Environment and Ecology 40 (3C) : 1612—1623, July—September 2022
ISSN 0970-0420

A Study on Trend and Forecast of Vegetable Cultivation of West Bengal

Bhola Nath, Debasis Bhattacharya

Received 13 May 2022, Accepted 17 June 2022, Published 26 August 2022

ABSTRACT

The present study was conducted to estimate the growth rate with the apparent fluctuations present in yield as well as area and production of vegetables grown in the state of West Bengal. The growth rates and instabilities in area, production and yield have been obtained by fitting an appropriate spline regression model. Linear, compound and linear spline models were found to be the best fit models to fit the data on yield, production and area, respectively. A significant positive growth rate in area, production and yield have been observed and the highest growth rate (2.74%) with the highest coefficient of variation (CV) (1.891%) was observed for production of vegetables. Growth rate of yield was 1.40% with CV of 1.889% followed by growth rate of area i.e., 1.34% with CV of 0.813%. The forecast values of area, production and yield of vegetables for subsequent five years have also been obtained by using appropriate ARIMA models. In case of area, production and yield of vegetables, ARIMA (1,1,0) model with constant was found to be the best fit model. The forecast values of all the three variables show a steadily increasing trend.

Keywords ARIMA, Coefficient of variation, Forecast, Growth rate, Instability.

INTRODUCTION

Peoples' health and sustainable development of the community can be ensured by ensuring the food safety. In India, majority of foods are prepared at household level. Vegetables are the most important part of our daily dietary plan; the state of West Bengal is accounted for 15.9% of the total vegetable production in India during the season 2018-19. The production of vegetables is followed by the state like, Uttar Pradesh (14.9%), Madhya Pradesh (9.6%), Bihar (9%) and Gujarat (6.8%). It can be noted that, the area under vegetable cultivation is increased by 3.36 million hectares from 2004-05 (6.74 million hectare) to 2018-19 (10.10 million hectare). The increased area under vegetables cultivation resulted into increment in production from 101.25 mt during 2004-05 to 185.88 mt during 2018-19 with an average yield of 18.4 tonnes per hectare.

Around 11.20% of the total vegetable production is contributed by the major vegetables viz., onion, brinjal, potato, tomato, cauliflower, peas, cabbage,

Bhola Nath[1]*, Debasis Bhattacharya[2]
[1]Research Scholar, [2]Professor and Head of the Department,
Department of Agricultural Statistics, Institute of Agriculture, Visva-Bharati, West Bengal 731236, India
Email : bnagstat@gmail.com
*Corresponding author

okra and production of the same is increasing day by day. West Bengal is one of the major potato cultivators' states in India, it produces around 110-115 lakh tonnes of tubers by using the area of 3.7 lakh hectares during the year 2019. It includes the major districts of Hooghly, Burdwan, Bankura, Midnapore. The main cause of decline in production of potato is delayed sowing that faces the altered weather conditions while tuber growth and maturity. Thus, potato production of Uttar Pradesh and West Bengal during 2020 is dropped by 20-25 lakh tonnes as compared to 2020, however the area under potato is increased by 10-12% during 2020.

The fluctuations in area, production and yield should be well studied in order to achieve sustainable development of the community. A well-known way of studying these fluctuations is by using regression approach. Usual regression approach works well for the data set which follows the assumptions of independence, homoscedasticity and normality, in a single line. Sometimes, the general regression cannot capture the actual behavior of the time series data (Nath *et al.* 2019). In that case, we use a piece wise regression by dividing the whole period of time according to the point of fluctuation which is generally termed as spline regression. Piece-wise polynomial for divided periods along with a continuous polynomial for the period as a whole is generally preferred (Dash *et al.* 2017(a)). These polynomials can be constructed by using dummy variable techniques. The pattern in the said variables is of interest while studying the growth rate and instability. Parametric and nonparametric approaches for fitting the data on yield along with area and production can also be used Dasyam *et al.* (2016). Nath *et al.* (2020) used parametric and nonparametric models according to the error distribution of the data set under study. Here, the growth rate and instability are studied for all the variables and forecast values for five years starting from 2015-16 to 2019-20 have also been obtained. Instability is measured by using CV. Instead of CV. Coppock's instability index can also be used to study the instability (Dash and Hansdah 2020).

Research methodology

Time series data related to yield as well as area and production of fruits and vegetables are used in the present study. Average growth rate and instability in area, production and yield of vegetables are obtained by fitting an appropriate spline model. Forecast values for subsequent five years of the same are also obtained by fitting an appropriate ARIMA model.

Spline regression models

Using dummy variables technique, the spline model with k = 9 can be constructed (Dash *et al.* 2017(a)). Dependent variable (area, production and yield) is denoted by Y_{t} and independent variable is time denoted by t. Using this notation, we can write the linear spline regression model with one knot as,

$$Y_{\mathrm{t}} = \beta_0 + \beta_1 (D_1 \times t + D_2 \times k) + \beta_1' (D_2 \times \mathrm{t} - D_2 \times k) + \varepsilon_{\mathrm{t}} \qquad (1)$$

For period I ($1 \leq t \leq 9$), $D_1 = 1$ and $D_2 = 0$; For period II ($10 \leq t \leq 18$), $D_1 = 0$ and $D_2 = 1$

The linear model for period I is given in (2). $Y_{\mathrm{t}} = \beta_0 + \beta_1 \times \mathrm{t} + \varepsilon_{\mathrm{t}}$, where, t = 1, 2, …, 9 (2)

For period II, the linear spline model is given in (3) $Y_{\mathrm{t}} = \beta_0 + \beta_1 \times k + \beta_1' (t - k) + \varepsilon_{\mathrm{t}}$, where, t = 10, 11, …, 18 and $k = 9$ (3)

To get the intercept of the model given in (3), the quantities β_0 and $\beta_1 \times \mathrm{k}$ is to be combined as, $\beta_0' = \beta_0 + \beta_1 \times \mathrm{k}$. The intercept (i.e., β_0) of the model (2) is altered by an amount of $\beta_1 \times \mathrm{k}$ in the model given in (3) which covers the period II and denoted by β_0'.

For the sake of simplicity, let us assume that the slope coefficient in model (2) is changed by an amount of A_1 and becomes $\beta_1' = \beta_1 + A_1$. It indicates that the slope of model (3) i.e., β_1' is the consequence of the change in the slope of the model (2) i.e., β_1. Alternatively, the model given in (3) can be written as,

$$\begin{aligned} Y_t &= \beta_0 + \beta_1 \times k + (\beta_1 + A_1)(t - k) + \varepsilon_{\mathrm{t}} \\ &= \beta_0 + \beta_1 \times k + \beta_1 \times t + A_1 \times t - \beta_1 \times k - A_1 \times k + \varepsilon_t \\ &= \beta_0 + \beta_1 \times t + A_1 (t - k) + \varepsilon_t, \text{ by using } \beta_1' = \beta_1 + A_1 \end{aligned}$$

This is the linear spline model for period II covering the period of 9 years starting from 2006-07 to 2014-15. Here, t takes the values 10, 11, …, 18 and k = 9. A continuous linear spline regression model for the period as a whole (i.e., t = 1, 2, …, 18) is expressed in (4).

$$Y_t = \beta_0 + \beta_1 \times t \times \mathrm{I}_{(1 \leq t \leq 9)} + \{\beta_1 \times t + A_1 (t - k)\} \times \mathrm{I}_{(10 \leq \mathrm{t} \leq 18)} + \varepsilon_t, \qquad (4)$$

where $I_{(m)}$ is the indicator function which is 1 if *m* holds and 0 otherwise. In the same way, four spline models viz., (i) power spline model, (ii) compound spline model, (iii) logarithmic spline model and (iv) quadratic spline models are obtained. These models are given below.

Power spline model

The model expressed in (5) is the power spline regression model.

$$Y_t = \beta_0 \times X_t{}^{\beta}{}_{\mathrm{t}} \times \mathrm{I}_{(\mathrm{I} \leq \mathrm{t} \leq 9)} \{{}^{\mathrm{t}\,\beta\mathrm{t}} \times (t\text{-}k)^{A1}\} \times \mathrm{I}_{(10 \leq + 18)} \times \exp(\varepsilon_{\mathrm{t}}) \tag{5}$$

This model can be transformed into linear form by using natural log transformation as:

$$\mathrm{In}\,(Y_t) = \mathrm{In}\,(\beta_0) + \beta_1 \times \mathrm{In}\,(t) \times I_{(1 \leq + \leq 9)} + \{\beta_{\mathrm{t}} \times \mathrm{In}\,(t) + A_1 \times \mathrm{In}\,(t\text{-}k)\} \times I_{(10 \leq \mathrm{t} \leq 18)} + \varepsilon_{\mathrm{t}}$$

Compound spline model

Compound spline regression can be expressed as,

$$\mathrm{Yt} = \beta_0 \times \beta_1 \times I_{(1 \leq t \leq 9)} \{\beta_1{}^{\mathrm{t}} A1^{(t\text{-}k)}\} \times I_{(10 \leq t \leq 18)} \times \exp(\varepsilon_{\mathrm{t}}) \tag{6}$$

The equation given in (6) can be transformed into linear form using natural log transformation as,

$$\mathrm{In}\,(Y_t) = \mathrm{In}\,(\beta_0) + t \times \mathrm{In}\,(\beta_1) \times I_{(1 \leq t \leq 9)} + \{\mathrm{t} \times \mathrm{In}\,(\beta_1) + (t\text{-}k) \times \mathrm{In}\,(A_1)\} \times I_{(10 \leq t \leq 18)} + \varepsilon_{\mathrm{t}}$$

Logarithmic spline model

The model given in (7) is known as logarithmic spline regression model.

$$Y_t = \beta_0 + \beta_1 \times \mathrm{In}\,(t) \times I_{(1 \leq t \leq 9)} + \{\beta_1 \times \mathrm{In}\,(t) + A_t\, \mathrm{In}\,(t\text{-}k)\} \times I_{(10 \leq t \leq 18)} + \varepsilon_t \tag{7}$$

Quadratic spline model

Quadratic spline regression model can be written as,

$$Y_t = \beta_0 + \{\beta_1\, t + \beta_2\, t^2\} \times I_{(1 \leq t \leq 9)} + \{\beta_1\, t + A_1 (t\text{-}k) + \beta_2\, t^2 + A_2 (t\text{-}k)^2\} \times I_{(10 \leq t \leq 18)} + \varepsilon_t$$

In all the four spine regression models discussed above, $\mathrm{I}_{(m)}$ is the indicator function as defined in linear spline regression model.

These models have been tried to fit the data related to yield as well as area and production of vegetables. The best fit models were selected by considering the model selection criteria. Growth and instability in yield as well as area and production of the same were obtained. Ordinary Least Squares (OLS) technique has been used in estimation of model parameters (i.e., β_0, β_1, A_1, β_2 and A_2) which are denoted by b_0, b_1, a_1, b_2 and a_2, respectively.

It is important to assume that residuals in the said models are independent, homoscedastic and normally distributed. The following tests have been used for testing the above assumptions:

(i) To test the independence - Durbin-Watson test (Montgomery *et al.* 2012)
(ii) To test the heteroscedasticity - Park's test (Gujarati and Porter 2012)
(iii) To test the normality - Shapiro-Wilk's test (Nath *et al.* 2020)

Model selection criteria for spline regression models

A specific model can be selected, if the model is overall significant and holds the assumptions imposed on the error term. The model fit statistics, like, R^2, adjusted R^2 and root mean square error (RMSE) are to be compared. The model with the highest values of R^2, adjusted R^2 and the lowest value of RMSE should be preferred.

Average growth rate

Growth rates can be obtained for three periods viz., period I (1997-98 to 2005-06), period II (2006-07 to 2014-15) and period as a whole (1997-98 to 2014-15). Annual growth rate can be obtained by using the following formula,

Annual Growth Rate for the year *t*, $\mathrm{GR}_{\mathrm{t}} = \left(\frac{Y_{t+1} - Y_t}{Y_t}\right) \times 100$, t=1, 2,........,18 and also, Max (t +1) = 18.

which can be estimated by $GR_t = \left(\frac{Y_{t+1} - T_t}{Y_t}\right) \times 100$, y_t is the observed value of the variable y at time t and y_{t+1} is the observed value of the variable y at time $t + 1$.

Average growth rate (AGR) can be obtained as the simple average of the annual growth rates (Nath *et al.* 2020). Thus, the AGR for the specified periods can be computed as,

for the period 1, $AGR_i = \frac{\sum_2^9 AGR_i}{8}$, for the period II ,

$AGR_2 = \frac{\sum_{10}^{18} AGR_t}{8}$ and for the period as a whole, AGR

$= \frac{\sum_2^{18} AGRt}{17}$, the difference between AGR and AGR_2 is obtained as, $\Delta AGR = AGR_2 - AGR_t$

The significance of average growth rates for the divided two periods in the population can be tested by using student's t-statistic. The significance of the differences between average growth rates of the two divided periods in the population can be tested by using methods described in Bhattacharya and Roychowdhury (2017) viz., Welch's t-statistic (if the variance of both the populations are unknown but unequal) and Fisher's t-statistic (If the variance of both the populations are unknown but equal).

Study of instability

Instability is measured by using CV for the detrended value (yD) of the series so that the effect of the trend can be eliminated (Dash *et al.* 2017(b)). The detrended values should be centered accordingly.

The CV is defined as, $CV = \frac{\sigma_{yD}}{Mean_{yD}} \times 100$ or $CV = \frac{\sigma_{yD}}{\mu_{yD}} \times 100$,

Where μ_{yD} and σ_{yD} are mean and standard deviation of the detrended series (YD),respectively.

Three CVs are obtained separately for period I, period II and period as a whole. The differences in CVs between period I and period II are obtained to study the magnitude of instability for these divided periods. Student's t-test is used for testing the significance of CV in the population (CV for the period as a whole). In this case, the hypotheses are constructed as, H_0: population CV = 0 and H_1: population CV > 0. Here, the zero value of population CV indicates that the population mean is too large. This hypothesis is tested using the test statistic defined as,

$$t_0 = \frac{\tilde{CV}}{s_e(CV)} \sim t_{(n-1)}$$

where CV = coefficient of variation, n = total number of observations and $_{se}$(CV) is the standard error of CV, which is obtained by $_{Se}(CV) = \frac{CV}{\sqrt{2n}}$ (Dash *et al.* 2017 (b)).

Significance of change in the CVs of population is also tested by using student's t-test. The hypotheses are taken as, H_0: ΔCV = 0 and H_1: ΔCV ≠ 0. The test statistic is defined as,

$$t_0 = \frac{\Delta CV}{S_c(\Delta CV)} \sim t_{(2n-2).}$$ Dash *et al.* (2017 (b)

where n denotes total number of observations in each period and $\mathbf{s}_e(\Delta CV)$ denotes the corresponding standard error of the difference in the estimates of CVs which is obtained as,

$S_e(CV) = \frac{CV}{\sqrt{(n_1 + n_2)^2}}$ where CV is the CV for both the periods jointly. which is obtained as,

$$CV^t = \left[\frac{\{(n_1 - 1) \times CV_1 + (n_2 - 1) \times CV_2\}\}}{(n_1 + n_2 - 2)}\right]$$

with n_1 =number of observations in period 1 and n_2= number of observations in period II.

Forecast using ARIMA models

Auto-regressive integrated moving average (ARIMA) is a time series model which was proposed by Box and Jenkins. The basic assumptions imposed on the residuals of the ARIMA model are independence and normality. For "*p*" AR components, "*q*" MA components and "*d*" number of differences, the model can be expressed as (8).

$$\text{ARIMA}\ (p,\ d,\ q) = (1\text{-}\sum_{i=1}^{p} \alpha_I \mathrm{B}^I)\ (1\text{-B})^d\ Y_t = (1+ \sum_{i=1}^{q} \theta_t \mathrm{B}^t)\ \varepsilon_t \qquad (8)$$

where, α_i = the coefficient of AR component at lag i, θ_i = the coefficient of MA component at lag *i*, B = the backshift operator, Y_t = the actual value of the series at time t, ε_t = white noise error at time *t*.

ARIMA model fitting

The following steps are to be followed while fitting an appropriate ARIMA model (Nath *et al.* 2019):

(i) **Model identification:** Deciding the parameters of ARIMA model.

(ii) **Estimation of parameter:** Finding the estimates of these parameters and testing their significance.

(iii) **Diagnostic check**: Checking the assumptions imposed on white noise error term of the model.

Forecast values of area, production and yield

Forecast of the future values is obtained by using the best fit models. For upcoming five years (i.e., 2015-16 to 2019-20) the forecast values are obtained with 95 and 99% prediction intervals. This process can forecast the future values up to a certain extent but for a long term forecast the prediction error increases significantly. Since forecast error increases significantly in forecasting with ARIMA models, then it is suggested that a short-term forecast is to be performed not a forecast for quite far future values (Sarika *et al.* 2011).

Empirical findings

Fitting of spline regression models

All the models fitted to the data on area under vegetables, show that the estimates of the parameters, β_0 and β_1 (estimated by b_0 and b_1, respectively) are significantly different from zero (Table 1). It indicates that these models can be tried for further estimation of growth rate and instability. It is also evident from Table 2, that at least one among Durbin-Watson statistic, Park's ln (t) statistic and Shapiro-Wilk's statistic is significant for power, compound, logarithmic and quadratic spline models which indicates that errors of these models do not hold the assumptions imposed on the residuals. The non-significant statistics of the residuals diagnostic checks for the linear spline model indicates the appropriateness of the model. It can also be seen that for area under vegetables, only the linear spline model is found to be suitable model which also has the highest values of R^2, adjusted R^2 and the lowest value of RMSE. The significance of change in the slope coefficient from period I to period II is tested by the coefficient A_1 (estimated by a_1) is also found to be significant for linear spline model.

Table 1. Parameter estimates of the models fitted to data on area under vegetables. Figures in the parentheses are the standard error of the estimates.

Model	Parameter estimate				
	b_0	b_1	a_1	b_2	a_2
Linear spline	767.44** (8.22)	13.98** (1.46)	-4.326* (1.498)		
Power spline	763.09** (8.34)	0.064 ** (0.006)	-0.017 (0.008)		
Compound spline	768.89** (5.20)	1.017** (0.01)	0.993** (0.01)		
Logarithmic spline	761.58** (8.69)	53.253** (5.515)	29.648 (35.593)		
Quadratic spline	749.66** (12.79)	23.67** (5.872)	79.695** (27.365)	-0.97** (0.57)	-4.18** (1.26)

Table 2. Model fit statistics and residual diagnostics of the models fitted to the data on area under vegetables.

Model	Model fit statistics				Residuals diagnostics		
	R^2	Adj. R^2	F-statistic	RMSE	Durbin-Watson statistic	Park's ln (*t*) statistic	Shapiro-Wilk's statistic
Linear spline	0.986	0.986	2854.50**	4.79	1.96	0.412	0.8992
Power spline	0.977	0.975	667.11**	10.19	1.56	-0.428	0.8546*
Compound spline	0.985	0.980	2693.15**	4.99	1.83	0.425	0.8465*
Logarithmic spline	0.448	0.435	34.14*	29.92	0.77**	2.545*	0.8520*
Quadratic spline	0.881	0.875	22.12 *	13.92	0.94**	2.225*	0.9286

In this study, * and ** have been used for indicating the significance at 5% and 1% level of significance, respectively.

Table 3. Parameter estimates of the models fitted to data on production of vegetables.

Model	Parameter estimate				
	b_0	b_1	a_1	b_2	a_2
Linear spline	9106.66** (93.43)	255.62** (16.60)	124.011* (46.372)		
Power spline	9101.556** (177.64)	0.091** (0.011)	0.034 (0.022)		
Compound spline	9159.117** (55.89)	1.025** (0.01)	1.005** (0.001)		
Logarithmic spline	9055.199** (189.804)	934.717** (120.449)	633.566 (763.575)		
Quadratic spline	9166.153** (174.26)	223.171* (80.011)	-2108.544** (601.572)`	3.24** (7.80)	197.97** (49.13)

Estimates of the parameters, obtained by b0 and b1 for β0 and β1, respectively are found to be significant for all the models fitted to the data on production of vegetables (Table 3). Residual diagnostics of these models are also performed which suggests that all the spline models have a significant F-statistic value (Table 4). For the compound spline model, all the statistics of residual diagnostics are non-significant. It suggests that this is the appropriate spline model because it holds the assumptions imposed on the error terms. The significance of change in the slope coefficient (estimated by a_1) from period I to period II for compound spline model is checked and found to be highly significant. Therefore, compound spline model is considered to be suitable spline model for estimation of growth rate and instability in the production of vegetables. The parameter estimates given in Table 5 reveals that both the estimates of parameters β0 and $β_1$ for all the spline models fitted to the data on yield of vegetables are significant except for quadratic

Table 4. Model fit statistics and residual diagnostics of the models fitted to the data on production of vegetables.

Model	Model fit statistics				Residuals diagnostics		
	R^2	Adj. R^2	F-statistic	RMSE	Durbin-Watson statistic	Park's ln (*t*) statistic	Shapiro-Wilk's statistic
Linear spline	0.981	0.980	2085.55	148.42	0.30	3.730*	0.9601
Power spline	0.957	0.955	359.80	364.64	0.58**	0.646	0.8585*
Compound pline	0.974	0.970	1560.26*	173.06	0.43	3.953	0.9292
Logarithmic spline	0.639	0.631	74.47*	641.86	0.67**	4.796*	0.8969*
Quadratic spline	0.704	0.689	7.12 *	581.84	0.90**	4.837*	0.8895*

Table 5. Parameter estimates of the models fitted to data on yield of vegetables.

Model	Parameter estimate				
	b_0	b_1	a_1	b_2	a_2
Linear spline	11.90** (0.13)	0.10** (0.02)	0.17** 0.052)		
Power spline	11.927** (0.17)	0.027* (0.009)	0.051** (0.017)		
Compound spline	11.912** (0.08)	1.008** (0.01)	1.012** (0.001)		
Logarithmic spline	11.921** (0.174)	0.332* (0.11)	0.263 (0.389)		
Quadratic spline	12.196** (0.20)	-0.062 (0.094)	-3.317** (0.791)	0.02** (0.01)	0.27** (0.07)

Table 6. Model fit statistics and residual diagnostics of the models fitted to the data on yield of vegetables.

Model	Model fit statistics				Residuals diagnostics		
	R^2	Adj. R^2	F-statistic	RMSE	Durbin-Watson statistic	Park's ln (*t*) statistic	Shapiro-Wilk's statistic
Linear spline	0.928	0.926	527.63**	0.17	53.44	-1.456	0.9671
Power spline	0.909	0.903	158.94*	0.31	0.57**	1.005*	0.9247
Compound spline	0.915	0.910	451.83*	0.18	0.44**	-1.392*	0.9617
Logarithmic spline	0.720	0.713	107.78*	0.33	0.59**	-0.846	0.8921*
Quadratic spline	0.067	0.065	1.21 *	0.80	0.87**	-0.348	0.8843*

spline model. From the model fit statistic and residual diagnostic (Table 6), it is clear that quadratic spline model does not has a significant F-statistic value. The non-significant statistics of residual diagnostic checks for the linear spline model suggest that this model can be used for estimation of growth rate and instability in yield of vegetables. It can also be noted that the values of R^2, adj. R^2 are the highest for linear spline model along with the least value of RMSE. The significance of the change in the slope coefficient from period I to period II is also found to be significant for this model which indicates the appropriateness of linear spline model.

Table 7. Average growth rates in area, production and yield of vegetables for state of West Bengal (in percent).

Particulars	GR	GR_1	GR_2	ΔGR
Area	1.34** (0.08)	0.64** (0.02)	1.03** (0.09)	0.39** (0.02)
Production	2.74** (0.07)	0.95** (0.05)	2.91** (0.07)	1.95** (0.02)
Yield	1.40** (0.14)	0.30** (0.01)	1.92** (0.03)	1.62** (0.03)

Note: Figures in the parentheses are standard error of the estimates, *(significant at 5% level of significance), **(significant at 1% level of significance); GR, GR_1, GR_2 and $\Delta GR = (GR_2 - GR_1)$ stand for growth rate for whole period, period I, period II and difference between growth rates from period II and period I respectively.

Table 8. CV in area, production and yield of vegetables for state of West Bengal (in percent).

Particulars	CV	CV_1	CV_2	ΔCV
Area	0.813** (0.136)	1.191** (0.281)	1.337** (0.315)	0.146 (0.195)
Production	1.891** (0.315)	1.092** (0.257)	2.959** (0.698)	1.867** (0.313)
Yield	1.889** (0.315)	1.301** (0.307)	1.482** (0.349)	0.181 (0.215)

Note: Figures in the parentheses are standard error of the estimates, *(significant at 5% level of significance), **(significant at 1% level of significance); CV, CV_1, CV_2 and $\Delta CV = (CV_2 - CV_1)$ stand for CV for whole period, period I, period II and difference between period II and period I, respectively.

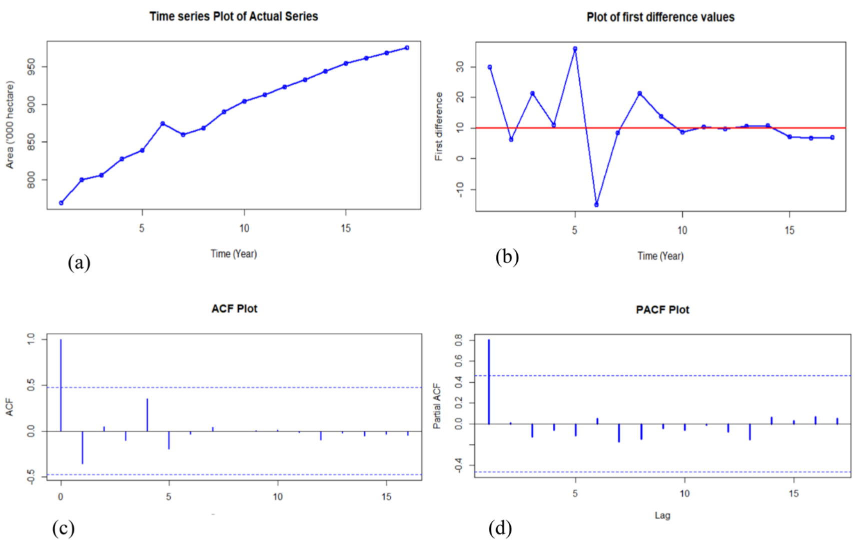


Fig. 1(a). Plot of actual values of the data on area under vegetables against time. **Fig. 1(b)**. Plot of the first difference values of the data on area under vegetables. **Fig. 1(c).** ACF plot of stationary series. **Fig.1(d).** PACF plot of stationary series.

Growth rate and instability of area, production and yield of vegetables

The average growth rates and instability (measured by CV) for area, production and yield of vegetables are obtained (Tables 7-8. respectively). It can easily be concluded that the average growth rates of area, production and yield of vegetables are significantly positive for the period as a whole, period I, period II and difference between period II and period I. Also, the CVs for area, production and yield of vegetables are significantly positive except for the area and yield in case of difference between the CVs from period II to period I.

Fitting of ARIMA models for forecast

Data on area under vegetables

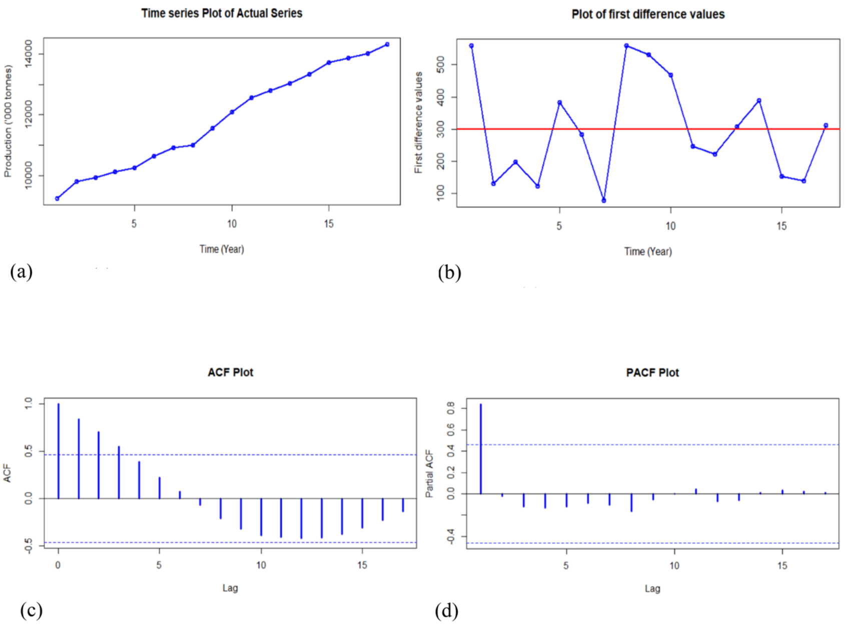
In the plot of actual series (Fig.1(a)). it is observed that the actual series is not stationary as it has an increasing trend. So, the first difference of the actual series is obtained and plotted (Fig.1(b)). which looks like a stationary process. The ACF (Fig.1(c)). and PACF (Fig.1(d)). plots are obtained for this stationary series and the possible values of p and q are then decided accordingly by looking at the significant spikes in these plots. The possible values are obtained as, *p* = 1 and *q* = 0, respectively. By using these possible values, ARIMA (1,1,0) model with and without constant is formed as candidate ARIMA models for forecasting area under vegetables.

Data on production of vegetables

Plot of the actual data points of production of vegetables for the state of West Bengal shows an increasing trend that means the actual series is not stationary (Fig.2(a). Thus, the first difference of the actual series is obtained and plotted (Fig. 2(b)) which appears to be stationary in mean and variance. From the ACF and PACF plots (Fig.2(c). and Fig. 2 (d) of the stationary series the possible values of p and q are obtained as, p = 1 and q = 0. Thus, two candidate ARIMA models are formed by using these possible values as, ARIMA(1,1,0) with constant and ARIMA(1,1,0) without constant for forecasting the production of vegetables.

Data on yield of vegetables

In case of yield of vegetables for the state of West Bengal, plot of the actual data points shows an upward trend indicating that the series is not stationary (Fig. 3(a)). So, the first difference of the actual series is obtained and plotted (Fig 3(b)) which appears to be

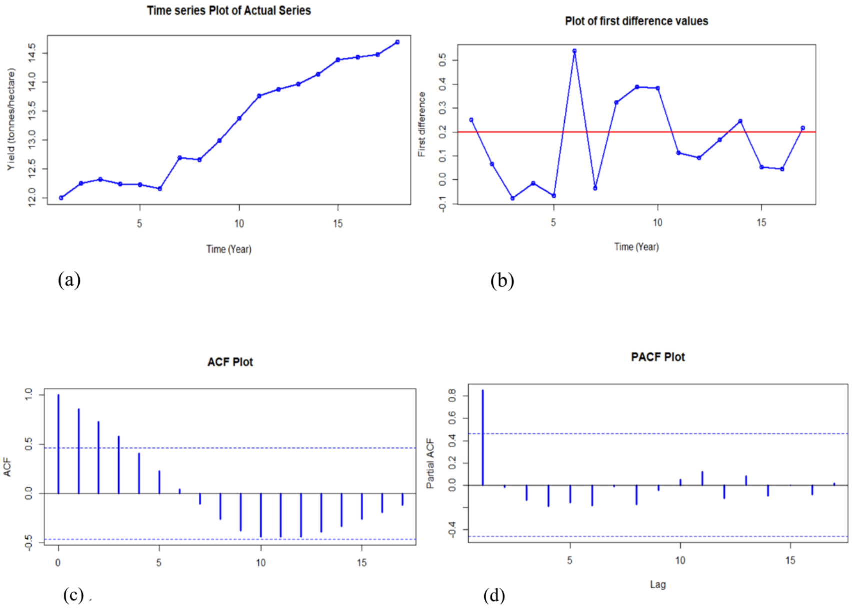


(a) (b)

(c) (d)

Fig. 2 (a). Plot of actual values of the data on production of vegetables against time. **Fig. 2(b).** Plot of the first difference values of the data on production of vegetables. **Fig. 2(c).** ACF plot of stationary series. **Fig. 2(d).** PACF plot of stationary series.

stationary in mean and variance. From ACF and PACF plots of the stationary series (Figures: 3(c) and 3(d)), the possible values of p and q are obtained as, p = 1 and q = 0. Using these possible values two tentative



(a) (b)

(c) . (d)

Fig. 3(a). Plot of actual values of the data on yield of vegetables against time. **Fig. 3(b)** Plot of the first difference values of the data on yield of vegetables. **Fig. 3(c)** ACF plot of stationary series. **Fig. 3(d)** PACF plot of stationary series.

Table 9. Estimates of AR and MA components of the fitted ARIMA models for forecasting the area, production and yield of vegetables in West Bengal.

Particulars	Model	Constant (μ)	Coefficient of AR components		Coefficient of MA components	
			α_1	α_2	θ_1	θ_2
Area	ARIMA (1,1,0)		0.3954* (0.2413)			
	ARIMA (1,1,0) with constant	11.8335** (1.7369)	-0.4040*			
Production	ARIMA (1,1,0)		0.8353** (0.1356)			
	ARIMA (1,1,0) with constant	300.4055**	0.0662* (0.2577)			
Yield	ARIMA (1,1,0)	(40.2436)	0.4059*			
	ARIMA (1,1,0) with constant	0.1582** (0.0406)	-0.0368* (0.2384)			

ARIMA models are formed as, ARIMA(1,1,0) with and without constant for forecasting the yield of vegetables.

The estimates of the parameters of the AR and MA components of the fitted ARIMA models to forecast area, production and yield of vegetables for the state of West Bengal are obtained and their significance is tested (Table 9). There are two candidate ARIMA models for forecasting area, production and yield of vegetables viz., ARIMA(1,1,0) model with and without constant, which is an autoregressive model of order 1 with constant term. The estimate of AR component including the constant term of these models are highly significant.

The model fit statistics and residual diagnostics (Table 10) reflects that neither Ljung-Box test nor Shapiro-Wilk's test is significant for all the models fitted to the data on area, production and yield of vegetables. It indicates that the residuals are serially uncorrelated and normally distributed. It means all the candidate models fit the data well. However, the lowest AICs and MAPEs are observed for ARIMA(1,1,0) with constant models for forecasting area, production and yield of vegetables. Thus, this model has been considered to be the best fit models for the purpose of forecasting.

Forecast values of area, production and yield of vegetables

The ARIMA (1,1,0) model with constant is the best fit ARIMA model for forecasting area, production and yield of vegetables. By using this model, forecast values for consecutive five years starting from 2015-16 to 2019-20 with 95 and 99% prediction intervals are obtained (Table 11).

Table 10. Model fit statistics and residuals diagnostics of the fitted ARIMA models for forecasting the area, production and yield of vegetables.

Particulars	ARIMA model	Model fit statistics			Residual diagnostics	
		AIC	RMSE	MAPE	Ljung-Box statistic	Shapiro-Wilk's statistic
Area	ARIMA(1,1,0)	144.37	14.5298	1.2444	7.6785	0.8841
	ARIMA(1,1,0) with constant	132.26	9.5922	0.8175	0.0012	0.9560
Production	ARIMA(1,1,0)	233.91	196.3366	1.2868	2.8716	0.9780
	ARIMA(1,1,1) with constant	225.62	150.1788	1.0987	0.0032	0.9462
Yield	ARIMA(1,1,0)	0.12	0.2086	1.1698	3.5641	0.9560
	ARIMA(1,1,1) with constant	0.05	0.1684	1.0723	0.0001	0.9597

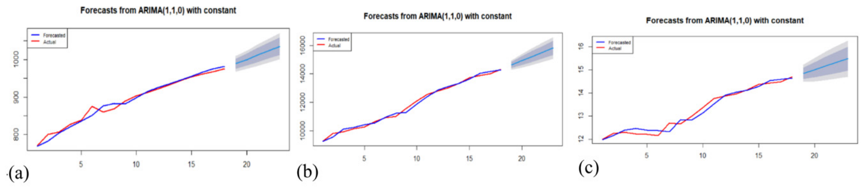


Fig. 4(a). Plot of forecasted values and actual values of area under vegetables. **Fig. 4(b).** Plot of forecast values and actual values of production of vegetables. **Fig. 4 (c).** Plot of forecast values and actual values of yield of vegetables.

Table 11. Final forecast of area, production and yield of vegetables with 95 and 99% prediction intervals.

Particulars	Year	Forecast	LCL 95%	UCL 95%	LCL 99%	UCL 99%
Area (‘000 hectares)	2015-16	14.8511	14.4895	15.2127	14.3759	15.3264
	2016-17	15.0094	14.5074	15.5115	14.3496	15.6693
	2017-18	15.1676	14.5563	15.7790	14.3642	15.9711
	2018-19	15.3259	14.6220	16.0297	14.4009	16.2508
	2019-20	15.4841	14.6986	16.2696	14.4518	16.5164
Production (‘000 tonnes)	2015-16	14628.82	14306.38	14951.26	14205.06	15052.57
	2016-17	14929.28	14457.93	15400.63	14309.82	15548.74
	2017-18	15229.69	14645.43	15813.94	14461.85	15997.52
	2018-19	15530.09	14851.42	16208.76	14638.17	16422.02
	2019-20	15830.50	15069.02	16591.97	14829.75	16831.25
Yield (tonnes per hectare)	2015-16	988.80	968.21	1009.40	961.73	1015.87
	2016-17	999.84	975.86	1023.81	968.33	1031.34
	2017-18	1011.99	983.37	1040.61	974.38	1049.61
	2018-19	1023.69	991.71	1055.68	981.66	1065.73
	2019-20	1035.58	1000.33	1070.84	989.25	1081.91

Fig. 4 (a), 4(b) and 4(c) show the observed values, fitted values and forecast values with 95 and 99% upper and lower prediction intervals of area, production and yield of vegetables, respectively. In all these plots, the line of forecast values is steadily increasing which indicates that area, production and yield of vegetables will increase steadily during the period of forecast.

CONCLUSION

In case of area and yield of vegetables, linear spline while in case of production of vegetables, compound spline model is found to be the best fit spline models. The highest growth rate has been observed for production (2.74%) followed by yield (1.40%) and area (1.34%) of vegetables. It can easily be seen that the higher growth rates are followed by the higher instabilities. The stability in production, yield and area can be achieved by proper scheduling of irrigation, proper management of fertilizers, improved varieties and package of practices. The ARIMA (1,1,0) model with constant is found to be suitable model for forecasting area, production and yield of vegetables. The forecast of the same show a steadily increasing trend in all the variables during the period of forecast i.e., 2015-16 to 2019-20.

REFERENCES

Bhattacharya D, Roychowdhury S (2017) Probability and Statistical Inference: Theory and Practice (3rd ed). U.N. Dhur and Sons, Pvt Ltd, Kolkata.

Dash A, Dhakre DS, Bhattacharya D (2017a) Fitting of appropriate model to study growth rate and instability of mango produc tion in India. *Agricult Sci Digest* 37(3): 191-196.

Dash A, Dhakre DS, Bhattacharya D (2017b) Study of growth and instability in food grain production of Odisha: A statisti cal modelling. *Environ Ecol* 35(4D):3341-3351.

Dash A, Hansdah R (2020) Analytical study of growth and instability of *rabi* oilseed production in Odisha. *J Pharmacog Phytochem* 9 (2): 2145-2149.

Dasyam R, Bhattacharyya B, Mishra P (2016) Statistical Modeling to area, production and yield of Potato in West Bengal, *Int J Agricult Sci* 8 (53): 2782-2787.

Gujarati DN, Porter DC (2012) Basic Econometrics (5th ed). McGraw Hill Publication.

Montgomery DC, Peck EA, Vining GG (2012) Introduction to linear regression analysis (5th ed). John Wiley & Sons, Inclusive, Publication.

Nath B, Dhakre DS, Bhattacharya D (2019) Forecasting wheat production in India: An ARIMA modelling approach.*J Pharmacog Phytochem* 8 (1): 2158-65.

Nath B, Dhakre DS, Sarkar KA, Bhattacharya D (2020) The challenge of selecting the best forecasting model for a time series data. J *Food Agric Environ* 18(2): 97-102.

Sarika Iquebal MA, Chattopadhyay C (2011) Modelling and forecasting of pigeon pea (*Cajanus cajan*) production using autoregressive integrated moving average methodology. *Ind J Agricult Sci* 81(6): 520–523.