International Journal of Plant & Soil Science

33(23): 15-24, 2021; Article no.IJPSS.75651
ISSN: 2320-7035

Forecasting of Kharif Cereal Production in Odisha by Using Spline Regression Technique

Harsha S. Basanaik [a#] and Abhiram Dash [a*†]

[a] Department of Agricultural Statistics, College of Agriculture, OUAT, Bhubaneswar, India.

Authors' contributions

This work was carried out in collaboration between both authors. Both authors read and approved the final manuscript.

Article Information

DOI: 10.9734/IJPSS/2021/v33i2330714
Editor(s):
(1) Prof. Hakan Sevik, Kastamonu University, Turkey.
Reviewers:
(1) A. Mohamed Ashik, National College, India.
(2) Muhammad Azka, Institut Teknologi Kalimantan, Indonesia.
Complete Peer review History: https://www.sdiarticle4.com/review-history/75651

Original Research Article

Received 25 August 2021
Accepted 04 November 2021
Published 09 November 2021

ABSTRACT

Cereals are prime determinant of agricultural status of the state mainly during kharif season. Forecasting of the production of kharif cereals is of utmost importance to formulate the agricultural policy and strategy of the state. The ARIMA model can be reliably used to forecast for short future periods because uncertainty in prediction increases when done for longer future periods. The predictions obtained from the ordinary regression model are valid only when the relationship between the independent variables and the dependent variable does not change significantly in the future period which can be rarely assumed. It is expected that the spline regression will overcome the respective discrepancies in both ARIMA and ordinary regression techniques of forecasting with the assumption that the future period which needs forecasting follows the same pattern as the last partitioned period.

The entire period of data is split into different periods based on the scatter plot of the data The suitable regression models, such as, linear, compound, logarithmic and power model are fitted to the data on area and yield of kharif cereals by using the training set data. Selection of best fit model is done on the basis of overall significance of the model, model diagnostic test for error assumptions and model fit statistics. The selected best fit model is then cross validated with the testing set data. After successful cross validation of the selected best fit models, they are used for forecasting of the future values for their respective variables.

[#]P.G. Scholar
[†]Assistant Professor
**Corresponding author: E-mail: abhidash2stat@gmail.com;*

The models found to be best fit and thus selected for cross validation purpose are compound spline model for both area and yield of kharif cereals respectively. Forecasting of area, yield and hence production of kharif cereals for six years ahead i.e., for the year 2020-21 to 2025-26 by using the selected best fit model after successful cross validation. The forecast values for production of kharif cereals are found to decrease despite increase in forecast values of yield which is due to decrease in forecast value of area.

Keywords: Cross validation; forecast; knots; portioning; spline regression.

1. INTRODUCTION

Cereals are important food grain crops of Odisha which shares almost 89.58% of the total food grains production respectively. Forecasting of cereal production is of utmost importance for framing of food policies and ensuring nutritional security of the state mainly during kharif season.

ARIMA models are commonly used for forecasting purpose. According to Ashik and Kanan [1], the ARIMA model performs better than exponential smoothing model in forecasting stock prices. But the drawback of ARIMA model is that it can give reliable forecast for a short future period as the uncertainty increases as prediction is made for periods which are quite far in future times [2]. The ordinary regression model has the demerit that the predictions obtained from the regression model are valid only when the relationship between the independent variables and the dependent variable does not change significantly. Spline regression is thought to get rid of these demerits as the technique fits different curves for different section of data range without losing the continuity of the curve. Hence, spline regression model can be used to obtain forecast for comparatively longer future period.

Keeping in view the above perspectives, the study has been made with following objectives:

- To divide the whole period of study into different segments on the basis of pattern of variability in the data on area and yield of kharif cereals in Odisha.
- To fit suitable spline regression models to the training set data on area and yield of kharif cereals in Odisha.
- To select the best fit model on the basis of model diagnostic test and model selection criteria.
- To forecast the area, yield and hence production of kharif cereals in Odisha by using the selected best fit spline regression model after successful cross validation.

2. MATERIALS AND METHODS

The data on area, yield and production of kharif cereals are collected from Odisha Agricultural Statistics published by Directorate of Agriculture and Food Production, Odisha [3] for the period 1970-71 to 2019-20 out of which the data for the period 1970-71 to 2015-16 are used for building the model and remaining data are kept for cross validation.

The scatter plot gives an idea for partitioning of the whole study period into different periods such that the data within a period follows a definite pattern and abruptly changes in the consecutive periods.

The partitioning of data into periods can be further ascertained by calculating CV of each period and testing whether the difference in CV of consecutive periods are significant or not.

Let S_i be the observed standard deviation of the i^{th} period, $\bar{x}_i$ be the observed mean of the i^{th} period, and let $m_i = n_i - 1$. Where, $n_i \rightarrow$ no. of years in i^{th} period,

Let CV^* be the estimate of population CV,

$$CV^* = \frac{\sum_{i=1}^{k} m_i \frac{\sigma_i}{\bar{x}_i}}{M} \quad \text{and} \quad M = \sum_{i=1}^{k} m_i$$

Where, k→ no. of periods

Test statistics, $\chi^2 = (CV^*)^{-2} (0.5 + (CV^*)^2)^{-1} [\sum_{i=1}^{k} m_i (\frac{\sigma_i}{\bar{x}_i})^2 - M (CV^*)^2]$

The Σ^2 value distributed as a central Σ^2 variables with k-1 degrees of freedom, from which the p-value can be computed. The Σ^2 value measures how far each sample CV is from the estimate of the population CV^* [4].

Fitting of the selected models by using spline regression technique:

The regression models such as linear, compound, power and logarithmic are fitted with spline regression technique with one knot placed at time period k_1 in the following manner:

Linear spline model:

$$Y_t = \beta_0 + \beta_1 . t. I_{(1\le t\le k_1)} + \{\beta_1 . t + A_1 (t - k_1)\}. I_{(k_1+1\le t\le n)} + \varepsilon_t$$

Logarithmic spline model:

$$Y_t = \beta_0 + \beta_1 . \ln(t) . I_{(1\le t\le k_1)} + \{ \beta_1 . \ln(t) + A_1 . \ln(t - k_1)\} . I_{(k_1+1\le t\le n)} + \varepsilon_t$$

Compound spline model:

$$Y_t = \beta_0 . \beta_1^{\,t} . I_{(1\le t\le k_1)} . \{ \beta_1^{\,t} . A_1^{\,(t-k_1)} \}. I_{(k_1+1\le t\le n)} . \exp(\varepsilon_t)$$

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

$$\ln(Y_t) = \ln \beta_0 + t . \ln(\beta_1) . I_{(1\le t\le k_1)} + \{ t . \ln(\beta_1) + (t - k_1) . \ln(A_1)\} I_{(k_1+1\le t\le n)} + \varepsilon_t$$

Power spline model:

$$Y_t = \beta_0 . t^{\beta_1} . I_{(1\le t\le k_1)} \{ t^{\beta_1} . (t - k_1)^{A_1} \}. I_{(k_2+1\le t\le n)} . \exp(\varepsilon_t)$$

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

$$\ln(Y_t) = \ln \beta_0 + \beta_1 . \ln(t) . I_{(1\le t\le k_1)} + \{ \beta_1 . \ln(t) + A_1 \ln(t - k_1)\} . I_{(k_1+1\le t\le n)} + \varepsilon_t ,$$

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

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

Linear spline model:

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

Logarithmic spline model:

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

Compound spline model:

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

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

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

Power spline model:

$$Y_t = \beta_0 . t^{\beta_1} . I_{(1\le t\le k_1)} \{ t^{\beta_1} . (t - k_1)^{A_1} \}. I_{(k_1+1\le t\le k_2)} . \{ t^{\beta_1} . t^{A_1} (t - k_2)^{A_2} \}. I_{(k_2+1\le t\le n)} \exp(\varepsilon_t)$$

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

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

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

Selection of best fit model:

The model to be considered for selection should have overall significance and must satisfy the assumptions regarding the errors.

The model fit statistics, viz., R^2, adjusted R^2, Root Mean square Error (RMSE), Mean absolute Percent Error (MAPE) are computed for the purpose of model selection. In spline regression models, the significance of the coefficients is tested by using t-test. F test is used to test the overall significance of the model.

The above models are fitted under the assumptions that errors are independently distributed, follow normal distribution and have constant variance i.e. homoscedastic.

The following statistical tests are considered for testing the assumptions regarding errors in the model:

(i) Durbin-Watson test for testing independence of residuals.
(ii) Shapiro-Wilk's test for testing normality of residuals.
(iii) Breusch-Pagan test for testing homoscedasticity of the errors

Durbin-Watson test: This test considers the first order autocorrelation among the residuals [5].

Null hypothesis is taken as, H_0: the errors are independent.

And the alternative hypothesis as, H_1: the errors are not independent.

Durbin-Watson test statistic (D-W statistic), $d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$,

Where, $e_t = y_t - \hat{y}_t$, y_t and $\hat{y}_t$ are respectively the actual and estimated values of the response variable at time t and *n* is the no. of observations. As the p-value of the test statistic d is greater than 0.05, the independency of errors can be assumed.

Shapiro-Wilk's test: This test is used for testing normality of the residuals.

Null hypothesis here is H_0: the errors follow normal distribution.

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 $e_{(1)} \leq e_{(2)} \leq \ldots \leq e_{(n)}$

The Shapiro-Wilk's (S-W) test statistic as given by, $W = \frac{s^2}{b}$

Where, $s^2 = \sum_{k=1}^{m} a(k) \left\{e_{(n+1-k)} - e_{(k)}\right\}^2$; $b = \sum_{t=1}^{n} (e_t - \bar{e})^2$ [6]

If *n* is even, then $m = \frac{n}{2}$. If n is odd, then $m = \frac{n-1}{2}$

The parameter k takes the values 1, 2, …, m.

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.

The values of coefficients are (k) for different values of *k* and particular values of *n* are obtained from the table of Shapiro-Wilk [7].

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 [7].

Breusch-Pagan test:

The homoscedasticity of errors obtained from the regression model can be tested by using Breusch-Pagan test [8].

Null hypothesis, H_0: Errors have constant variance i.e. homoscedastic.

Alternative hypothesis, H_0: Errors have non-constant variance i.e. heteroscedastic.

Breusch-Pagan test statistic is given as, BP = n x 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, in the present study).

BP statistic follows chi-square distribution with 'k' degrees of freedom.

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, model having overall significance and model satisfying the diagnostics tests, model having highest R^2, highest adjusted R^2 and lowest MAPE is 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,

$$SSM = \sum_{t=1}^{n} \left(\hat{y}_t - \bar{y}\right)^2 , \quad SSE = \sum_{t=1}^{n} \left(y_t - \hat{y}_t\right)^2 ,$$

Where y_t and $\hat{y}_t$ are respectively the actual and estimated values of the response variable at time t, and $\bar{y}$ is the mean of y_t.

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. To check the significance of R^2 and adjusted R^2, F-value is calculated. If the f-statistic has a p-value below 0.05, the test is significant.

$$\text{F-statistic} = \frac{R^2/(p-1)}{(1-R^2)/(n-p)} ,$$

Where, 'p' is the no. of coefficients involved in the model and 'n' is the no. of observations.

To check the significance of adjusted R^2, in use adjusted R^2 values in place of the R^2 values in the above mentioned f-statistic formula.

Mean Absolute Percent Error, $MAPE = (\sum_{i=1}^{n} \frac{|P_i - O_i|}{O_i} \times 100)/n$, where P_i and O_i are respectively the predicted and observed values for the i^{th} year, i= 1, 2,…, n.

After exploring the best fit model from each group, cross validation is done for the selected models by obtaining the forecast values for the time period 2016-17 to 2019-20 as the observations were left out for the validation purpose. From the actual and forecast values of the dependent variable, the absolute percentage error (APE) value is obtained for each observation in the left out period. The APE for the i^{th} year of validation period is obtained as, $APE_i = \frac{|P_i - O_i|}{O_i} \times 100$, where P_i and O_i are respectively the predicted and observed values for the i^{th} 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.

The R software has been used for fitting the models, obtaining model diagnostic criteria, model fit statistics.

3. RESULTS AND DISCUSSION

Figs. 1 and 2, show respectively the scatter plot of area and yield of kharif cereals for the period 1970-71 to 2019-20. The study of these scatter plots suggests the suitable regression models that would fit the data. The regression models found to be suitable are linear, compound, logarithmic and power model.

The scatter plot gives an idea for partitioning of the whole study period into different periods such that the data within a period follows a definite pattern and abruptly changes in the consecutive periods.

The scatter plot of area under kharif cereals as shown in the Fig. 1 shows that the area undergoes three different phases in the entire period from 1970-71 to 2019-20 with first knot at the year 1987-88 and second knot at the year 2001-02 which corresponds to the time, t = 18 and 32 respectively. Thus the entire period of study is divided into three sub-periods: sub-period I (1970-71 to 1987-88), sub-period II (1988-89 to 2001-02) and sub-period III (2002-03 to 2019-20).

The scatter plot of yield of kharif cereals as shown in the Fig. 2 shows that the yield undergoes two different phases in the entire period from 1970-71 to 2019-20 with first knot at the year 2001-02, which corresponds to the time, t = 32. Thus the entire period of study is divided into two sub-periods: sub-period I (1970-71 to 2001-02), sub-period II (2002-03 to 2019-20).

The partitioning of data into segments can be further ascertained by calculating Σ^2 value for CV of each segment and testing whether the difference in CV of consecutive segments are significant or not.

Table 1 shows the partitioning of data on area and yield of kharif cereals based on the testing of Coefficient of Variation. The difference in CV of consecutive periods is found significant for area and yield of both kharif cereals with p-value less than 0.05. The area of kharif cereals is divided into three periods, whereas yield is divided into two periods.

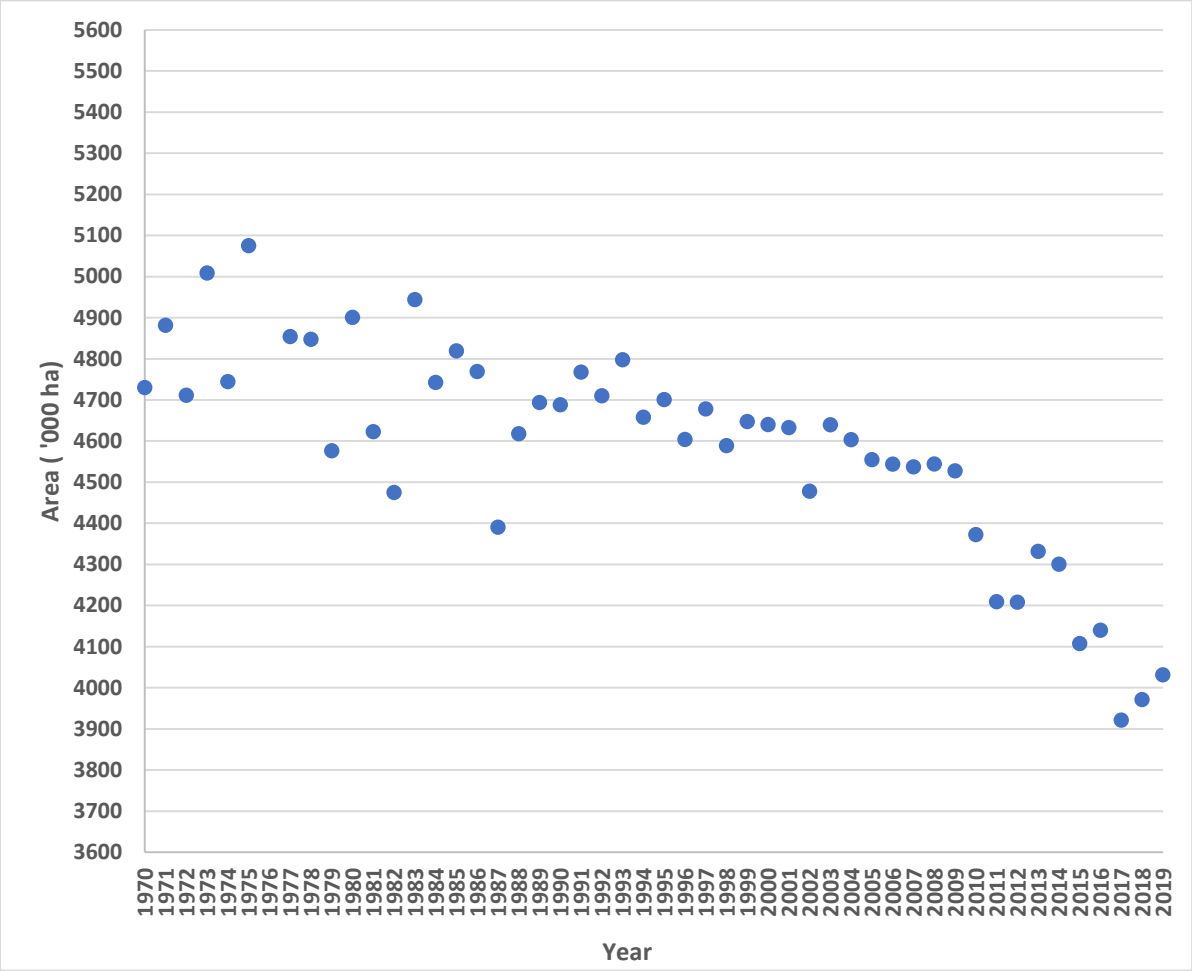

Fig. 1. Scatter plot of area under kharif cereals

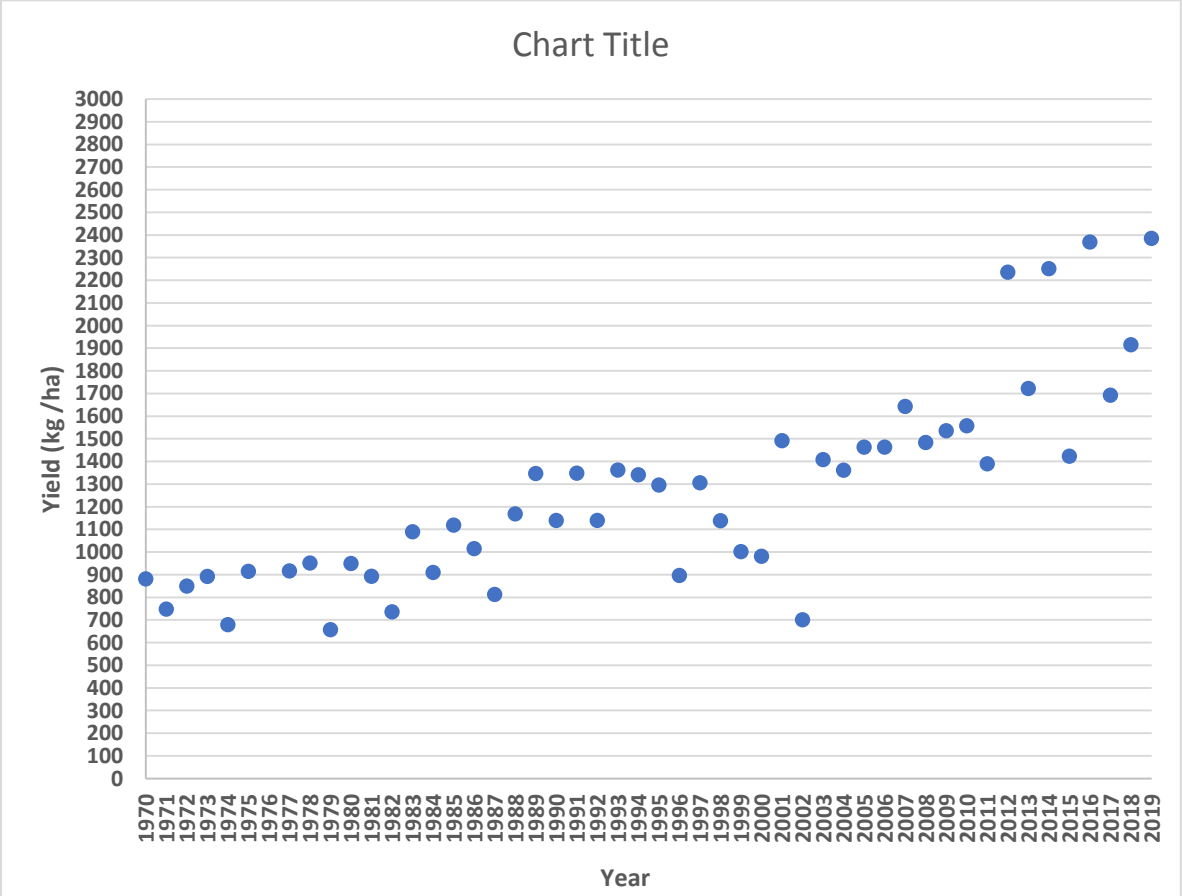

Fig. 2. Scatter plot of yield of kharif cereals

Table 1. Partitioning of data on area and yield of kharif cereals based on the testing of Coefficient of Variation

Season	Variable	Periods	Standard deviation	Mean	C.V. (%)	D'AD	
						I – II	II – III
Kharif	Area	I	261.84	4772.45	5.49	19.52***	19.098***
		II	59.38	4673.00	1.27	(≤0.001)	(≤0.001)
		III	232.45	4334.37	5.36		
	Yield	I	226.88	1020.9	22.22	0.431*	
		II	428.07	1667.19	25.68	(0.041)	

The figures in the parentheses represents the p-value
**** p-value ≤ 0.001; ** 0.001 < p-value ≤ 0.01; * 0.01 < p-value ≤ 0.05*

The study of Table 2 shows the results obtained by fitting of selected spline regression models of the type linear spline, logarithmic spline, compound spline and power spline model. The linear spline model, logarithmic spline model and power spline model does not satisfy the assumption of normality of errors as the SW-statistic used for testing the assumption of normality is found to be significant. The compound spline model is found to satisfy all the three assumptions of errors and also have moderately high value of R^2 and adjusted R^2 which are found to be significant and low value of RMSE, MAPE as compared to linear spline, logarithmic spline and power spline model. So, the compound model is selected for the purpose of cross-validation.

The study of Table 3 shows the results obtained by fitting of selected spline regression models of the type linear spline, logarithmic spline, compound spline and power spline model. The linear spline model, compound spline model and logarithmic spline model does not satisfy the

assumption of normality of errors as the SW-statistic used for testing the assumption of normality is found to be significant. The power spline model is found to satisfy all the three assumptions of errors and also have moderately high value of R^2 and adjusted R^2 which are found to be significant and low value of RMSE, MAPE as compared to linear spline, logarithmic spline and compound spline model. So, the power model is selected for the purpose of cross-validation.

Cross validation of the selected models for area and yield of kharif cereals in Odisha for the year from 2016-17 to 2019-20 is shown in Table 4. The absolute percentage error for the selected compound model for area under kharif cereals is found to be below 15% for all the years included in the testing data and the value of MAPE obtained is 11.50%. The absolute percentage error for the selected power model for yield of kharif cereals is found to be below 15% for all the years included in the testing data and the value of MAPE obtained is 14.88%. These values of MAPE obtained in both the cases are sufficiently low to accept compound spline model for area and power spline model for yield of kharif cereals as the best fit model. So, these models can be used respectively for forecasting of area and yield of kharif cereals in Odisha for the future years from 2020-21 to 2025-26. The forecast values of area and yield can be used for forecasting the production of kharif cereals of Odisha for the future years from 2020-21 to 2025-26.

Table 2. Estimated parametric coefficients, model diagnostics measures and model fit statistics of spline regression models fitted to training data on area under kharif cereals in Odisha

Model →	**Linear Spline**	**Compound Spline**	**Logarithmic Spline**	**Power Spline**
Estimated Parametric Coefficients				
b_0	4451.08***	3901.046***	4562.81***	4084.685***
	(≤0.001)	(≤0.001)	(≤0.001)	(≤0.001)
b_1	9.33	1.007	-5.05	1.019
	(0.591)	(0.388)	(0.978)	(0.845)
a_{11}	-2.813	-0.002	52.87	0.029
	(0.879)	(0.785)	(0.692)	(0.673)
a_{12}	-38.428	-0.011	-171.45	-0.053
	(0.454)	(0.671)	(0.307)	(0.544)
Model Diagnostics Criteria				
DW Statistic	2.301	2.204	2.285	2.18
	(0.572)	(0.818)	(0.617)	(0.886)
SW Statistic	0.364***	0.980	0.353***	0.936*
	(≤0.001)	(0.607)	(≤0.001)	(0.014)
BP Statistic	2.481	2.428	1.588	1.557
	(0.478)	(0.488)	(0.662)	(0.67)
Model Fit Statistics				
F Value	0.554	0.879*	0.424	0.214
	(0.648)	(0.038)	(0.739)	(0.886)
R^2	0.038	0.462***	0.029	0.015
	(0.649)	(≤0.001)	(0.741)	(0.887)
Adjusted R^2	-0.03	0.421***	-0.04	-0.055
	(0.73)	(≤0.001)	(0.63)	(0.493)
MAPE	21.17	21.159	21.364	22.532
AIC	600.422	600.008	600.83	605.42
AICc	601.398	600.984	601.81	606.396

**** p-value ≤ 0.001; ** 0.001 < p-value ≤ 0.01; * 0.01 < p-value ≤ 0.05*

Table 3. Estimated parametric coefficients, model diagnostics measures and model fit statistics of spline regression models fitted to training data on yield of kharif cereals in Odisha

Model →	Linear Spline	Compound Spline	Logarithmic Spline	Power Spline
Estimated Parametric Coefficients				
b_0	1291.7*** (≤0.001)	699.244*** (≤0.001)	1248.15* (0.016)	855.769*** (≤0.001)
b_1	-4.96 (0.75)	1.007 (0.241)	-20.58 (0.911)	1.085 (0.249)
a_{11}	60.87 (0.22)	0.026 (0.183)	229.05 (0.192)	0.16* (0.021)
Model Diagnostics Criteria				
DW Statistic	2.132 (0.892)	2.124 (0.915)	2.123 (0.916)	2.144 (0.858)
SW Statistic	0.395*** (≤0.001)	0.823*** (≤0.001)	0.367*** (≤0.001)	0.817 (0.06)
BP Statistic	2.158 (0.34)	1.926 (0.381)	1.301 (0.521)	1.204 (0.547)
Model Fit Statistics				
F Value	1.184 (0.316)	6.759** (0.003)	1.159 (0.323)	7.39** (0.002)
R^2	0.052 (0.317)	0.24** (0.003)	0.051 (0.324)	0.256** (0.002)
Adjusted R^2	0.008 (0.841)	0.204** (0.007)	0.007 (0.859)	0.221** (0.004)
MAPE	26.35	17.373	24.561	16.838
AIC	628.78	630.262	628.834	629.97
AICc	629.35	630.833	629.405	630.54

**** p-value ≤ 0.001; ** 0.001 < p-value ≤ 0.01; * 0.01 < p-value ≤ 0.05*

Table 4. Cross validation of the selected compound spline model and power spline model for area and yield of kharif cereals respectively in Odisha

Year	Area			Yield		
	Actual value	Predicted Value	APE	Actual value	Predicted Value	APE
2016-17	4139.76	4510.47	8.95	2369.4	2177.15	8.11
2017-18	3921.13	4487.51	14.44	1693	2188.28	29.25
2018-19	3971.15	4464.66	12.43	1915.85	2199.25	14.79
2019-20	4031.29	4441.93	10.18	2385.58	2210.04	7.36
Mean Absolute Percentage Error			11.50			14.88

Table 5. Forecast values of area and yield and hence production of kharif and rabi cereals in Odisha for the year from 2020-21 to 2025-26

Year	Kharif		
	Area ('000 ha)	Yield (kg/ha)	Production ('000 MT)
2020-21	4419.32	2220.67	9813.85
2021-22	4396.82	2231.15	9809.96
2022-23	4374.44	2241.48	9805.22
2023-24	4352.16	2251.65	9799.54
2024-25	4330.01	2261.69	9793.14
2025-26	4307.96	2271.59	9785.92

The forecast values of area and yield and hence production for both kharif and rabi cereals in Odisha for the future years from 2020-21 to 2025-26 are presented in Table 5. The forecast values

shows that the future values of area under kharif cereals is expected to decrease The future values for yield of kharif cereals is expected to increase. The future forecast values of production for kharif cereals is found to decrease which might be due to the decrease in forecast values for area. The forecast values of area and yield of kharif cereals obtained by fitting ARIMA model also shows the similar result but for a shorter period from 2016-17 to 2018-19 [9].

4. SUMMARY AND CONCLUSION

The selected models are used for forecasting of area and yield of kharif cereals in Odisha. The forecast values of area and yield can be used for forecasting the production of kharif cereals in Odisha for the future years from 2020-21 to 2025-26. The spline regression model thus helped us in providing relaible forecast values for production of kharif cereals for relatively longer future period than could be possible using ARIMA model.

The forecasted future value under area under kharif cereals is found to be decreasing which might be due to the shifting of cultivation area from cereals to some other foodgrains or non foodgrain crops deuring kharif season. The forecasted future values of yield of kharif cereals is found to be increasing may be because of adoption of new technologies and high yielding varieties. Despite expected increase in future forecast values of yield the forecast values of production for kharif cereals is found to be decreasing which might be due to the decrease in forecast values for area.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Ashik A, Kannan K. Time series model for national stock price prediction. Research & Reviews: Journal of Statistics. 2018;7(1): 85s–90sp
2. Sarika J, Iquebal MA, Chattopadhyay C. Modelling and forecasting of pegionpea (*Cajanus cajan*) production using Autoregressive integrated Moving Average methodology, Indian Journal of Agricultural Sciences. 2011;81(6):520–523.
3. Odisha Agricultural Statistics 2020, Directorate of Agriculture and Food Production, Odisha.
4. Feltz, Miller. An Asymptotic test for the equality of coefficients of variation from k populations, Statistics in Medicine. 1996; 15(6):647-658.
5. Montgomery DC, Peck EA, Vining GG. Introduction to linear Regression Analysis, 3rd Edition, New York, John Wiley & Sons, USA; 2001.
6. Lee R, Qian M, Shao Y. On Rotation Robustness of Shapiro-Wilk type. Tests for Multivariate Normality, Open Journal of Statistics. 2014;4(11):964-969.
7. Hanusz Z, Tarasinska J. Tables for Shapiro-Wilk W statistic according to royston approximation, Colloquium Biometricum. 2011;41:211-219.
8. Zaman A. The Inconsistency of the Breusch-Pagan Test, Journal of Economic and Social Research. 2000;2(1):1-11.
9. Dash, Abhiram, Mangaraju A, Mishra P, Nayak H. Using Autoregressive Integrated Moving Average (ARIMA) Technique to Forecast the Production of Kharif Cereals in Odisha (India), Current Journal of Applied Science and Technology. 2020; 39(9):104-113.

© 2021 Basanaik and Dash; This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Peer-review history:
The peer review history for this paper can be accessed here:
https://www.sdiarticle4.com/review-history/75651