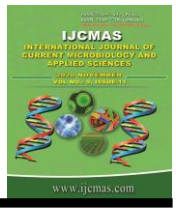
Int.J.Curr.Microbiol.App.Sci (2020) *9*(11): 84-93

International Journal of Current Microbiology and Applied Sciences
***ISSN: 2319-7706* Volume 9 Number 11 (2020)**
Journal homepage: http://www.ijcmas.com



Original Research Article https://doi.org/10.20546/ijcmas.2020.911.009

Statistical Evaluation of Stepwise Regression Method and Autoregressive Integrated Moving Average Method for Forecasting of Groundnut (*Arachis hypogaea* L.) Productivity in Junagadh District of Gujarat

K. Sathees Kumar[1*] and Mayur Shitap[2]

[1]Department of Agricultural Statistics, Bidhan Chandra Krishi Viswavidyalaya, Mohanpur, Nadia741 252, West Bengal, India
[2]Department of Agricultural Statistics, Junagadh Agricultural University, Junagadh-362 001, Gujarat, India

**Corresponding author*

Keywords

ARIMA, Stepwise regression, Weather indices

Article Info

Accepted:
04 October 2020
Available Online:
10 November 2020

A B S T R A C T

In India, the productivity of various crops is unstable mainly due to climatic factors, price volatility and resource availability. The pre-harvest forecasting of the crop productivity is a major priority to know about the market demand of the crops. The present study focused the ability of pre-harvest forecasting performance of stepwise regression method and the ARIMA method. In stepwise regression method, two approaches were developed namely (1) using week-wise original weather variable and (2) weather indices using correlation coefficient as weight. Among the two approaches studied, the correlation coefficient as a weighted approach had more expedient to pre-harvest forecasting of groundnut. Eventually, after the good interrogation, stepwise regression method had better puissance than the ARIMA method for forecasting the groundnut productivity in the Junagadh district of Gujarat.

Introduction

Oilseeds play a significant role in Indian agriculture, as a second major crop (next to cereals) possessing 11 % of the total cultivated area and devoted 9 % production out of the total agricultural production. India oozed at the period of late 1990s as one of the world's largest edible oils importers and consumer of oilseeds and their products. The cash crops share about 6.87 % GDP and oilseed crops share about 1.5 % to agricultural GDP (Anon., 2017). Groundnut (*Arachis hypogaea* L.) is most imperatives oilseed crop among others. Our country is one of the world's premier producers having the largest area under groundnut in the globe.

For ensuring lucrative prices to the growers, not only requires improved crop technology for increasing the production but also need long term and short-term policy decision for yearly export and import. This naturally demands a believable and valid pre-harvest forecast of the production of crops. Multifarious methods are available for

forecasting the crop productivity were stepwise regression method and ARIMA method used for profuse investigation and the forecasting the crop productivity. Earlier studies investigated to compare the forecasting performance of stepwise regression and ARIMA for forecasting of mango and banana productivity (Rathod and Mishra, 2018). When ARIMA applies better from stepwise regression, the current productivity is more dependent on the previous productivity than the weather parameters and *vice versa*. The present study emulated the above investigation to compare the ability of stepwise regression and ARIMA model for groundnut productivity forecasting in the Junagadh district of Gujarat.

Materials and Methods

Data description

Year wise data of groundnut productivity and historical weather data of groundnut growing season (24th Standard Meteorological Week to 37th Standard Meteorological Week) including maximum temperature (MAX T), Minimum temperature (MIN T), weekly total rainfall (RF), morning relative humidity (RH_1), afternoon relative humidity (RH_2) of Junagadh district of Gujarat for the years 1985 to 2015 were collected from Director of Agriculture and Agro-meteorology Cell, College of Agriculture, Junagadh Agricultural University, Junagadh, respectively.

The data from 1985 to 2009 used for analysis and the data from 2010 to 2015 were used for model validation by percent deviation. To seek the possibility of early forecasts before 6, 4, and 2 weeks of the harvest of groundnut crop, three models were fitted using generated weather variables for the period of 10 (24th SMW to 33rd SMW), 12 (24th to 35th SMW), and 14 (24th to 37th SMW) weeks crop periods.

Stepwise regression method

In stepwise regression method, two approaches were developed based on using the explanatory variables.

Week-wise approach using original weather variables.

Weather indices using correlation as weight

Year-wise groundnut productivity data and weekly weather data were used for the above approaches and utilized for stepwise regression (Montgomery *et al.*, 2003) to avoid the multicollinearity consequences.

Week-wise approach using original weather variables

In this approach, the weekly average data as per the original scale were used as the explanatory variables and time period was also considered as the explanatory variable (Draper and Smith, 1981).

The mathematical expression of this model is,

$$Y = A_0 + \sum_{i=1}^{p}\sum_{j=1}^{w} a_{ij}X_{ij} + bT + e$$

Where,

Y = Groundnut productivity of the Junagadh district (Kg/ha)
A_0 = Constant
X_{ij} = Observed value of ith weather variable in jth week, i = 1, 2, … …, p and j = 1, 2, … …,w (p=5, w = 10, 12 and 14)
T = Year number includes correcting for long term upward or downward trend in productivity
a_{ij} and b are partial regression coefficients associated with each X_{ij} and time trend respectively.
e= Error term

Weather indices using correlation as weight

This methodology proposed by the Indian Agricultural Statistical Research Institute (Agrawal *et al*., 1980), New Delhi, modified the IASRI model was used to convey how the crop productivity was affected by the weather variables as a function of the correlation coefficient between respective weather variables and the crop productivity. In this approach, two types of weather indices developed for getting the explanatory variables (First order weather indices and second order weather indices). First order weather indices (Z_{ij}) were determined by each weather variables and second order weather indices ($Z_{ii'j}$) were determined by the interaction of the weather variables (multiplication of two possible weekly weather variables).For first order weather indices, two possible indices were developed. One is unweighted weather indices (simple aggregation of weather variable which means the power of the correlation coefficient is zero) another one is weighted weather indices (power of correlation coefficient is one). Similarly, for second order indices, two kind indices (unweighted and weighted) were made by weekly products of weather variables. The time period was also considered as the explanatory variable.

$$Y = A_0 + \sum_{i=1}^{p}\sum_{j=0}^{1} a_{ij} z_{ij} + \sum_{i\neq i'=1}^{p}\sum_{j=0}^{1} a_{ii'j} z_{ii'j} + bT + e$$

Where,

Y = Groundnut productivity of the Junagadh district (Kg/ha)
A_0 = constant
T = Year number included correcting for long term upward or downward trend in productivity
a_{ij}, $a_{ii'j}$, and b are estimated partial regression coefficients associated with z_{ij}, $z_{ii'j}$, and time trend respectively.
e = Error term
p = Number of weather variables

$Z_{ij} = \sum_{j=0}^{1}\sum_{w=1}^{m} r_{iw}^{j} x_{iw}$ and $Z_{ii'j} = \sum_{j=0}^{1}\sum_{w=1}^{m} r_{ii'w}^{j} x_{iw} x_{i'w}$

Where,

w = weeks (w = 1, 2, …m = 10, 12, and 14)
m = No. of weeks up to the time of the forecast
X_{iw} = value of i^{th} weather variable at w^{th} week of groundnut growing season
Z_{ij} and $Z_{ii'j}$ are generated first order and second-order variables
r_{iw}^{j}= correlation coefficient between the groundnut productivity and i^{th} weather variable at w^{th} period
$r_{ii'w}^{j}$= correlation coefficient between the groundnut productivity and the multiplication of i^{th} and i'^{th} weather variables at w^{th} period

Autoregressive Integrated Moving Average Model (ARIMA)

An ARIMA model, time-series data of groundnut productivity data used for prediction purposes. Box-Jenkins time-series models *i.e.* ARIMA is known as "Univariate Box-Jenkins technique" (Box and Jenkins, 1976). This method is contrary to stepwise regression method which explains the relation between successive observation and previous observations of the successive observation. Univariate Box-Jenkins ARIMA (p, d, q) revealed as follows,

$$\emptyset(B)(1-B)^{d}Y_t = \theta(B)\varepsilon_t$$

Where,

$$\emptyset(B) = 1 - \emptyset_1 B - \emptyset_2 B^2 \ldots - \emptyset_p B^p$$

(Autoregressive parameter)

$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 \ldots - \theta_p B^q$ (Moving average parameter)

ε_t- Error term, d- degree of differencing to make a series of stationary, B – Backshift operator, *i.e.*$B^a Y_t = Y_{t-a}$

The ARIMA model has three steps, *viz.* Identification, estimation and diagnostic checking. Identification of d is necessary to transform a nonstationary time series in to stationary. Testing of stationarity was employed by estimates of mean, autocovariance and autocorrelation of the data. Identification of p and q were decided by PACF's correlogram and ACF's correlogram respectively.

The estimation of parameters was calculated by the maximum likelihood technique. In diagnostic checking, the goodness of fit for the ARIMA model (Sarda and Prajneshu, 2002**)** were checked by Akaike's Information Criterion (AIC) and Schwartz-Bayesian Criterion (SBC) and testing of error independence were checked by Chi-square test (Ljung and Box, 1978). If the model is not satisfied with the above criteria, the above three stages are repeated until getting the satisfactory ARIMA model for forecasting.

Comparison of MLR and ARIMA models

Coefficient of determination (R^2)

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

Where R^2 indicates the value of variation in the dependent variable accounted for due to the model.

Y_i– Observed productivity
$Y_i\,\hat{}$–Estimated productivity

Adjusted coefficient of determination ( )

$$\overline{R}^2 = 1 - \frac{(n-1)(1-R^2)}{(n-k-1)}$$

Root Mean Square Error (RMSE)

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

Mean Absolute Error (MAE)

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

The fitted models, which had higher values of R^2 and with lower values of RMSE and MAE, were considered to be better.

Test of forecasting values by using percent deviation

Forecast values for remaining years by using selected models were tested based on percentages of forecasting error. Percentages of forecasting error were calculated as under

$$\%\ of\ PD = \left(\frac{Y - \hat{Y}}{Y}\right) \times 100$$

Where PD is the percent deviation

Y is the observed value of the remaining years

$\hat{Y}$ is the forecast value of the remaining years

Results and Discussion

Week-wise approach using original weather variables

In a fitted model using data for 10, 12 and 14 weeks crop period (Table 1), the set of explanatory variables entered in the equation consisted *viz.*, X_{59} (afternoon relative

humidity of 9th week), X_{44} (morning relative humidity of 4th week), X_{26} (minimum temperature of 6th week) and X_{58} (afternoon relative humidity of 8th week). These variables explained about 58.10% of the variation in productivity of groundnut. The results indicated that the partial regression coefficients of all included variables were positive and significant. The forecasted productivity of groundnut productivity for the fitted equation (Table 2) showed 8.66 to 53.50 percent deviation from observed productivity.

Correlation coefficient as weight using generated weather variables

The variables recorded in the model (Table 1) for 10 weeks data were Z_{141} (weight of correlation coefficient to the product of maximum temperature and morning relative humidity), Z_{121} (weight of correlation coefficient to the product of maximum temperature and minimum temperature), Z_{51} (weight of correlation coefficient of afternoon relative humidity), Z_{131} (weight of correlation coefficient to the product of maximum temperature and weekly total rainfall) and Z_{31} (weight of correlation coefficient of rainfall) explained about 86.30% of the variation in the groundnut productivity.

The results referred that the partial regression coefficients of Z_{141}, Z_{121}, Z_{51}, and Z_{131} were positive and significant, whereas, the partial regression coefficient of Z_{31} was negative and significant. The deviations of actual productivity and forecasted productivity (Table 2) ranged between 12.95 to 15.72 %. The model (Table 1) for 12 weeks data comprised of only were Z_{51} (weight of correlation coefficient of afternoon relative humidity) explaining about 52.60% of the variation in productivity of groundnut crop. The partial regression coefficient of Z_{51} was positive and significant with the deviations of actual productivity and forecasted productivity (Table 2) ranging between 18.96 to 64.38 %. The variable entered in the model (Table 1) for 14 weeks data were Z_{131} (weight of correlation coefficient to the product of maximum temperature and weekly total rainfall), Z_{241} (weight of correlation coefficient to the product of minimum temperature and morning relative humidity), Z_{31} (weight of correlation coefficient of weekly total rainfall) and Z_{11} (weight of correlation coefficient of maximum temperature) explained about 88.00% of the variation in productivity of groundnut crop.

The results mentioned that the partial regression coefficients of Z_{131}, Z_{241}, and Z_{11} were positive and significant, whereas the partial regression coefficient of Z_{31} was negative and significant. The deviations of actual productivity from forecasted productivity (Table 2) ranged from 11.82 to 16.87 %. Looking to higher (88.00%), lower deviations (11.82 to 16.87 %), RMSE (144.64) and MAE (116.51) in prediction, the model of 14 weeks could be considered as pre-harvest forecast model which can predict the productivity at 2 weeks before harvest with R^2 value 90.60%. The model for 10 weeks data has low deviations (12.95 to 15.72 %) as compared to the model of 14 weeks which can predict at 6 weeks before harvest. Also, there is no much difference in (86.30%), RMSE (149.68) and MAE (119.41), model with10 weeks data as compared to a model of 14 weeks ($\overline{R}^2$=88.00%, RMSE=144.64 and MAE=116.51).

In stepwise regression models, among the two approaches using correlation coefficient as weight approach gave the highest and lower RMSE, MAE and lower deviations than week-wise using original weather variable. So, the use of correlation coefficient as weight approach gave better performance for

forecasting the groundnut productivity of Junagadh district. The present study was identical to the result of groundnut yield forecasting in Kolhapur, Maharashtra (Dhekale *et al.*, 2014).

Table.1 Fitted Step-wise regression models

APPROACHES	WEEKS	STEPWISE REGRESSION EQUATIONS	R^2 (%)	$\bar{R}^2$ (%)	RMSE	MAE
Correlation coefficient as weight using generated weather variables	10	$Y=-3536.76+1.24 *Z_{141}+5.95* Z_{121}+15.27* Z_{51}+1.37* Z_{131}-41.73* Z_{31}$	89.90	86.30	149.68	119.41
	12	$Y=-2240.21+31.31* Z_{51}$	55.10	52.60	316.21	245.43
	14	$Y=-1301.31+1.08* Z_{131}+0.65* Z_{241}-31.49* Z_{31}+100.79* Z_{11}$	90.60	88.00	144.64	116.51
Week-wise approach using generated weather variables	10	$Y=10910.35+18.12* X_{59}+43.97* X_{44}+231.71* X_{26}+15.95* X_{58}$	66.90	58.10	271.32	196.02
	12					
	14					

Table.2 Forecasted productivity based on Step-wise regression and it's percent deviations from observed productivity

Year	Observed Productivity (Kg/ha)	Correlation coefficient as weight using generated weather variables			Week-wise approach using generated weather variables
		10 weeks	12 weeks	14 weeks	10, 12 and 14 weeks
2010	2162	1882.03 (12.95)	1752.00 (18.96)	1888.07 (12.67)	1842.74 (14.77)
2011	1774	1524.27 (14.08)	1370.94 (22.72)	1564.23 (11.82)	1620.40 (8.66)
2013	3590	3025.65 (15.72)	1278.76 (64.38)	2984.28 (16.87)	1669.35 (53.50)
2014	3123	2693.49 (13.75)	1636.89 (47.59)	2739.19 (12.29)	1673.62 (46.41)

Figures in parentheses indicate percent deviation of forecasted productivity from observed productivity (productivity of 2012 and 2015 are outliers)

Table.3 Fitted ARIMA models

ARIMA	C	AR(Φ_1)	MA(θ_1)	MAθ_2)	MA(θ_3)	R^2 (%)	$\bar{R}^2$ (%)	RMSE	MAE	AIC	BIC
(0,1,1)	71.77 (105.45)	-	0.99 (7.71)	-	-	69.30	58.80	544.92	430.38	13.04	13.10
(0,1,2)	4.99 (65.15)	-	1.40 (2.65)	-0.42 (0.95)	-	64.10	59.80	525.65	387.67	13.06	13.15
(0,1,3)	93.11 (94.69)	-	1.29 (2.63)	-0.37 (0.91)	0.05 (0.43)	63.30	56.50	549.58	402.95	13.35	13.39
(1,1,1)	34.91 (67.55)	-0.40 (0.28)	1.00 (14.59)	-	-	65.20	61.10	517.19	398.62	13.04	13.12
(1,1,2)	32.30 (80.82)	-0.75 (0.44)	0.53 (32.57)	0.47 (15.56)	-	65.70	59.30	531.42	391.16	13.27	13.33
(1,1,3)	46.72 (96.84)	-0.75 (1.03)	0.60 (40.24)	0.49 (15.23)	-0.09 (2.97)	66.10	57.10	548.53	392.06	13.49	13.54

Table.4 Forecasted productivity based on ARIMA and it's percent deviations from observed productivity

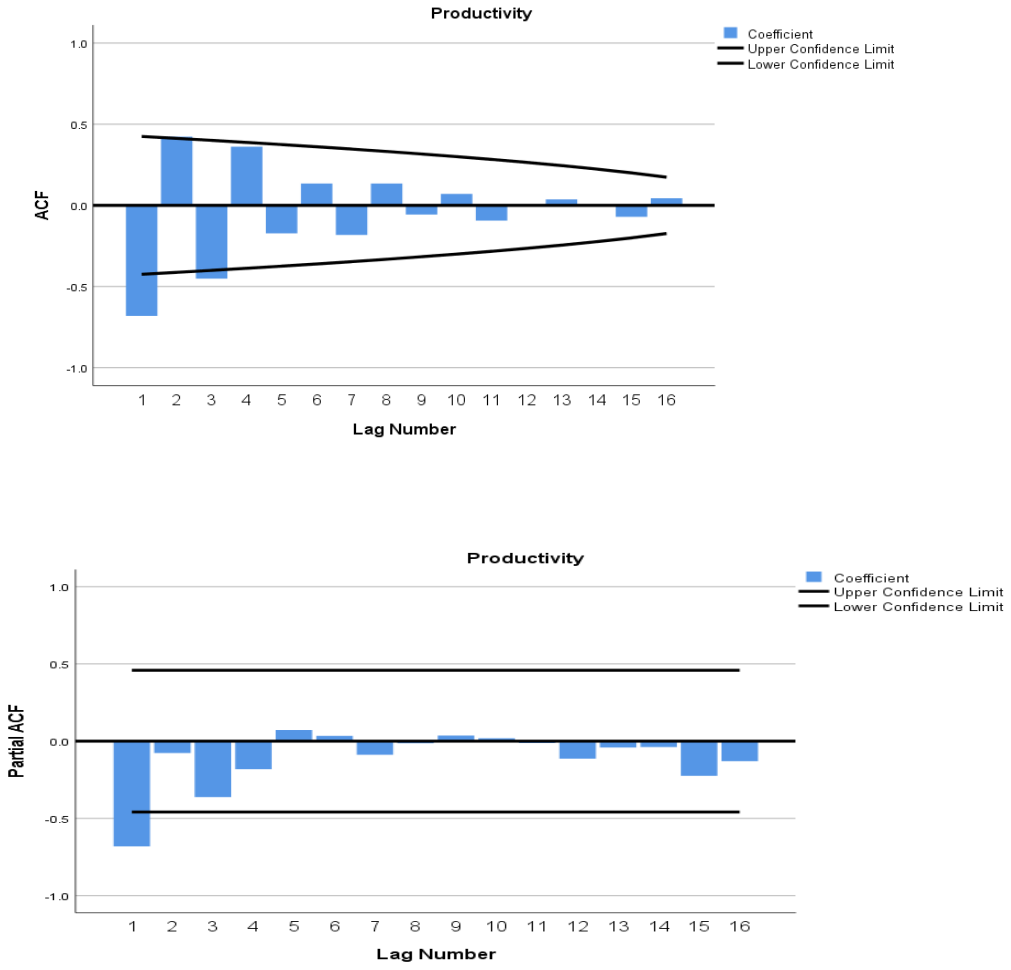
Year	Observed Productivity (Kg/ha)	ARIMA model					
		(0,1,1)	**(0,1,2)**	**(0,1,3)**	**(1,1,1)**	**(1,1,2)**	**(1,1,3)**
2010	2162	1457 (32.61)	2077 (3.93)	1796 (16.93)	1941 (10.22)	1809 (16.33)	1796 (16.93)
2011	1774	1414 (20.29)	1795 (-1.18)	1460 (17.70)	1525 (14.04)	1456 (17.93)	1460 (17.70)
2013	3590	1365 (61.98)	1828 (49.08)	1628 (54.65)	1692 (52.87)	1714 (52.26)	1628 (54.65)
2014	3123	1312 (57.99)	1862 (40.38)	1470 (52.93)	1622 (48.06)	1511 (51.62)	1470 (52.93)

Figures in parentheses indicate percent deviation of forecasted productivity from observed productivity (productivity of 2012 and 2015 are outliers)

Table.5 Comparison between the fitted Step-wise regression models and ARIMA model

Methods	R^2(%)	$\bar{R}^2$ (%)	RMSE	MAE
Regression model using the correlation coefficient as the weight with generated weather variables(model of 10 weeks)	89.90	86.30	149.68	119.41
Regression model using the week-wise approach with original weather variables(model of 10 weeks)	66.90	58.10	271.32	196.02
ARIMA (1, 1, 1)	65.20	61.10	517.19	398.62

Fig.1 ACF and PACF of the groundnut productivity of Junagadh district



Autoregressive Integrated Moving Average Model

Time series data of groundnut productivity data was found as a non-stationary series. The new generated variable X_t was made by taking the difference of one (*i.e.* d=1) to make the series from non-stationary to stationary. The new series X_t (d=1) was found to be stationary for groundnut productivity data. Partial autocorrelation function (PACF) and autocorrelation function (ACF) of various orders of X_t was computed to identify the values of p and q respectively. The ACF (Υ_k) of the transformed variable was tail-off towards zero with cut off at third spike and PACF (φ_{kk}) of the transformed variable tails off towards zero with cut off at first spike (Fig 1). This suggested that the algebraic family of ARIMA on p=0, 1 d=1, and q=0, 1, 2, 3 can be used. The results are given in Table 3 and forecasted productivity presented in Table 4. The assumptions of residuals *i.e.* independence of residuals were tested by Box-Ljung (Q) test indicated that all ARIMA models satisfied the assumption of residuals. Among the fitted models, ARIMA (1, 1, 1) model gave highest $\overline{R}^2$ (61.10%) and lowest RMSE (517.19), Whereas the lowest MAE (387.67) was observed in ARIMA (0, 1, 2) and lowest BIC (13.10) was observed in ARIMA (0,1,1) but ARIMA (1,1,1) which

gave highest $\overline{R}^2$ and lowest RMSE had slightly greater MAE (398.62) and BIC (13.12) as compared to other models. The deviations of forecasted productivity from observed productivity was ranged from 10.22 to 52.87%. The ARIMA (1, 1, 1) was found to be best fitted model for forecasting the groundnut productivity in Junagadh district. The finding was contrary to Rajarathinam and Dixit (2007), they studied the groundnut yield trends in long-term fertilizer experiment at Junagadh.

Comparison between the fitted Stepwise regression models and ARIMA model

The comparison was made among these selected models and the fitted ARIMA models based on the coefficient determination (R^2), Adjusted coefficient of determination ($\overline{R}^2$), Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). It is seen from the Table 5 that both the regression models *viz.*, model using correlation coefficient as the weight with generated weather variables (10 weeks), and model using the week-wise approach with original weather variables (10 weeks) gave higher R^2 (89.90 % and 66.90 %, respectively), adjusted R^2 (86.30 % and 58.10 %, respectively) with lower RMSE (149.68 and 271.32, respectively), MAE (119.41 and 196.02, respectively) having lower per cent deviation between forecasted and observed productivity (12.95 to 15.72 % and 8.66 to 53.50 %, respectively) as compared to ARIMA (1, 1, 1) using 25 years data (R^2 = 65.20 %, $\overline{R}^2$ = 61.10 %, RMSE = 517.19, MAE = 398.62 and percent deviations = 10.22 to 52.87 %). Stepwise regression had more potential than ARIMA for forecasting the groundnut productivity in Junagadh. Similar results were found, when forecasted the mango area and production in Karnataka region (Rathod and Mishra, 2017).

Based on the comparisons, the stepwise regression models performed better than the ARIMA models for forecasting the groundnut productivity in the Junagadh district of Gujarat. The study further identified that groundnut productivity was more influenced by weather parameters than previous records of groundnut productivity.

References

Agrawal, R., R. C. Jain, M. P. Jha and Singh, D. 1980. Forecasting of rice yield using climatic variables. Indian Journal of Agricultural Science, 50(9):680-684.

Anonymous. 2017. Division of Agricultural Economics, Indian Agricultural Research Institute, New Delhi. Available at http://www.iari.res.in>accessed 10 august, 2019.

Box, G. E. P. and Jenkins, G. M. 1976. Time Series Analysis Forecasting and Control, Second Edition, Holden Day. PP: 88-122.

Draper, N.R. and Smith, H. 1981. Applied regression analysis, second edition, John Wiley and sons, New York.

Ljung, G. M and Box, G. E. P. 1978. On a Measure of Lack of fit in Time Series Models. Biometrika, 65: 297-303.

Dhekale, B. S., M. S. Sheraz, T. P. Dalvi, and Sawant, P. K. 2014. Forecast Models for Groundnut using Meteorological Variables in Kolhapur, Maharashtra. Journal of Agrometerology, 16 (2): 238-239.

Montgomery, D. C., E. A. Peck, and Vining G. G. 2003. Introduction to Linear Regression Analysis. John Wiley & sons, Inc, PP: 221-258.

Sarda, C and Prajneshu. 2002. Modeling and Forecasting Country's Pesticide/ Consumption Data using ARIMA Time Series Approach. Annals of Agricultural Research, 23(4): 719-722.

Rajarathinam, A and Dixit, S. K. 2007. Fitting of Groundnut Yield Trends in Long-Term Fertilizer Experiment- A Time-Series Model Approach. Crop Research, 34: 92-96.

Rathod, S and Mishra, G.C. 2017. Weather Based Modeling for Forecasting Area and Production of Mango in Karnataka. International Journal of Agriculture, Environment and Biotechnology, 10(1):149-162.

Rathod, S and Mishra, G.C. 2018. Statistical Models for Forecasting Mango and Banana Yield of Karnataka, India. Journal of Agricultural Science and Technology, 20: 803-816.

How to cite this article:

Sathees Kumar, K. and Mayur Shitap. 2020. Statistical Evaluation of Stepwise Regression Method and Autoregressive Integrated Moving Average Method for Forecasting of Groundnut (*Arachis hypogaea* L.) Productivity in Junagadh District of Gujarat. *Int.J.Curr.Microbiol.App.Sci.* 9(11): 84-93. doi: https://doi.org/10.20546/ijcmas.2020.911.009