt, Yt-1, Yt-2,… are the stationarized values of time series for time points t, t-1, t-2,… which may be the original values of the series or the values obtained after first or second order differencing, μ is the constant term; θ1, θ2,… are the AR coefficients; Φ 1, Φ 2,…are the MA coefficients and εt-1,εt-2,…, εt-q are the error terms at lags 1, 2, …, q respectively.

After identifying the values of p (the order of AR terms) and q (the order of MA terms), the parameters of the autoregressive and moving average terms are estimated using simple least square techniques. Next, after determining the values of p, q and d the parameters associated with AR and MA terms are estimated. Later the constant and the coefficients of the AR and MA terms are tested for their significance. If the constant term appears not to be significantly different from zero, then ARIMA model without constant is fitted. After testing the significance of the model parameters, the diagnostic test for the residuals of the selected model is done.

The randomness of the residuals is tested by using Box-Pierce test (McElroy and Monsell 2014).

Here the null hypothesis is set as H0: The errors are distributed randomly and the alternative hypothesis H_1: The errors are non-random.

The Box-Pierce Q-statistic is given by:

$$BP(k) = n\sum_{k = 1}^{L} \rho^2 e,k, \text{ where:}$$

$\rho^2 e,k$ is the autocorrelation coefficient at lag k of the residuals e_t, where et = $Y_t - Y_t$

n is the number of terms in the differenced series;

k is the maximum lag being considered which is usually 2 in case of annual data

If the residuals are random, they will be distributed as Chi-Square with (k-m) degrees of freedom, where m is the number of parameters in the model which has been fitted to the data.

The normality of the residuals is tested by Shapiro-Wilk's test. After the diagnostic checking of the model and its parameters, the evaluation of the model is done. Among the models satisfying the tests for residual diagnostics, the best fit model is chosen using any one of the criteria like RMSE and MAPE. The model having lowest value of any of these measures is considered to be the best fit ARIMA model for the given data.

After exploring the best fit model from each group, cross validation is done for the selected models.The forecast values of the dependent variable obtained from the fitted model for the time period for which the observations were left out for the validation purpose are used for this purpose. From the actual values and the forecast values of the dependent variable for the time period left out for validation, the absolute percentage error (APE) value is obtained for each observation in the left outtime period. The APE for the ith year of validation period is obtained as,

$$APE_i = \frac{pi - oi}{oi} \times 100$$

where Pi and Oi are respectively the predicted and observed values for the ith year, i= 1, 2, …, 9. Low value of APE ensures the appropriateness of the selected model for forecasting. After successful cross validation of the selected model, it is used for the purpose of forecasting.

R software has been used for the regression analysis including Durbin Watson-test, Shapiro-Wilk's test and Breusch Pagan test. Also the R software has been used for Augmented Dickey-Fuller test, ACF and PACF plots and fitting of ARIMA model to the time series data on production of *rabi* food grains in Odisha.

RESULTS AND DISCUSSION

The scatter of data on production of *rabi* food grains in Odisha as shown in Fig. 1 shows that the production undergoes three different phases in the entire period

The functions in R used for the analysis along with their respective packages are:

Name of the function	Use of the function	Name of the package that contains the function
lm()	Fit regression model	base
dwtest()	Durbin –Watson test for the residuals	lmtest
shapiro.test()	Shapiro-Wilk's test to test the normality of the residuals	dplyr
bptest()	Breusch-Pagan test to test the homoscedasticity of the residuals	lmtest
adf.test()	Augmented Dickey-Fuller test to check the stationarity of the time series data	tseries
acf()	To obtain ACF plot of the time series data	forecast
pacf()	To obtain PACF plot of the time series data	forecast
Arima()	Fit the ARIMA model to the time series data	forecast
Box.test()	To test the randomness of the residuals	stats
AICc()	To find AICc of the fitted model	MuMIn

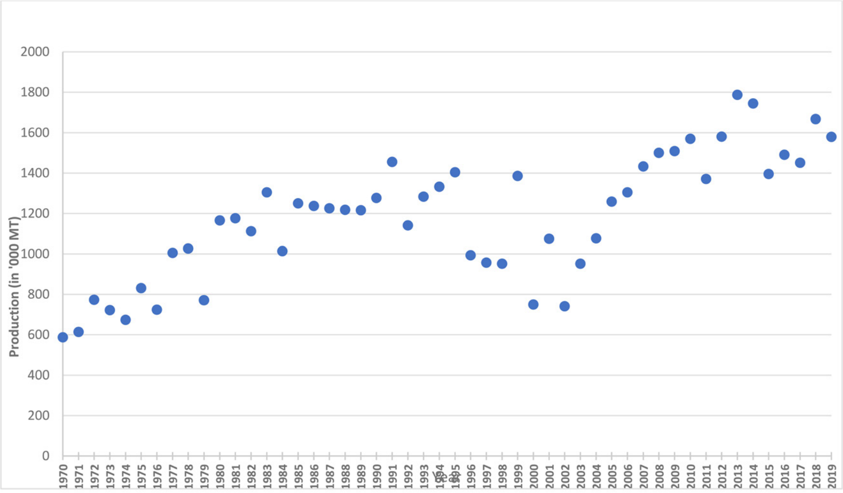

Fig. 1. Scatter of production of *rabi* food grains in Odisha from 1970-71 to 2019-20.

from 1970-71 to 2019-20 with knots at two places – first at the year 1985-86 and second one placed at the year 2002-03 which corresponds to the time, t = 16 and 33 respectively. Thus the entire period of study is divided into three sub-periods: Sub - period I (1970-71 to 1985-86), sub - period II (1986-87 to 2002-03) and sub - period III (2003-04 to 2019-20).

The study of Table 1 shows the results obtained by fitting of spline regression models of the type linear spline, logarithmic spline, compound spline and power spline model. All the fitted models show overall significance by having highly significant F-value with p-value less than 0.01and Also the parametric coefficients of all models have p-value less than 0.01 and are thus significant. The linear spline model does not satisfy the assumptions of normality and homoscedasticity of errors as the p-value of the

Table 1. Estimated model parameters model diagnostics and model selection criteria of spline regression models fitted to data on production of *rabi* food grains of Odisha. The figures inside the parentheses represent the p-value.

	Linear spline	Logarithmic spline	Compound spline	Power spline
b_0	703.613 (5.58e-12)	406.123 (0.000675)	695.734 (< 2e-16)	492.864 (< 2e-16)
b_1	26.836 (8.13e-06)	287.611 (1.31e-06)	1.028 (6.29e-06)	1.371 (4.47e-08)
a_1	-22.432 (0.00034)	-109.691 (0.003103)	0.979 (0.000571)	0.889 (0.000827)
a_2	78.579 (2.10e-05)	230.132 (3.51e-06)	1.065 (0.000474)	1.200 (4.15e-05)
D-W Statistic	1.5804 (0.05425)	1.661 (0.09883)	1.473 (0.021)	1.792 (0.231)
S-W Statistic	0.939 (0.019)	0.986 (0.8388)	0.933 (0.012)	0.977 (0.5174)
BP Statistics	8.8942 (0.031)	7.326 (0.0622)	5.953 (0.114)	7.5895 (0.0553)
R^2	0.674	0.710	0.618	0.714
Adj. R^2	0.649	0.693	0.590	0.689
RMSE	176.1223	165.973	190.059	161.678
MAPE	13.918	12.143	14.859	11.335
AICc	471.301	465.959	478.155	463.600
F	28.21 (4.641e-10)	33.48 (4.173e-11)	22.12 (1.114e-08)	34.051 (3.279e-11)

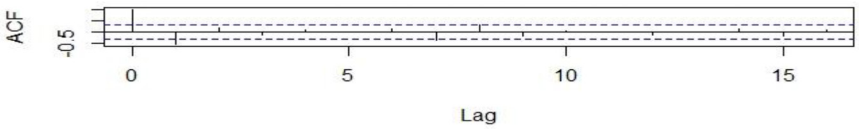

Fig. 2. ACF of the first order difference data on production of *rabi* food grains in Odisha.

S-W statistic and BP-statistic used respectively for testing the assumptions are found to be less than 0.05. The compound spline model also does not satisfy the assumptions of normality and independence of errors as the p-value of the respective test statistic i.e., S-W statistic and DW-statistic used for testing the assumptions are found to be less than 0.05. The logarithmic spline and power spline models are found to satisfy all the three assumptions of errors and also have moderately high value of R^2 and adjusted R^2 with low value of RMSE, MAPE and AICc as compared to linear spline and compound spline models. But these model fit statistics of both the models are very close to each other and the difference between AICc of both the models is less than 2. So, both are selected for cross-validation purpose.

To fit the ARIMA model for the data on production of *rabi* food grains in Odisha the Augmented Dickey-Fuller test has been conducted by using R software. The result of the test is as follows:
Null Hypothesis, H^0: Non-stationary data
Alternative nypothesis, H^1: Stationary data
For original data
Dickey-Fuller Statistic = -2.383, p-value = 0.41
For first difference data
Dickey-Fuller = -6.157, p-value = 0.01
Alternative hypothesis: Stationary data

The Dickey-Fuller statistic is significant for first difference data. Thus order of differentiation (d) considered for building the model is 1, i.e., d =1.

The order of moving average (q) and auto regression (p) are determined with help of ACF and PACF graphs of the first order difference data as shown in Figs 2, 3 respectively. The Fig. 2 which shows the ACF plot gives the order of moving average (q) which is considered to be zero i.e., q = 0. The Fig. 3 which shows the PACF plot gives the order of autoregression (p) which is considered to be zero i.e., p = 0.

The fitting of ARIMA (1,1,0) has beend done by using R software. In the first attempt, the constant is included in the model. But as seen from the Table 3 which provides the result of analysis for fitting of ARIMA(1,1,0) model to data on production of *rabi* food grains in Odisha, the constant is found to be non-significant having p-value much higher than 0.05, ARIMA(1,1,0) without constant has also been fitted. As seen from the Table 3, the coefficient of AR(1) is signifcant with p-value <0.01. Also the model satisfies the randomness and normality of errors as the Box-Pierce Statistic and S-W Statistic used for the respective purposes are both non-sognificant having p-value >0.05.

The study of Tables 1, 2 also reveals that the selected possible best fit models from spline category i.e., logarithmic spline and power spline model and from ARIMA category i.e., ARIMA(1,1,0) without constant model have very close values of RMSE, MAPE and AICc. Now the decision for declaring the best fit models among these three depend on the result of cross-validation for which the MAPE is used.

The cross-validation of the selected model from each of the three models shown in Table 3 shows

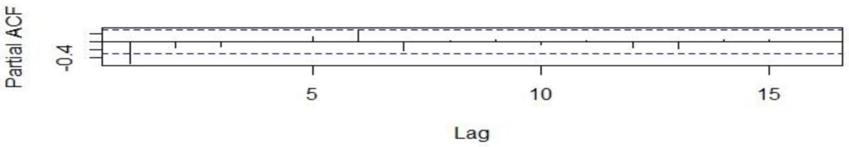

Fig. 3. PACF of the first order difference data on production of *rabi* food grains in Odisha.

Table 2. Estimated model parameters, model diagnostics and model selection criteria of selected ARIMA models fitted to data on production of *rabi* food grains of Odisha. Figures inside the parentheses indicate the p -value.

	ARIMA (1,1,0) (with constant)	ARIMA (1,1,0) (without constant)
Constant	26.835 (0.192)	-
b_1	-0.528 (0.001)	-0.503 (0.001)
Box-Pierce Statistic	0.309 (0.578)	0.507 (0.4762)
S-W Statistic	0.968 (0.253)	0.968 (0.245)
RMSE	173.617	178.209
MAPE	12.577	12.922
AICc	465.416	464.566

that the selected model (s) from each group have low values of MAPE. This shows that all are almost efficient for forecasting the area under *kharif* food grains in Odisha. Since among these three models, the logarithmic spline model provides the lowest MAPE (3.171 %) this model is selected to best model among the two groups of models fitted to the data on production of *rabi* food grains in Odisha. Thus, logarithmic Spline model is used for forecasting of production of rabi food grains in Odisha for the future years from 2020-21 to 2024-25.

The forecast values of production of *rabi* food grains in Odisha for the future years from 2020-21 to 2024-25 are presented in Table 4. The line graph of actual and estimated values of *rabi* food grain production in Odisha is also shown in Fig. 4. The fitted logarithmic spline regression curve is found to run at par with the actual curve except at few points. This shows that logarithmic spline curve fits well to the data on production of *rabi* food grains in Odisha which would ensure the reliability of the forecast values obtained for the future years.

Table 4. Forecast values of production of *rabi* food grains in Odisha for the year from 2020-21 to 2024-25 by using the selected best fit logarithmic spline model.

Year	2020-21	2021-22	2022-23	2023-24	2024-25
Production (in ‘000 tonnes)	1685.98	1696.28	1712.59	1732.74	1750.44

CONCLUSION

Thus, it is concluded from the study that among the spline regression models, both logarithmic spline model and power spline model fit well to the data as they fulfill all the selection criteria for a model to be a good fit. ARIMA (1,1,0) without constant model is found to be the best fit model among ARIMA models. The result of cross-validation of the three selected models, viz., logarithmic spline model, power spline model and ARIMA(1,1,0) without constant model by using the actual data for the period from 2016-17 to 2019-20 yield MAPE of 5.47 %, 7.96 % and 6.96 % respectively. It is found that ARIMA model is performing better than power spline model in terms of MAPE during cross validation but the logarithmic spline model is found to perform better than the selected ARIMA(1,1,0) without constant model in terms of MAPE. Thus, logarithmic spline model yielding the lowest MAPE is used for forecasting the production of *rabi* food grains in Odisha for the future years from 2020-21 to 2024-25. Apart from this, the use of spline regression model also enables us to get forecast values for longer time period which would not have been possible by use of ARIMA model as

Table 3. MAPE values for the selected best fit models for production of *rabi* food grains in Odisha among the possible best fit spline and ARIMA models.

		Predicted values			Absolute percentage error		
Year	Actual value	Logarithmic spline model	Power spline model	ARIMA (1,1,0 without constant model)	Logarithmic spline model	Power spline model	ARIMA (1,1,0 without constant model)
2016-17	1491	1554.32	1596.43	1600.76	4.25	7.07	7.36
2017-18	1451.9	1572.56	1623.45	1622.07	8.31	11.82	11.72
2018-19	1667.11	1689.79	1709.45	1643.37	1.36	2.54	1.42
2019-20	1580.26	1706.13	1744.52	1664.68	1.64	4.07	5.34
		Mean absolute percentage error			5.47	7.96	6.46

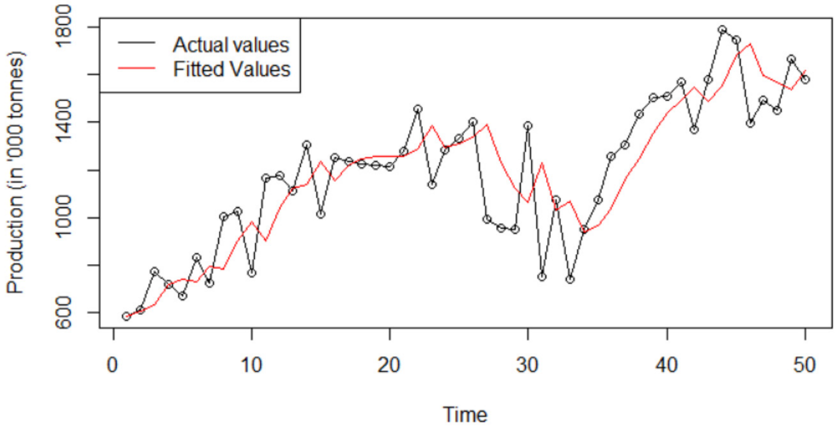

Fig. 4. Actual and predicted values of production of *rabi* food grains by using logarithmic spline model.

they are suitable for short term forecasting. Thus it is seen that for forecasting purposes, spline regression model can be better choice than ARIMA model.

It is seen from the forecast that the production of *rabi* food grains in Odisha is likely to increase in future years. The cause may be attributed to increase in yield by increasing the area under assured irrigation and HYV. The actual reason of increase in production could be inspected in further study regarding trend and variability in area and yield of *rabi* food grains in Odisha along with the study of effect of area under assured irrigation and HYV on production.

REFERENCES

Barton K (2009) Mu-MIn: Multi-model inference. R Package Version 0.12.2/r18. http://R-Forge.R-project.org/projects/mumin/

Bhattacharya D, Roychowdhory S (2010) Statistics: Theory and Practice, UN Dhur and Sons Private Limited, Kolkata,177-178.

Breusch TS, Pagan AR (1979) A simple test for heteroscedasticity and random coefficient variation. *Econometrica* 47(5): 1287-1294.

Burnham KP, Anderson DR (2002) Model selection and multi model inference: A practical information-theoretic approach. 2nd edn. New York, Springer-Verlag.

Hyndman RJ, Athanasopoulos G, Bergmeir C, Caceres G, Chhay L, O'Hara-Wild M, Petropoulos F, Razbash S, Wang E, Yasmeen F (2020) Forecast: Forecasting functions for time series and linear models. R package version 8.13. https://pkg.robjhyndman.com/forecast/>

Hyndman RJ, Khandakar Y (2008) Automatic time series forecasting: The forecast package for R. *J Stat Software* 27(3): 1-22. https:// www. Jstatsoft.org/article/view/v027i03

Lee R, Qian M, Shao Y (2014) On rotational robustness of Shapiro-Wilk type tests for multivariate normality. *Open J Stat* 4(11): 964-969.

McElroy T, Monsell B (2014) The multiple testing problem for Box-Pierce statistics. *Elect J Stat* 8: 497-522.

Montgomery DC, Peck EA, Vining GG (2001) Introduction to Linear Regression Analysis. 3rd edn New York, John Wiley and Sons, USA.

Raftery Adrian E (1996) "Bayesian model selection in social research (With Discussion)." Sociol Methodology 25: 111-195.

R Core Team (2020) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Trapletti A, Hornik K (2019) *t-series*: Time Series Analysis and Computational Finance. R package version 0.10-47.

Wickham H, François R, Henry L, Müller K (2020) Dplyr: A Grammar of Data Manipulation. R package version 1.0.2. https://CRAN.R-project.org/package=dplyr

Xiao Z, Philips P (2014): An ADF coefficient test for a unit root in ARMA models of unknown order with empirical applications to the US economy. *Economet J* 1: 27-43.

Zeileis A, Hothorn T (2002). "Diagnostic Checking in Regression Relationships." *R News* 2(3): 7–10. https://CRAN.R-project.org/doc/Rnews/