International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
Maths 2024; SP-9(5): 149-157

www.mathsjournal.com
Received: 15-07-2024
Accepted: 14-08-2024

DM Padalia
M.Sc. (Agricultural Statistics), Department of Agricultural Statistics, C. P. College of Agriculture, SDAU, Sardarkrushinagar, Gujarat, India

GK Chaudhary
Associate Professor, Department of Agricultural Statistics, C. P. College of Agriculture, SDAU, Sardarkrushinagar, Gujarat, India

PB Marviya
Assistant Professor, Department of Agricultural Statistics, C. P. College of Agriculture, SDAU, Sardarkrushinagar, Gujarat, India

LK Kumawat
M.Sc. (Agricultural Statistics), Department of Agricultural Statistics, C. P. College of Agriculture, SDAU, Sardarkrushinagar, Gujarat, India

IB Savaliya
M.Sc. Scholar, Department of Agricultural Statistics, C. P. College of Agriculture, SDAU, Sardarkrushinagar, Gujarat, India

Tejas S
M.Sc. Scholar, Department of Agricultural Statistics, C. P. College of Agriculture, SDAU, Sardarkrushinagar, Gujarat, India

Corresponding Author:
DM Padalia
M.Sc. (Agricultural Statistics), Department of Agricultural Statistics, C. P. College of Agriculture, SDAU, Sardarkrushinagar, Gujarat, India

Forecasting models for area, production and productivity of wheat in North Gujarat

DM Padalia, GK Chaudhary, PB Marviya, LK Kumawat, IB Savaliya and Tejas S

DOI: https://dx.doi.org/10.22271/maths.2024.v9.i5Sc.1829

Abstract
The present study was carried out to estimate the trends of area, production and productivity of wheat in North Gujarat. Yearly time series data on area, production and productivity of wheat of North Gujarat for the period of 1991-92 to 2021-22 were collected from the published reports by Directorate of Agriculture, Gujarat state, Gandhinagar. Polynomial and Autoregressive Integrated Moving Average (ARIMA) models were applied to analyze the area, production and productivity of wheat in North Gujarat. First, second and third-degree polynomial models were fitted to the original data and moving averages, with the best model selected based on adjusted R^2, Root Mean Square Error (RMSE), Mean Absolute Error (MAE), normality by the Shapiro-Wilk (1965) test and randomness of residuals by the Run test. ARIMA models were evaluated for stationarity and different orders (p, d, q) were assessed using autocorrelation function (ACF), partial autocorrelation function (PACF) and the models were selected on the basis of significant autoregressive and moving average term, lower value of Akaike's Information Criteria (AIC) and Schwartz-Bayesian Criteria (SBC) and normality of residuals by Shapiro-Wilk (1965) test and Box-Ljung (1978) test. The cubic polynomial model on a four year moving average was most suitable for wheat production trends, while ARIMA (0, 2, 1) and ARIMA (0, 2, 2) were optimal for wheat area and production patterns.

Keywords: Forecasting, Polynomial models, ARIMA, ACF, PACF, AIC, SBC, adjusted R^2, S-W test, B-L test

Introduction

In India, wheat occupied area of 31.13 million hectare with production of 109.59 million tonnes and productivity of 35.21 quintal per hectare (Anonymous, 2022a) [1]. In Gujarat, wheat occupied an area of 12,53,834 hectare with production of 40,18,534 MT and productivity of 32.05 quintal per hectare (Anonymous, 2022b) [2].

The most important and widely used time series model is the ARIMA model. The feature of ARIMA modelling over other univariate time series models is that it delivers the least mean squared forecast error variances in addition to revealing the fundamental behaviour in the time series variables. The popularity of the ARIMA model is due to its statistical properties as well as the well-known Box Jenkins methodology in the model building process. ARIMA model is highly efficient in short term forecasting.

Forecasting plays a main role in decision-making to make their future decisions more correct. Government needs forecasting to make various policy and policy decision on pricing, import-export, storage *etc*. A continuous change in productivity of agricultural crops due to many factors such as variations in economic, precipitation, agricultural conditions and technologically. Study the change in that productivity is useful in evaluating efforts to increase agricultural production. That's why the forecasting of productivity of various agricultural crops allows the accurate prediction of production levels to be made in the future years. Thus, forecasting is one of the leading tools in the field of agricultural production to make an effective growth policy and economic plans.

Methodology

Yearly time series data (1991-92 to 2021-22) on wheat area, production and productivity were obtained from the Directorate of Agriculture, Gandhinagar, Gujarat. Due to geographical variations across districts, productivity was calculated using the weighted mean method. The dataset was split into a testing set (1991-92 to 2017-18) and a validation set (2018-19 to 2021-22).

Linear regression approach (Rangaswamy, 2006) [11]

Regression analysis was used to model the relationship between the dependent variable (Y: area, production, productivity) and the independent variable (time, t), applied to both original and moving average data.

$$\hat{Y} = a + bt$$

Where, a and b are the regression constant and regression coefficient, respectively to be estimated.

Quadratic regression approach (Montgomery *et al*., 2003) [8]

The fitted regression equation was as under.

$$\hat{Y} = a + bt + ct^2$$

The parameter *viz.*, a, b and c were estimated by using 'Principle of least square'.

Third degree polynomial approach (Montgomery *et al*., 2003) [8]

The model for the third-degree polynomial fitted to the data was as under.

$$\hat{Y} = a + bt + ct^2 + dt^3$$

The constant (a) and coefficient (b, c and d) were estimated using 'least square method'.

Goodness of fit of the models (Montgomery *et al*., 2003) [8]

To test the goodness of fit of the fitted polynomial model, the Coefficient of Determination (R^2) defined as the proportion of total variation in the response variable being explained by the explanatory variables and was calculated as under.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

Where, $\hat{Y}_i$ = Predicted value of dependent variable
$\bar{Y}$ = Average value of dependent variable

$$RMSE = \left[\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 / n\right]^{1/2}$$

$$MAE = \sum_{i=1}^{n} |Y_i - \hat{Y}_i| / n$$

Additionally, Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) were computed to evaluate model adequacy (Liew *et al*., 2000) [5]. Models with lower R^2, RMSE, and MAE values were considered better fits.

Test the randomness of the residuals (Sidney and Castellan, 1988) [14]

The sample must be random to arrive at conclusion about the population by using the information in the sample.

Suppose m be the number of elements of one kind (+ve sign residuals) and n be the number of elements of the other kind (-ve sign residuals) in sequence of N = m + n. To use one sample run test, first observe the m and n events in which they occurred and determined the value of r (no. of runs). If m or n is larger than 20, determine the value of Z as under.

$$\text{Mean} = \mu = \frac{2mn}{N} + 1$$

$$\text{Standard Deviation} = \sigma_r = \left[\frac{2mn(2mn-N)}{N^2(N-1)}\right]^{1/2}$$

$$Z = \frac{r - \mu_r}{\sigma_r} = \frac{r + h - (\frac{2mn}{N}) - 1}{\sqrt{\frac{2mn(2mn-N)}{N^2(N-1)}}}$$

where $h = +0.5$ if $r < \frac{2mn}{N} + 1$ and $h = -0.5$ if $r > \frac{2mn}{N} + 1$

Use the normal table (Sidney and Castellan, 1988) [14] for testing Z value. The non-significant of Z value indicates randomness of the residual.

Test for normality of the residuals (Shapiro and Wilk, 1965) [13]

The Shapiro - Wilk (1965) statistic was used to test whether the residuals are normally distributed or not. The test is based on n residuals. These are arranged in non-decreasing sequence and it is designated by $e_{(1)}, e_{(2)}, e_{(3)} \ldots e_{(n)}$. The following hypothesis is to be tested.

H_0: The residuals are normally distributed
H_1: The residuals are not normally distributed
The required test statistic W is defined as

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

$$\text{Where } S^2 = \sum a(k)[e(n+1-k) - e(k)]$$

The parameter k takes the values

$$K = \begin{cases} 1, 2, 3, \ldots . \frac{n}{2} & \text{when n is even} \\ 1, 2, 3, \ldots . \frac{(n-1)}{2} & \text{when n is odd} \end{cases}$$

$$b = \sum_{i=1}^{n} (e_i - \bar{e})^2$$

The values of coefficients "a(k)" for different values of n and k are given by Shapiro and Wilk (1965). When the calculated value of W is non-significant *i.e.*, very close to unity, the null hypothesis regarding normality of residual was accepted.

Autoregressive (AR) and moving average (MA) models (Pankratz, 1983) [9]

Correlation measures the relationship between observations in a time series. In time series models, the correlation between the current value (Y_t) and past values ($Y_{t-1}, Y_{t-2}, \ldots$) is examined.

Autoregressive (AR) process

$$Z_t = C + \varphi_1 Y_{t-1} + a_t$$

Where, Z_t = time sequenced random variable
C = constant term related to mean (μ) such that C =μ (1-φ_1)
φ_1 = relationship of Y_t with Y_{t-1}
a_t = random shock element at time t
Moving average (MA) process

$Z_t = C - \theta_1 a_{t-1} + a_t$

Where, C = constant term related to mean μ and

θ = relationship of a_t with a_{t-1}
a_t = random shock element at time t

The AR process relies on past Y values, while the MA process depends on past error terms.

Fitting of Box-Jenkins ARIMA models
Box-Jenkins time-series models *i.e.*, ARIMA (p, d, q) is known as "UBJ technique (Univariate Box-Jenkins) technique" (Box *et al.*, 1976) [3]. An ARIMA model is an algebraic statement telling how observations on a variable are statistically related to past observation.
This model amalgamates three types of process, namely autoregressive of order p; differencing to make a series stationary of degree d and moving average of order q. This method applies only to a stationary time series data. When the data is non-stationary then it has to be brought into stationary by the method of differencing.

Test for stationarity
The stationarity requirement ensures that one can obtain useful estimates of the mean, variance and ACF from a sample. If a process has a mean that is changing in each time period, one could not obtain useful estimates since only one observation available per time period. This necessitates testing any observed series of data for stationarity.
There are three ways to determine whether the above mentioned stationarity requirement is met.

i) Examine the realization visually to see if either the mean or the variance appear to change over time.
ii) Examine the estimated AR coefficient to see if it satisfies the stationary condition. In case of AR (1) process the condition for stationary is that absolute value of φ_1 must be less than one, or symbolically, $|\varphi_1| < 1$. In practice one do not know φ_1, therefore, one apply the condition to $\widehat{\varphi 1}$ (*i.e.*, estimate of φ_1) rather than φ_1.
iii) For an MA (1) process the corresponding condition is that the absolute value of θ_1 must be less than one. Which is called the condition of invariability, or in symbols $|\theta_1| < 1$.

To find out the t-value of the estimated auto-correlation Barlett's approximate expression for the standard error of the sampling distribution of r_k values can be used. The estimated standard error, designated as $S(r_k)$ is calculated using the following expression.

$$S(r_k) = \left(1 + 2\sum_{j=1}^{k-1} r_j^2\right)^{1/2} n^{-1/2}$$

The following null hypothesis is to be tested
$H_0 : \rho_k = 0$ for k = 1, 2, 3, … using the test statistics

$$t_{rk} = \frac{r_k - \rho_k}{S(r_k)}$$

If the value of t comes out to be significant, we reject H_0 at the level of significance and conclude that $\rho \neq 0$.

Unit root test (Augmented Dickey-Fuller test) (Harris, 1992) [4]
The mean, variance and autocovariance of a time series are time-invariant, meaning they remain constant over time. Testing for stationarity is crucial because a non-stationary series limits analysis and forecasting to the specific time period observed. The Augmented Dickey-Fuller (ADF) test is commonly used to determine whether a time series is stationary or contains a unit root.
The presence of unit root (non-stationarity) in the underlying series is tested by performing Augmented Dickey-Fuller test using the following regression:

$$\Delta Y_{it} = \alpha + \beta_i T + \delta_i Y_{it-1} + b_i \sum_{i-1}^{p} \Delta Y_{i-1} + e_i i{-}1$$

Where:
Y_{it} = Price of a commodity in a given market 'i' at a time 't';
$\Delta Y_{i-1} = Y_{t-1} - Y_{t-2}$ (t-1 – lagged prices and Δ – differenced series);
T = Time trend;
α = Drift parameter;
β_i, δ_i and b_i = Coefficients;
and e_i = Pure white noise error-term.
p is the optimal lag value which is selected on the basis of Schwartz information criterion (SIC).

H_0: $\delta = 0$

H_1: $\delta < 0$

The possibility of acceptance and rejection of H_0 is based on the tau statistic or test (τ), and the estimation procedure of tau statistic (τ) is as follows:

1. Estimate the equation by OLS method.
2. Divide the estimated coefficient of Y_{t-1} by its standard error and refer to the DF (Dickey-Fuller) table.
3. If the computed absolute value of the ($|\tau|$) exceeds the absolute DF or MacKinnon 18 critical tau values, then the null hypothesis (δ = 0) is rejected, in which case the time series if stationary.
4. On the other hand if the computed absolute value of the ($|\tau|$) does not exceeds the absolute DF or MacKinnon critical tau values, then the null hypothesis (δ = 0) is accepted, in which case the time series if non stationary.

Steps in ARMA model:
a) Identification
b) Estimation
c) Diagnostic checking

a) Identification
In the identification stage of ARIMA modeling, the estimated ACF and PACF are compared with theoretical patterns to select a tentative model. The principle of parsimony is applied, favoring models with fewer parameters. Key characteristics are:

1. A stationary AR process has an ACF that gradually decays to zero, while its PACF cuts off after a few spikes, with the last PACF spike indicating the AR order (p).

2. A MA process has an ACF that cuts off after a few spikes, with the last spike indicating the MA order (q), while the PACF gradually decays.

b) Estimation stage

During the identification stage of ARIMA modeling, one or more tentative models are selected based on their statistical adequacy for the data. The next step is to obtain precise parameter estimates using the least squares method, as proposed by Box and Jenkins. Software packages such as JMP (student version) and SPSS were utilized for estimating the relevant parameters through iterative procedures in the analysis.

c) Diagnostic checking

A statistically adequate model has residuals that are independent and not autocorrelated. To test this, the residual autocorrelation function (ACF) is calculated using the residuals ($\hat{a}_t$) from the estimated model, rather than the observed data (z_t). The residual autocorrelation coefficient is computed as:

$$r_k(\hat{a}) = \frac{\sum_{t=1}^{n-k}(\hat{a}_t-\bar{a})(\hat{a}_{t+k}-\bar{a})}{\sum_{t=1}^{n-k}(\hat{a}_t-\bar{a})^2}$$

Where, $r_k(\hat{a})$ is the residual autocorrelation.

For an ideal ARIMA model, the residual autocorrelation coefficients should be statistically zero, indicating uncorrelated random shocks.

Goodness of fit of ARIMA model (Sarada and Prajneshu, 2002) [12]

Autoregressive (AR)

$$Z_t = C + \varphi_1 Y_{t-1} + a_t$$

Where, n denotes number of effective observation and $\hat{\sigma}^2$ is the white noise variance.

Moving average (MA)

$$Z_t = C - \theta_1 a_{t-1} + a_t$$

Where, m is the sum of number of AR and MA parameters *i.e.*, m = p + q.

The model with lower values of AIC and SBC was considered better model.

An ARIMA model is commonly denoted as (p, d, q), where p is the number of the autoregressive terms, q denotes the number of the moving average terms and d indicates the number of differences required for stationarity.

Test for independence of errors (Chi – squared test)

Ljung and Box (1978) [6] suggested a test statistic based on all the residual autocorrelations used to test whether the residual autocorrelations is independent or not. The following joint null hypothesis about the correlations among the random shocks was tested

$$H_0: \rho_1(a) = \rho_2(a) = \ldots = \rho_k(a) = 0$$

With the test statistic

$$Q = n(n+2)\sum_{k=1}^{k}(n-k)^{-1} r_i^2(\hat{a})$$

Where, n is the number of observations

The statistics Q approximately follows a Chi - square (χ^2) distribution with (K - m) degrees of freedom where K is the number of residual autocorrelation and m is the number of parameters estimated in the ARIMA model. Non-significant χ^2 (or Q value) indicated that residuals are independent.

If the model is found to be inadequate, the three stages are repeated until satisfactory ARIMA model is selected for representing the time-series observations under consideration.

Criteria for selection of model

In case of polynomial models, the model was selected if it full fills the following characteristics.

1. The residuals should be normally and independently distributed.
2. The regression coefficient in the model should be significant.
3. The model should possess significant F value for coefficient of determination.
4. The coefficient of determination (R^2) should be higher.
5. The Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) should be lower.

In the case of stochastic time-series (ARIMA) models, the model was selected if its full fills the following characteristics (Makridakis *et al.*, 2003) [7].

1. It is parsimonious (uses the smallest number of coefficients needed to explain the available data).
2. It is stationary (AR coefficients which satisfy some mathematical inequalities).
3. It is invertible (MA coefficients which satisfy some mathematical inequalities).
4. The estimated coefficients should be significant (absolute t – values above 2.0 or larger).
5. It should have statistically independent and normally distributed residuals.

Results and Discussion

Trends on wheat area in North Gujarat

Fitting of polynomial models

The results of fitted polynomial models are given in Table 1. The results indicated that in all the polynomial models the coefficient of determination (R^2) was found to be highly significant. The regression constants value of all three linear, quadratic and cubic models in all data approaches were found to be highly significant. Linear regression coefficients of all data approaches of linear model and moving averages of cubic model were found to be highly significant and original data approach of cubic model found to be significant and all data approaches of quadratic model found to be non-significant. All the quadratic regression coefficients except original and three year moving average data approach of quadratic model were found to be highly significant. While, quadratic regression coefficient for three years was found significant. Cubic regression coefficient was found to be significant for original data approach and highly significant for three, four and five year moving averages data approach of cubic model. In original data approach the value of adjusted R^2 was improved by 2.20 per cent in quadratic regression as compared to linear regression while, in case of cubic regression, it was improved by 5.10 per cent over quadratic model. By taking moving average of five year the improvement in adjusted R^2 was observed to be 7.70, 14.10 and 12.80 per cent in case of first, second- and third-degree polynomial models, respectively over original data approach.

The third degree model showed comparatively lower values of root mean square error (RMSE) and mean absolute error (MAE). Among these the least RMSE and MAE were observed in case of model based on five years moving averages. The criteria for testing normality (Shapiro-Wilk test) of residuals indicated that, all models had normally distributed residuals except linear model of three-year moving average. The test of randomness of residuals (Run test) indicated that none of the model was randomly distributed.

Table 1: Fitted linear and non-linear models for area of wheat in North Gujarat

Model	Moving average	Regression constant	Regression coefficients			R^2 (%)	Adj. R^2	RMSE	MAE	S-W test	Run test |Z|
		a	b	c	d						
Linear	Original	1066.31**	107.22**	-	-	73.90**	72.80	496.85	422.04	0.95	3.13**
	3 Years	1092.80**	113.57**	-	-	78.20**	77.20	432.79	369.83	0.90*	2.85**
	4 Years	1103.49**	117.20**	-	-	79.70**	78.80	409.51	350.74	0.91	3.54**
	5 Years	1106.88**	121.29**	-	-	81.40**	80.50	384.85	334.32	0.92	3.41**
Quadratic	Original	1488.93**	19.78	3.12	-	76.90**	75.00	467.18	381.49	0.97	2.74**
	3 Years	1641.87**	-8.43	4.69*	-	83.70**	82.20	374.03	321.29	0.94	3.26**
	4 Years	1723.80**	-25.94	5.72**	-	86.90**	85.70	328.31	270.71	0.95	3.13**
	5 Years	1776.07**	-39.30	6.69**	-	90.10**	89.10	281.33	229.77	0.95	2.98**
Cubic	Original	2241.91**	-276.34*	29.08**	-0.61*	82.40**	80.10	407.91	356.51	0.92	2.74**
	3 Years	2364.05**	-312.36**	33.35**	-0.73**	89.10**	87.50	306.38	265.02	0.93	2.44*
	4 Years	2400.02**	-320.85**	34.62**	-0.77**	91.70**	90.50	261.23	222.07	0.95	3.13**
	5 Years	2372.68**	-309.31**	34.22**	-0.76**	93.80**	92.90	221.42	193.34	0.94	2.98**

*,** indicates significant at 5 per cent and 1 per cent level of significance, respectively

Fitting of ARIMA models

As the series was found non stationary, the new variable X_t was constructed by taking difference of two (*i.e.*, d = 2) to make the data stationary.

The autocorrelation (ACF) and partial autocorrelation (PACF) coefficient of various order of X_t were computed to identify the value of p and q. From the observed model ARIMA (0, 2, 1) had lowest AIC (Akike Information Criterion) and SBC (Schwartz-Bayesian Criterion), RMSE and MAE value and highly significant MA term with highest adjusted R^2 (79.89%) (Table 2). The assumptions of residuals tested by the Shapiro-Wilk test and Box-Ljung (Q) test indicated that the selected ARIMA model satisfied the assumptions of residuals.

In the fitting of polynomials none of the model was found suitable while, in fitting of ARIMA models, ARIMA (0, 2, 1) was found suitable to predict the pattern of wheat area in North Gujarat.

Thapa *et al.* (2022) [15] also found ARIMA (0, 2, 1) for area of vegetable crops in Nepal. Thus these results are in agreement with these authors.

In order to check the validity of these forecasted values, they were compared with the actual values of wheat area during the post sampled forecast period for ARIMA (0, 2, 1) model which are presented in Table 3 and graphically depicted in Fig. 1.

Table 2: Fitted ARIMA models for area of wheat in North Gujarat

Model	AIC	SBC	Constant	AR (φ)		MA (θ)		Adj. R^2 (%)	RMSE	MAE	S-W test	B-L test (Q)
				AR (1)	AR (2)	MA (1)	MA (2)					
ARIMA (0, 2, 1)	378.49	380.92	0.25	-	-	0.99**	-	79.89	439.47	345.06	0.98	8.96
ARIMA (1, 2, 2)	379.31	384.18	2.80	1.00**	-	2.00	-1.00	79.52	459.81	329.48	0.98	8.91
ARIMA (0, 2, 2)	380.30	383.95	0.09	-	-	1.10**	-0.10	79.27	447.63	339.56	0.98	9.02
ARIMA (1, 2, 1)	380.35	384.00	0.11	-0.07	-	0.99**	-	79.19	447.68	341.30	0.98	9.10
ARIMA (2, 2, 1)	381.83	386.70	0.28	-0.09	-0.14	0.99**	-	78.80	451.50	340.46	0.97	8.69

*,** indicates significant at 5 per cent and 1 per cent level of significance, respectively

Table 3: Testing of forecast values for four year by using best fitted ARIMA model (0, 2, 1) for wheat area ('00 ha) of North Gujarat

Year	Observed	Predicted		
		ARIMA (0, 2, 1)	Deviation	Per cent deviation
2018-19	2935.07	3767.51	-832.44	28.36
2019-20	3746.85	3855.71	-108.86	2.90
2020-21	3727.23	3944.17	-216.94	5.82
2021-22	3452.88	4032.89	-580.01	16.79
Mean				13.47

It is observed that the mean per cent deviation between forecasted and actual area based on ARIMA (0, 2, 1) was 13.47 per cent. This proved that the best fit model in validation test for forecasting the area of wheat in North Gujarat.

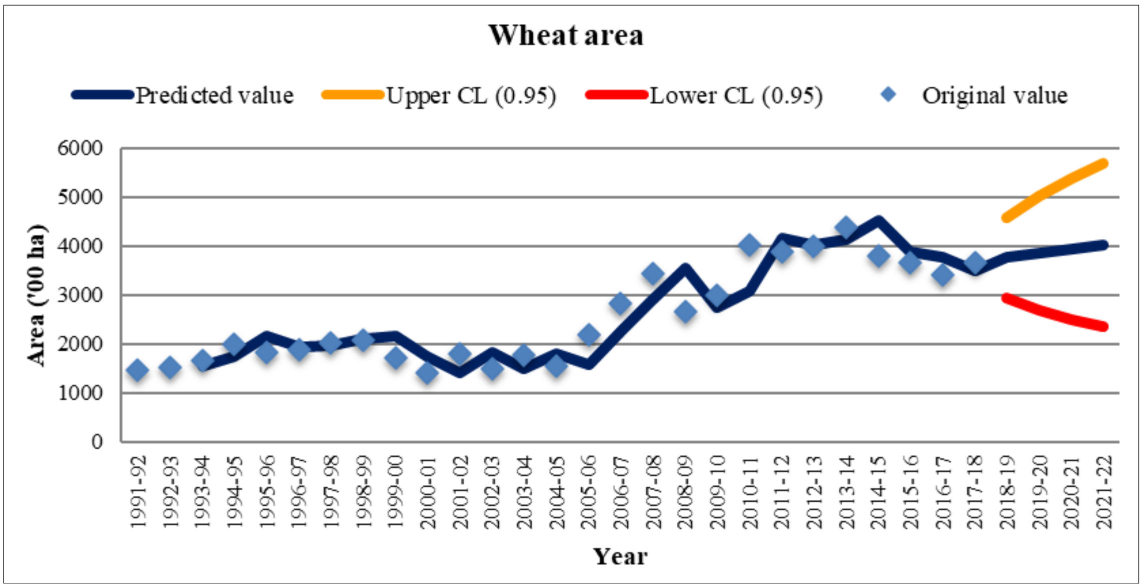

Fig 1: Forecasted wheat area of North Gujarat fitted with ARIMA (0, 2, 1)

Trends on wheat production in North Gujarat

Fitting of polynomial models

The results of fitted polynomial models are given in Table 4. The results indicated that in all the polynomial models the coefficient of determination (R^2) was found to be highly significant. The regression constants value of all three linear, quadratic and cubic models in all approaches were found to be highly significant. Linear regression coefficients were found to be highly significant in all data approaches of linear model, in cubic model three year moving average found to be significant and other moving averages found to be highly significant and in quadratic model and cubic model, original data approach was found to be non-significant. While, quadratic regression coefficients were found to be highly significant for all moving averages data approach, except original data approach of quadratic model. Cubic model of original data approach was found to be significant. Cubic regression coefficient was found to be significant for original and three year moving average data approach and for four- and five-year moving average found to be highly significant. In original data approach the value of adjusted R^2 was improved by 3.30 per cent in quadratic regression as compared to linear regression, in case of cubic regression, it was improved by 3.40 per cent over quadratic model. By taking moving average of five year the improvement in adjusted R^2 was observed to be 9.90, 17.70, 16.90 per cent in case of first, second- and third-degree polynomial models, respectively over original data.

The third degree model showed comparatively lower values of root mean square error (RMSE) and mean absolute error (MAE). Among these the least RMSE and MAE were observed in case of model based on five years moving averages. The criteria for testing normality (Shapiro-Wilk test) of residuals indicated that, all models had normally distributed residuals. The test of randomness of residuals (Run test) indicated that four year moving average of cubic model was randomly distributed.

Table 4: Fitted linear and non-linear models for production of wheat in North Gujarat

Model	Moving average	Regression constant	Regression coefficients			R^2 (%)	Adj. R^2 (%)	RMSE	MAE	S-W test	Run test |Z|
		a	b	c	d						
Linear	Original	2194.32**	346.18**	-	-	70.70**	69.60	1734.94	1470.81	0.96	3.13**
	3 Years	2293.72**	364.17**	-	-	76.90**	75.90	1438.04	1238.48	0.94	2.85**
	4 Years	2345.21**	374.74**	-	-	78.40**	77.50	1359.66	1173.69	0.94	2.71**
	5 Years	2342.19**	388.40**	-	-	80.40**	79.50	1271.39	1089.71	0.93	3.41**
Quadratic	Original	3840.55**	5.58	12.16	-	74.90**	72.90	1605.03	1317.90	0.96	2.35*
	3 Years	4286.84**	-78.74	17.03**	-	83.90**	82.40	1201.34	996.23	0.97	2.44*
	4 Years	4619.64**	-150.12	20.99**	-	87.80**	86.70	1021.41	811.91	0.96	3.13**
	5 Years	4770.07**	-194.28	24.27**	-	91.40**	90.60	841.82	679.63	0.97	2.98**
Cubic	Original	5988.53**	-839.14	86.23*	-1.79*	79.00**	76.30	1467.67	1251.63	0.84	1.95*
	3 Years	6357.08**	-949.97*	99.18**	-2.10*	88.10**	86.40	1031.97	911.27	0.95	2.44*
	4 Years	6551.06**	-992.46**	103.53**	-2.20**	91.60**	90.30	848.89	725.68	0.97	1.46
	5 Years	6386.83**	-925.97**	98.89**	-2.07**	94.10**	93.20	698.20	602.01	0.95	2.12*

*,** indicates significant at 5 per cent and 1 per cent level of significance, respectively

Fitting of ARIMA models

As the series was found non-stationary, the new variable X_t was constructed by taking difference of two (*i.e.*, d = 2) to make the data stationary.

The autocorrelation (ACF) and partial autocorrelation (PACF) coefficient of various order of X_t were computed to identify the value of p and q. This suggested that algebraic family of ARIMA (0, 2, 1), ARIMA (0, 2, 2), ARIMA (1, 2, 1), ARIMA (2, 2, 1) and ARIMA (1, 2, 2) were the candidate models for fitting of ARIMA model, the result are given in Table 5. From the observed model ARIMA (0, 2, 2) had lowest AIC (Akike Information Criterion) and SBC (Schwartz-Bayesian Criterion), RMSE and MAE value and highly significant MA term with highest adjusted R^2

(68.99%). The assumptions of residuals tested by the Shapiro-Wilk test and Box-Ljung (Q) test indicated that the selected ARIMA model satisfied the assumptions of residuals.

In polynomial models, the cubic model on four year moving average and in fitting of ARIMA models, ARIMA (0, 2, 2) were found suitable to predict the pattern of wheat production in North Gujarat. Rahman *et al.* (2022) [10] also found ARIMA (0, 2, 2) for production of potato in Bangladesh. Thus these results are in agreement with these authors.

In order to check the validity of these forecasted values, they were compared with

the actual values of wheat production during the post sampled forecast period for cubic model on four year moving average data approach and ARIMA (0, 2, 2) model which are presented in Table 6 and graphically depicted in Fig. 2 and Fig. 3, respectively.

Table 5: Fitted ARIMA models for production of wheat in North Gujarat

Model	AIC	SBC	Constant	AR (φ)		MA (θ)		Adj. R^2 (%)	RMSE	MAE	S-W test	B-L test (Q)
				AR (1)	AR (2)	MA (1)	MA (2)					
ARIMA (0, 2, 1)	449.84	452.28	8.40	-	-	1.00**	-	67.43	1857.01	1475.08	0.98	11.97
ARIMA (0, 2, 2)	449.92	453.58	5.62	-	-	1.30**	-0.30	68.99	1815.19	1377.99	0.99	10.68
ARIMA (1, 2, 1)	450.40	454.05	5.57	-0.23	-	0.99**	-	68.27	1828.85	1379.43	0.99	11.60
ARIMA (2, 2, 1)	451.56	456.44	4.66	-0.29	-0.17	0.99**	-	67.90	1928.47	1404.23	0.98	9.16
ARIMA (1, 2, 2)	453.37	458.25	-92.90	0.35	-	1.69**	-0.99**	65.76	1839.47	1406.15	0.94	13.05

*,** indicates significant at 5 per cent and 1 per cent level of significance, respectively

Table 6: Testing of forecast values for four year by using best fitted cubic model on four year moving average data approach and ARIMA models (0, 2, 2) for wheat production (00' MT) of North Gujarat

Year	Observed	Predicted					
		Cubic (4 year)	Deviation	Per cent deviation	ARIMA (0, 2, 2)	Deviation	Per cent deviation
2018-19	9301.84	12057.62	-2755.78	29.62	11354.14	-2052.30	22.06
2019-20	12217.50	12051.24	166.26	1.36	11716.64	500.86	4.10
2020-21	11560.35	11908.57	-348.22	3.01	12084.76	-524.41	4.54
2021-22	11201.63	11616.41	-414.78	3.70	12458.52	-1256.89	11.22
Mean				8.75			8.43

It is observed that the mean per cent deviation between forecasted and actual production based on cubic model of four year moving average data approach and ARIMA (0, 2, 2) models were 8.75 and 8.43 per cent, respectively. This proved that both the models found to be best fit model in validation test for forecasting the production of wheat in North Gujarat.

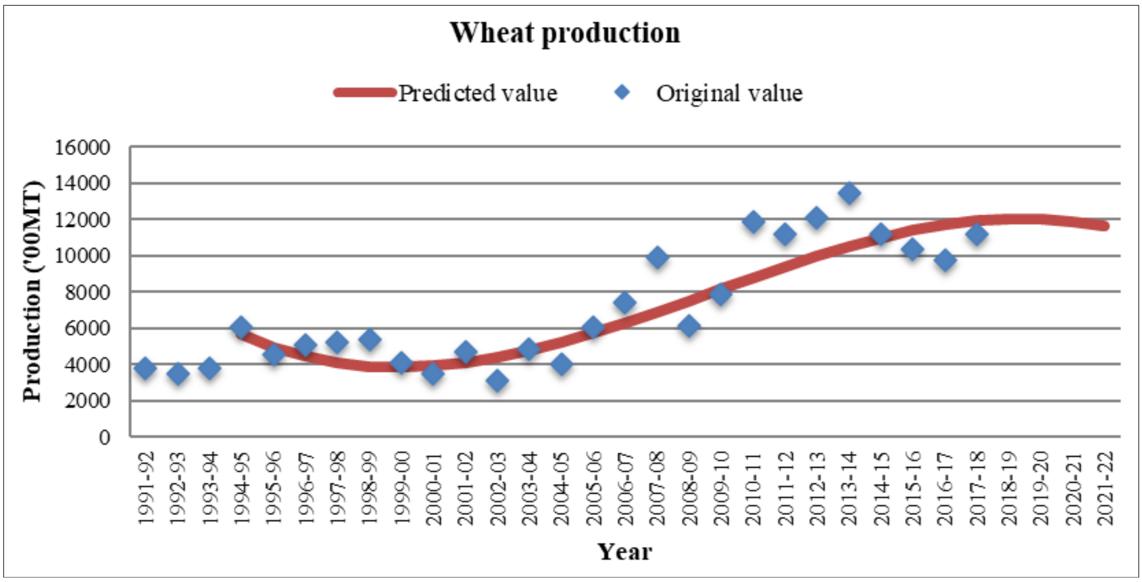

Fig 2: Forecasted wheat production of North Gujarat fitted with cubic model of four year moving average data approach

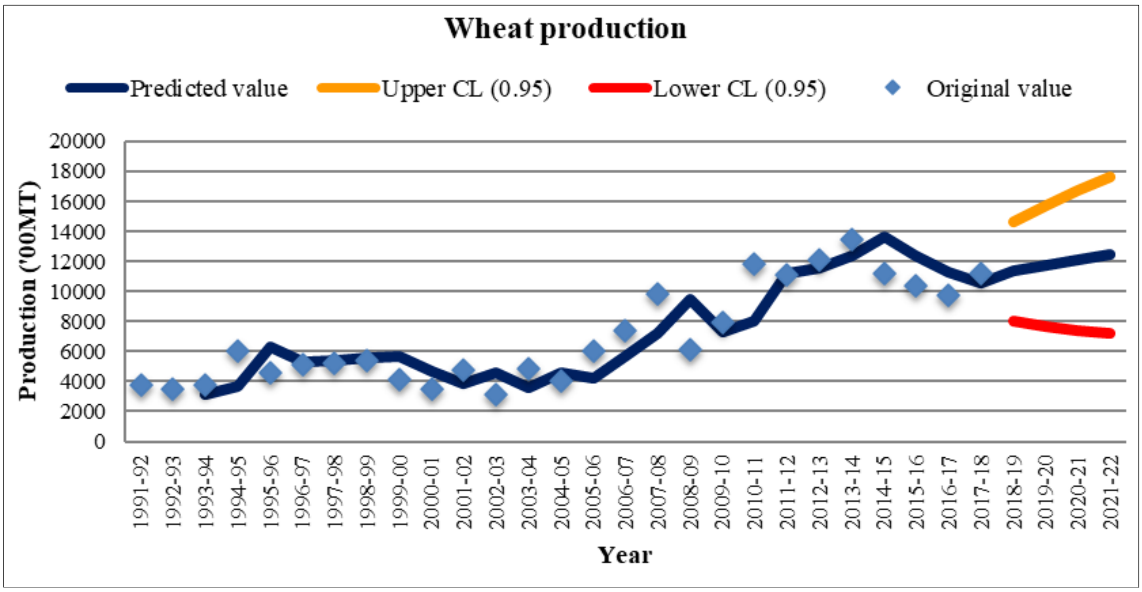

Fig 3: Forecasted wheat production of North Gujarat fitted with ARIMA (0, 2, 2)

Trends on wheat productivity in North Gujarat

Fitting of polynomial models

The results of fitted polynomial models are given in Table 7. The results indicated that in all the polynomial models the coefficient of determination (R^2) was found to be highly significant. The regression constants value of all three linear, quadratic and cubic models in all the data approaches were found to be highly significant. Linear regression coefficients were found to be highly significant in all data approaches of linear and found to be significant for four and five year moving average of cubic models. Quadratic regression coefficient was found significant for four year moving average and highly significant for five year moving average of quadratic model and significant for all moving averages of cubic model. Cubic regression coefficient was significant in four year moving average of cubic model. In original data approach the value of adjusted R^2 was decrease by 1.60 per cent in quadratic regression as compared to linear regression while, in case of cubic regression, it was improved by 4.60 per cent over quadratic model. By taking moving average of five year the improvement in adjusted R^2 was observed to be 25.20, 41.80, 40.30 per cent in case of first, second- and third-degree polynomial models, respectively over original data.

The third degree polynomial models showed comparatively lower values of root mean square error (RMSE) and mean absolute error (MAE). Among these the least RMSE and MAE were observed in case of model based on five years moving averages. The criteria for testing normality (Shapiro-Wilk test) of residuals indicated that, all models had normally distributed residuals. The test of randomness of residuals (Run test) indicated that original data approaches of all linear, quadratic and cubic models were randomly distributed.

Table 7: Fitted linear and non-linear models for productivity of wheat in North Gujarat

Model	Moving average	Regression constant	Regression coefficients			R^2 (%)	Adj. R^2 (%)	RMSE	MAE	S-W test	Run test |Z|
		a	b	c	d						
Linear	Original	1105.22**	49.09**	-	-	34.60**	32.00	525.91	418.73	0.98	1.56
	3 Years	1108.52**	53.78**	-	-	46.00**	43.70	420.03	342.77	0.98	2.85**
	4 Years	1107.69**	56.85**	-	-	51.80**	49.60	379.84	315.83	0.98	3.54**
	5 Years	1085.12**	61.41**	-	-	59.20**	57.20	338.35	289.29	0.95	3.41**
Quadratic	Original	1277.51**	13.45	1.27	-	35.70**	30.40	521.37	424.06	0.98	1.56
	3 Years	1428.91**	-17.41	2.73	-	51.00**	46.50	400.35	308.45	0.98	2.44*
	4 Years	1544.61**	-43.97	4.03*	-	61.70**	58.10	338.46	256.22	0.97	3.13**
	5 Years	1617.96**	-66.46	5.32**	-	74.80**	72.20	266.00	218.76	0.98	2.98**
Cubic	Original	1836.73**	-206.47	20.55	-0.45	42.50**	35.00	493.17	423.87	0.97	1.56
	3 Years	1967.62**	-244.12	24.11*	-0.54	58.80**	52.90	366.97	310.24	0.97	2.44*
	4 Years	2033.28**	-257.09*	24.91*	-0.55*	68.60**	63.90	306.44	258.12	0.96	3.13**
	5 Years	1976.20**	-228.58*	21.86*	-0.45	78.60**	75.30	244.74	215.53	0.96	2.98**

*,** indicates significant at 5 per cent and 1 per cent level of significance, respectively

Fitting of ARIMA models

As the series was found non stationary, the new variable X_t was constructed by taking difference of one (*i.e.*, d = 2) to make the series stationary.

The autocorrelation (ACF) and partial autocorrelation (PACF) coefficients of various orders of X_t were computed to identify the value of p and q. However, on the basis of goodness of fit criteria some models were selected and are given in Table 8.

From the observed model ARIMA (0, 2, 1) had the lowest AIC (Akike Information Criterion), SBC (Schwartz-Bayesian Criterion) value, RMSE, MAE and MA term was highly significant and with highest adjusted R^2 (25.28%).

None of the polynomial as well as ARIMA model was found suitable to explain the trend of wheat productivity in North Gujarat.

Table 8: Fitted ARIMA models for productivity of wheat in North Gujarat

Model	AIC	SBC	Constant	AR (φ)		MA (θ)		Adj. R^2 (%)	RMSE	MAE	S-W test	B-L test (Q)
				AR (1)	AR (2)	MA (1)	MA (2)					
ARIMA (0, 2, 1)	391.59	394.03	-2.27	-	-	0.99**	-	25.28	575.49	421.45	0.95	9.35
ARIMA (0, 2, 2)	392.97	396.63	-3.00	-	-	1.19**	-0.19	24.67	580.80	420.01	0.97	9.29
ARIMA (1, 2, 1)	393.17	396.83	-2.77	-0.13	-	0.99**	-	23.77	581.80	417.20	0.96	9.50
ARIMA (2, 2, 1)	394.47	399.35	-3.23	-0.16	-0.16	0.99**	-	22.56	586.40	433.25	0.96	7.66
ARIMA (1, 2, 2)	395.59	400.47	-2.27	-1.00**	-	0.00	1.00**	18.16	602.08	421.45	0.95	9.49

*,** indicates significant at 5 per cent and 1 per cent level of significance, respectively

Conclusion

Fitting of trend for wheat in North Gujarat

Area

None of the polynomial model on area of wheat in North Gujarat was selected, on original data approach as well as moving average data approaches due to fails to fulfill the assumption of randomness.

ARIMA (0, 2, 1) model was selected on the basis of highly significant MA coefficient, lower value of AIC, SBC, RMSE, MAE, fulfill the assumptions of residuals and success in validation test, which is as under.

$$Y_t = 0.25 + 0.99^{**}a_{t-1} + a_t$$

(Adj. R^2 = 79.89%)

Production

The cubic model on four year moving average data approach was satisfied all the criteria for selection of model. The highly significant positive quadratic term and highly significant negative linear and cubic term was observed in cubic trend for production with highest adjusted R^2 (90.30%) compare to linear and quadratic models. It was also, fulfilled the assumptions of randomness, normality of the residuals and success in validation test, cubic model on four year moving average data approach was considered as suitable. The selected model is,

$$\hat{Y} = 6551.06^{**} - 992.46^{**}t + 103.53^{**}t^2 - 2.20^{**}t^3$$

(Adj. R^2 = 90.30%)

ARIMA (0, 2, 2) model was selected on the basis of highly significant MA coefficient, lower value of AIC, SBC, RMSE, MAE, assumptions of residuals and success in validation test, which is as under.

$$Y_t = 5.62 + 1.30^{**}a_{t-1} – 0.30a_{t-2} + a_t$$

(Adj. R^2 = 68.99%)

Productivity

None of the polynomial model on productivity of pearl millet in North Gujarat was selected on original data approach as well as moving average data approaches due to lack of one or more criteria for selection of models.

In ARIMA family none of the model was found suitable to explain the trend of wheat productivity in North Gujarat due to poor adjusted R^2.

References

1. Anonymous. Agricultural statistics at a glance 2022. Government of India, Ministry of Agriculture and Farmers Welfare, Directorate of Economics and Statistics Division; 2022.
2. Anonymous. District wise area, production and yield of important food and non-food crops in Gujarat State, Directorate of Agriculture, Gujarat state, Gandhinagar; c2022.
3. Box GEP, Jenkins GM, Reinsel GC. Autocorrelation function and spectrum of stationary processes and analysis of seasonal time series. In: Time Series Analysis: Forecasting and Control. 2nd ed. San Francisco: Holden-Day; c1976. p. 21-43.
4. Harris RI. Testing for unit roots using the augmented Dickey-Fuller test: Some issues relating to the size, power and the lag structure of the test. Economics Letters. 1992;38(4):381-386.
5. Liew VKS, Shitan M, Hussain H. Time series modeling and forecasting of Sarawak black pepper price. Jurnal Akademik; c2000. p. 39-55.
6. Ljung GM, Box GE. On a measure of lack of fit in time series models. Biometrika. 1978;65(2):297-303.
7. Makridakis S, Wheelwright SC, Hyndman RJ. Forecasting methods and applications. 3rd ed. John Wiley & Sons; c2003. p. 312-372.
8. Montgomery DC, Peck EA, Vining GG. Introduction to Linear Regression Analysis. John Wiley & Sons; c2003. p. 221-258.
9. Pankratz A. Detecting and treating outliers in dynamic regression models. Biometrika. 1983;80(4):847-854.
10. Rahman MM, Islam MA, Mahboob MG, Mohammad N, Ahmed I. Forecasting of potato production in Bangladesh using ARIMA and mixed model approach. Scholars Journal of Agriculture and Veterinary Sciences. 2022;10:136-145.
11. Rangaswamy R. A text book of Agricultural statistics. New Age International (P) Limited Publishers; c2006. p. 158-210.
12. Sarda C, Prajneshu. Modeling and forecasting country's pesticide consumption data using ARIMA time-series approach. Annals of Agricultural Research. New series. 2002;23(4):719-722.
13. Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). Biometrika. 1965;52(3/4):591-611.
14. Sidney S, Castellan NJ. The permutation test for paired replicates. In: Nonparametric Statistics for the Behavioral Sciences: The Case of One Sample, Two Measures or Paired Replicates. 2nd ed. Singapore: McGraw-Hill Book Co.; c1988. p. 95-101.
15. Thapa R, Devkota S, Subedi S, Jamshidi B. Forecasting area, production and productivity of vegetable crops in Nepal using the Box-Jenkins ARIMA model. Turkish Journal of Agriculture - Food Science and Technology. 2022;10(2):174-181.