Nonlinear Modeling of Area and Production of Sugarcane in Tamil Nadu, India

P. Dinesh Kumar*, Bishvajit Bakshi and V. Manjunath

Department of Agricultural Statistics, Applied Mathematics and Computer Sciences, UAS, GKVK, Bengaluru-65, Karnataka, India
*Corresponding author

A B S T R A C T

The present investigation was carried out to model the trend of area and production of sugarcane in Tamil Nadu. It was obtained by using the secondary data of area and production over a period of 30 years (1984-85 to 2014-15). For this purpose, Different nonlinear models such as Logistic, Gompertz, Rational, Gaussian, Weibull, Hoerl and Sinusoidal models were employed. Levenberg-Marquardt technique was used to obtain the estimates of the unknown parameters of the nonlinear regression models. To select a best fitted model for the area and production of sugarcane in Tamil Nadu, the model adequacy statistics such R^2, RMSE, MAE and residual assumption tests such as Runs test, Shapiro-Wilks test and Durbin-Watson test were carried out. For area of sugarcane, it was found that Logistic model had the lowest Root Mean Square Error (27.770), Mean Absolute Error (18.737) and the highest R^2 value (74.7 per cent). Hence, Logistic model is the most suitable among the fitted nonlinear model which can be used for further trend analysis on the area under sugarcane. For production of sugarcane, Gaussian model had the lowest Root Mean Square Error (2.604), Mean Absolute Error (2.760) and the highest R^2 value (78.2 per cent). Hence, Gaussian model is the most suitable among the fitted nonlinear model which can be used for further trend analysis on the production of sugarcane.

Introduction

Sugarcane, a traditional crop of India plays an important role in agricultural and industrial economy of the country. It is cultivated in most of the states and though it covers an insignificant share in gross cropped area of the country, its share in the country's economic growth has become significant. The crop is grown in more than 120 countries, of which, Brazil (736 million tonnes), India (352 million tonnes) and China (126 million tonnes) are the top three countries in production (Anon, 2015). In 2015, Uttar Pradesh recorded the highest area of sugarcane of about 42.25 per cent, followed by Maharashtra (20.33%), Karnataka (9.47%), Tamil Nadu (5.19%), Gujarat (4.11%) and Andhra Pradesh (2.74%) contributing about 84 per cent of the total area in India. Currently in Tamil Nadu, 0.263 million hectares are under cane cultivation and this is increasing annually due to the increased consumption of sugar and also the growing demand from mills for sugar cane as a raw material. Because of its diversified uses in different industries, this crop is considered as ''Karpagavirucham'' and in modern terminology as ''wonder cane'' (Mohan *et al.,*

2007). From the above justified facts, it is evident that there is a considerable scope to study the trend in area and production of sugarcane crop in Tamil Nadu.

Materials and Methods

The present study is conducted with the overall objective of estimating suitable regression model that explains the trend of area and production of sugarcane in Tamil Nadu. For this study, A secondary data of area, production and productivity of sugarcane in Tamil Nadu for the period of 30 years from 1985 to 2014 were collected from the Department of Economics and Statistics, Government of Tamil Nadu.

Non-linear regression models

Statistical modelling essentially consists in constructing a model, represented by a set of equations to describe the input-output relationship among the variables of interest. From a realistic point of view, such a relationship among variables in agriculture and biological sciences is '*nonlinear*' in nature. In such a model, a unit increase in the value of independent variable(s) may not result in an equivalent unit increase in the dependent variable. A nonlinear regression model is one in which at least one of the parameters appears nonlinearly. A nonlinear model, which can be transformed into linear model by some transformation is called '*intrinsically linear*', else it is called as '*intrinsically nonlinear*'. Mathematically, in nonlinear models at least one of the derivatives of the expectation function with respect to at least one parameter is a function of parameter(s). The model is a nonlinear regression model as the derivatives of *Y* with respect to *a* and *b* are both functions of *a* and / or *b*. Like in linear regression, parameters in a nonlinear model can also be estimated by the method of least squares. However, due to the

difficulty in the procedure of computation, the common practice is to work with the log transformed model.

$$Y = ab^x + e$$

The log transformation is valid only when error term '*e*' in the above equation is multiplicative in nature. Thereafter, method of least square is used to estimate the unknown parameters. Furthermore, R^2 value is calculated to measure the goodness of fit of the model.

The log transformed procedure suffers from some important drawbacks.

Original structure of the error term got disturbed due to transformation.

R^2 values computed, assess the goodness of fit of the transformed model and not of the original nonlinear model.

Proceeding further to carryout residual analysis for the residuals generated by the transformed model, will result in erroneous conclusion.

As a remedy to these pitfalls, nonlinear regression procedures are already developed in literature which necessitates computer intensive tools to find solution for the parameters (Venugopalan and Shamasundaran, 2003). The following nonlinear models are considered in the present investigation.

Where *Y* is the area/production during the time *X; A, B, C* and *D* are the parameters, and '*e*' is the error term. The parameter '*C*' is the intrinsic growth rate and the parameter '*A*' represents the carrying capacity for each model. Symbol '*B*' represents different functions of the initial value *Y*(0) and '*B*' is

the added parameter. In addition to the above nonlinear models some other nonlinear models also are employed as per the data need.

To obtain estimates of the unknown parameters of a nonlinear regression model, Levenberg-Marquardt technique was used. In this method, the following steps are carried out.

Step I: Starting with a good initial guess of the unknown parameters, a sequence of θ's which hopefully converge to θ is computed.

Step II: Error sum of squares or objective function expressed as

$$S(\theta) = \sum_{x=1}^{n} [Y_x - F_i(\theta)]^2$$

is minimized with respect to the current value of θ. The new estimates are obtained.

Step III: By feeding the recently obtained estimates as the initial guess for the next iteration, objective function $S(\theta)$ is minimized again to obtain fresh estimates. This procedure is continued till the successive iteration yielded parameter estimate values are close to each other.

Choice of starting values of the parameters for various models

All the iterative procedures require initial values θ_{r0} ($r = 1, 2, 3…, k$) of the parameter θ_r. The choice of good initial values can spell the difference between success and failure in locating the fitted value or between rapid and slow convergence to the solution. Also, if multiple minima exist in addition to absolute minimum, poor starting values may result in convergence to an unwanted stationary point of the sum of squares surface. This unwanted point may have parameter values which are physically impossible or which does not provide the true minimum value of $S(\theta)$. There are number of ways to determine initial parameter values for nonlinear models. The

most obvious method for making the initial guesses is by the use of prior information. Estimates calculated from previous experiments, known values from similar systems, values computed from theoretical considerations: all these form ideal initial guesses. In this study the Curve expert Ver.1.3 software package is used to estimate the initial values.

Model adequacy checking

To test the goodness of fit of the fitted polynomial model, the co-efficient of determination R^2 defined as the proportion of total variation in the response variable (time) being explained by the fitted model is widely used.

$$R^2 = 1 - \frac{\sum\limits_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{\sum\limits_{i=1}^{n} (Y_i - \bar{Y})^2}, \qquad 0 \le R^2 \le 1$$

To test the overall significance of the model, the F test is used.

$$F = \frac{\left(\dfrac{R^2}{k}\right)}{\left(\dfrac{1-R^2}{n-k-1}\right)}$$

Which follows F distribution with k (number of parameter in the model), (n-k-1) degrees of freedom.

Adjusted R^2 is a modification of R^2 that adjusts for the number of explanatory terms in a model. Unlike R^2, the adjusted R^2 increases only if the new term improves the model more than would be expected by chance. The adjusted R^2 can be negative and will always be less than or equal to R^2. The adjusted R^2 is defined as

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

Where,

'k' is the number of parameters in the equation
'n' is the is the total number of observations.

In addition to the above, two more reliability statistics viz., Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) are generally utilized to measure the adequacy of the fitted model and it can be computed as follows:

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

$$MAE = \frac{\sum\limits_{i=1}^{n}\left|Y_i-\hat{Y}_i\right|}{n}$$

and

The lower the values of these statistics, the better are the fitted model.

Assumptions of error term

An important assumption of nonlinear regression is that the residual 'ε', or the dependent variable 'Y' follows normal distribution.

This assumption is required for test of hypothesis about the regression coefficients. This assumption was verified using,

Shapiro-Wilk test was used to test for normality. The test statistic value of 'W' ranges from 0 to 1. When $W = 1$ the given data are perfectly normal in distribution (Shapiro, *et al.,* 1968).

When 'W' is significantly lesser than 1, the assumption of normality is not met. The test statistic is

$$W = \frac{\left(\sum\limits_{i=1}^{n}a_i\,x_{(i)}\right)^2}{\sum\limits_{i=1}^{n}(x-\bar{x})^2}$$

Where, $x_{(i)}$ is the ith order statistic, i.e., the ith smallest number in the sample; $\bar{x}$ is the sample mean; The constants a_i are given by

$$(a_1,a_2,...,a_n) = \frac{m^T V^{-1}}{\sqrt{(m^T V^{-1} V^{-1} m)}}$$

Where,

$m^T = (m_1,m_2,...,m_n)^T$ and $m_1,m_2,...,m_n$ are the expected values of the order statistics of independent and identically-distributed random variables sampled from the standard normal distribution, and V is the covariance matrix of those order statistics. Then values a_i, coefficients are tabulated by Shapiro and Wilk (1965).

Durbin-Watson test is used to test the presence or absence of autocorrelation in residuals. Durbin-Watson is the ratio of the distance between the errors to their overall variance. The test statistic is

$$d = \frac{\sum\limits_{i=2}^{n}(e_i - e_{i-1})^2}{\sum\limits_{i=1}^{n}e_i^2} = 2(1-\hat{\rho})$$

Where $e_i = y_i - \hat{y}_i$ and y_i and $\hat{y}_i$ are, respectively, the observed and predicted values of the response variable for individual i. Thus, DW is equal to 2 minus two times the correlation of e_t and e_{t-1}.

Durbin-Watson is used both as diagnostic for autocorrelation and as estimate of ρ. DW statistic is a correlation and thus depends on values of independent variables as $-1 \le \rho \le +1$ thus $0 \le DW \le 4$.

The runs test (Bradley, 1968) was used to decide if a data set is from a random process. The test statistics is

$$z = \frac{r - \mu_r}{\sigma_r} \sim SND(0,1)$$

Where, Mean $\mu_r = \frac{2n_1 n_2}{n_1 + n_2} + 1$,

Standard deviation

$$(\sigma_r) = \sqrt{\frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}}$$

With n_1 and n_2 denoting the number of positive and negative values in the series respectively.

The runs test rejects the null hypothesis, if

$$|Z| > Z_{1 - \frac{\alpha}{2}}.$$

Results and Discussion

Three parameter mechanistic growth models such as Logistic, Gompertz, Gaussian and Hoerl models, and four parameter mechanistic growth models such as Ration function, Weibull and Sinusoidal models were used for studying area and production of sugarcane in Tamil Nadu. The Levenberg-Marquardts procedure is the most efficient iteration procedure described in the methodology, which was used for solving nonlinear normal equations. The results are discussed in the followings.

Model based trend analysis for area under sugarcane in Tamil Nadu

For the area under the cultivation of sugarcane, the nonlinear models such as Logistic, Rational, Gompertz, Sinusoidal and Weibull models were fitted which were graphically represented in the Figure 1 and 2. The results presented in the Table 1 which reveals that, among the different nonlinear models fitted, the maximum R^2 value of 74.7

per cent was observed in the logistic model with the minimum values of RMSE (27.770) and MAE (18.737) on comparison with all other nonlinear models. The next best nonlinear model was the Rational model with 73 per cent of R^2 value.

The *p* value of Shapiro-Wilks test statistic (0.920) and the Run test statistic (0.436) indicates that the residuals of the logistic model were normal and random respectively. The Durbin-Watson statistic recorded the value of 1.577, which indicated that there was no serial correlation among the residuals and were independent. The scatter diagram and normal plot for the residuals of the logistic model confirmed those assumptions.

For the best fitted logistic model, all the model coefficients were highly significant at 1 per cent. The parameter estimates of the logistic model were with a carrying capacity of 322.627 and the intrinsic growth rate of 0.176.

Among the nonlinear models fitted for the area under sugarcane, obtained suitable logistic function was as follows,

$$\hat{Y} = \frac{322.627}{1 + 1.003 \times \exp\{-0.176\,X\}} \quad R^2 = 74.7 \text{ per cent}$$

Model based trend analysis for production of sugarcane in Tamil Nadu

For the production of sugarcane, the nonlinear models such as Logistic, Rational, Gompertz, Sinusoidal, Weibull and Gaussian models were fitted which were graphically represented in the Figure 3 and 4. The results presented in the Table 2 revealed that, among the different nonlinear models fitted, the maximum R^2 value of 78.2 per cent was observed in the Gaussian model with the minimum RMSE (2.604) and MAE (2.760) values on comparison with all other nonlinear models.
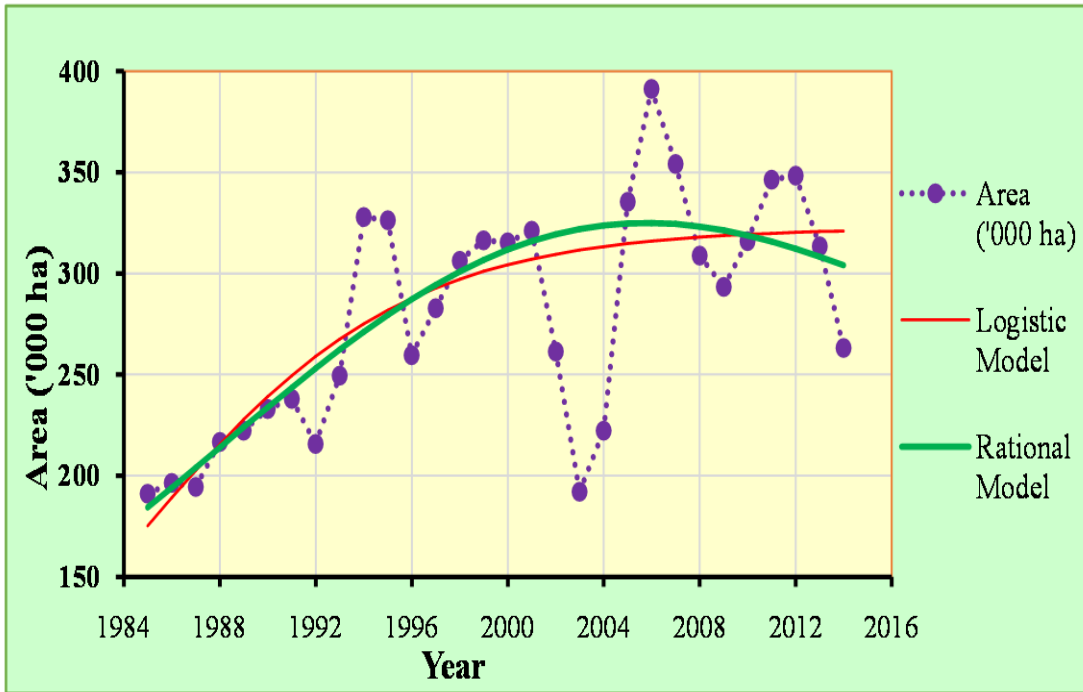
Fig. 3.1: Graph of the actual values and fitted models for the area undersugarcane in Tamil Nadu
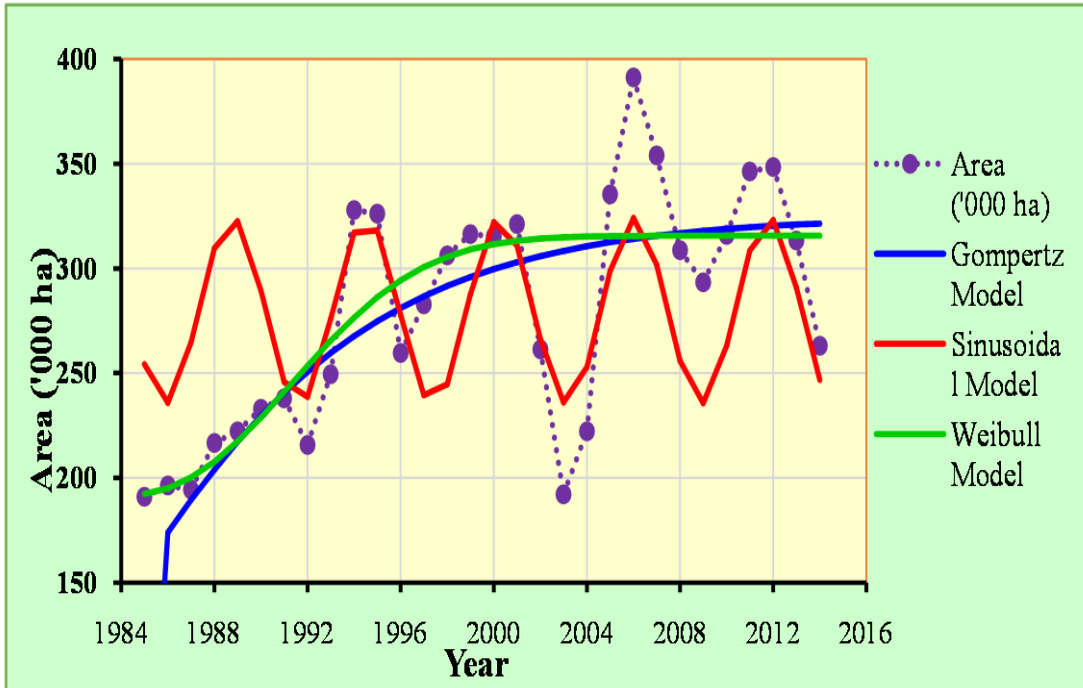

Fig. 3.2: Graph of the actual values and fitted models for the area under sugarcane in Tamil Nadu
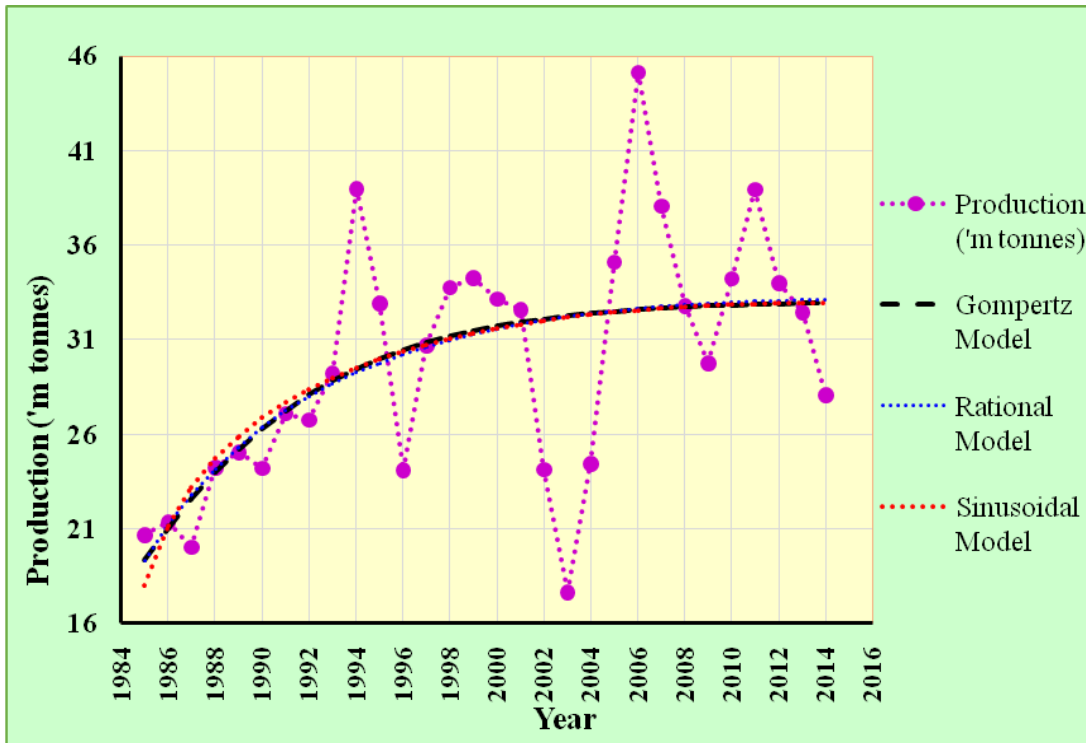
Fig. 3.3: Graph of the actual values and fitted models for the production of sugarcane in Tamil Nadu
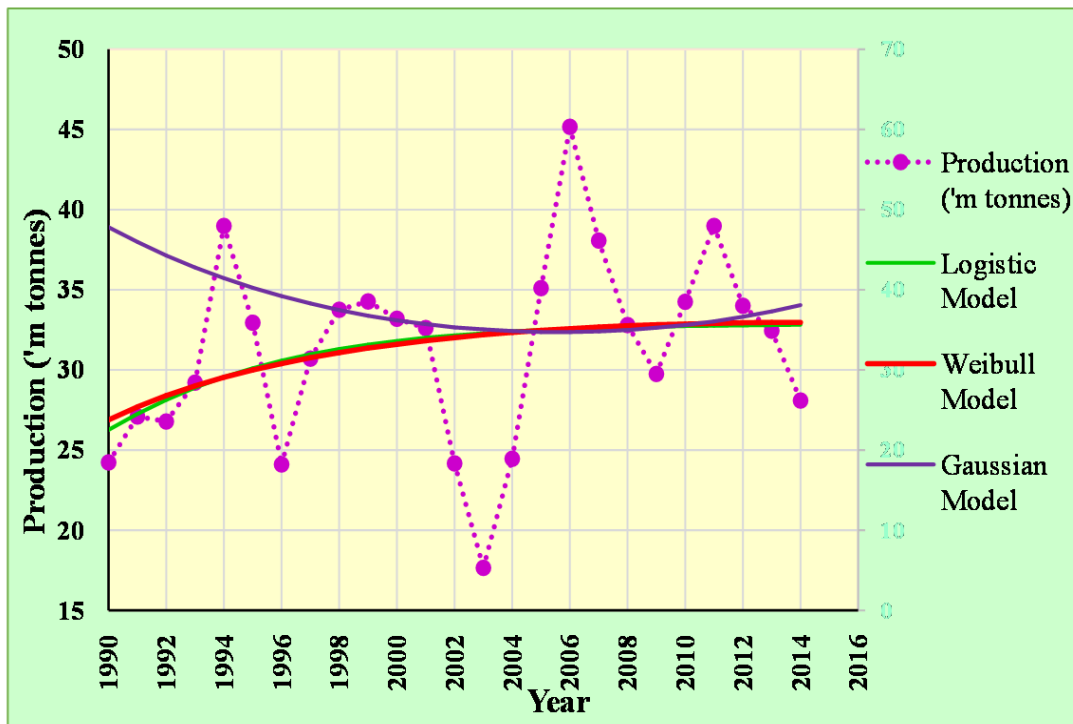

Fig. 3.4: Graph of the actual values and fitted models for the production of sugarcane in Tamil Nadu

Table.2 Estimates of the parameters along with model adequacy of fitted nonlinear models for area under sugarcane (1985-2014)							
	Estimates	**Nonlinear Models**					
		Logistic	**Gompertz**	**Rational**	**Sinusoidal**	**Weibull**	**Gaussian**
Parameters	**Carrying Capacity / Intercept (A)**	322.627**	324.796**	174.628**	279.756**	315.601**	331.641**
		(11.638)	(13.230)	(20.208)	(9.107)	(8.659)	(8.279)
	Function of initial value (B)	1.003**	-0.324	4.617	44.373**	124.032**	21.916**
		(0.226)	(0.177)	(9.189)	(12.866)	(22.722)	(1.455)
	Intrinsic growth rate / slope (C)	0.176**	0.147**	-0.028	1.091**	0.006	0..190
		(0.049)	(0.045)	(0.027)	(0.033)	(0.015)	(2.135)
	Added Parameter (D)	-	-	0.001**	1.090	2.308*	-
				(0.0001)	(0.594)	(1.093)	
Goodness of fit	**R^2**	0.747*	0.727*	0.730*	0.314**	0.732*	0.741*
	S-W test (p value)	0.920	0.894	0.552	0.205	0.342	0.426
	Run Test (p value)	0.436	0.436	0.436	0.005**	0.700	0.847
	D-W Statistic	1.577	1.551	1.662	0.368	1.680	1.546
	RMSE	27.770	28.246	28.600	49.653	27.997	29.163
	MAE	18.737	21.351	21.609	38.958	20.933	19.851
* Significant at 5% level; ** Significant at 1% level							
RMSE: Root Mean Square Error; MAE: Mean Absolute Error							
Values in parentheses () indicate standard errors							

Table.3 Estimates of the parameters along with model adequacy of fitted nonlinear models for production of sugarcane (1985-2014)							
Estimates		**Models**					
		Logistic	**Gompertz**	**Rational**	**Sinusoidal**	**Hoerl**	**Gaussian**
Parameters	**Carrying Capacity / Intercept (A)**	32.885**	33.090**	17.009*	30.365**	18.122**	34.735**
		(1.803)	(2.018)	(8.181)	(0.889)	(2.608)	(0.814)
	Function of initial value (B)	0.845*	-0.456	4.088	-5.482**	0.992**	21.620**
		(0.410)	(0.385)	(10.044)	(1.292)	(0.008)	(1.296)
	Intrinsic growth rate / slope (C)	0.202	0.169	0.093	1.054**	0.247*	0.236**
		(0.117)	(0.104)	(0.330)	(0.027)	(0.099)	(2.027)
	Added Parameter (D)			0.000	-0.958		
				(0.002)	(0.491)		
Goodness of fit	**R^2**	0.359**	0.359**	0.358	0.442	0.497*	0.782*
	S-W test (p value)	0.074	0.087	0.116	0.156	0.524	0.186
	Run Test (p value)	0.193	0.041*	0.041*	0.436	0.083	0.993
	D-W Statistic	1.081	1.081	1.082	0.872	1.520	2.247
	RMSE	5.371	5.371	5.477	4.548	4.089	2.604
	MAE	3.429	3.441	3.460	3.587	2.682	2.760
* Significant at 5% level; ** Significant at 1% level							
RMSE: Root Mean Square Error; MAE: Mean Absolute Error							
Values in parentheses () indicate standard errors							

Table.1 Nonlinear regression models		
S. No.	**Name of the model**	**Model**
I	Logistic model	$Y = \dfrac{A}{1 + B\exp\{-C\,X\}} + e$
II	Gompertz Relation model	$Y = A\exp\{-\exp\{B - C\,X\}\} + e$
III	Rational Function	$Y = \dfrac{A + B\,X}{1 + C\,X + D\,X^2} + e$
IV	Gaussian Model	$Y = \dfrac{A\exp\{-(B - X)^2\}}{2\,C^2} + e$
V	Weibull Model	$Y = A - Be^{-C\,X^D} + e$
VI	Hoerl model	$Y = A\,B^X\,X^C + e$
VII	Sinusoidal model	$Y = A + B\cos(C\,X - D) + e$

The next best nonlinear model was the Hoerl model with 49.7 per cent of R^2 value.

The *p* value of Shapiro-Wilk test statistic (0.186) and the Run test statistic (0.993) to test for assumptions indicates that the residuals of the Gaussian model were normal and random respectively. The Durbin-Watson statistic recorded the value of 2.247 which indicated that there was no serial correlation among the residuals and was independent. The scatter diagram and normal plot for the residuals of the Gaussian model in support of numerical test confirmed the liability of the residual assumptions (Table 1–4).

For the best fitted Gaussian model, all the coefficients were showing significant at 1 per cent level of significance. The parameter estimates of the Gaussian model were with a carrying capacity of 34.735 and the intrinsic growth rate of 0.236.

Gaussian model which was found to be the suitable model for the production of sugarcane is as follows,

$$\hat{Y} = \frac{34.735 \times \exp\{-(21.620 - X)^2\}}{0.111} \qquad R^2 = 78.2 \text{ per cent.}$$

Sugarcane is one of the important cash crops in Tamil Nadu. Due to the climatic and many other reasons, there is a lot of fluctuations in the area and production of Tamil Nadu. So, there is a necessity to study the trend in area and production of sugarcane and the impact of precipitation on the productivity of sugarcane in different agro-climatic zones. It was observed that nonlinear models are more appropriate to visualize the temporal trend of area and production of sugarcane in Tamil Nadu. Logistic and Gaussian models were the most suitable fitted models which clearly explained the trend of area and production of sugarcane in Tamil Nadu.

References

Anonymous, 2015.*Food and Agriculture Organization of United Nations statistics 2015*. Food and Agriculture Organisation of United Nations, Rome, Italy. http://www.fao.org/faostat/en/#data.

Bradley, J. V., 1968. *Distribution-free Statistical Tests*. Prentice-Hall, Englewood Cliffs, NJ, USA.

Mohan, S., Rajendran, K., Sivam, D. and Saliha, B., 2007. Sugar –The wonder

cane. *Co-operative Sugar,* 38(10): 21–24.

Shapiro, S. S. and Wilk, M. B., 1965. An analysis of variance test for normality (complete samples). *Biometrika,* 52(3-4): 591-611.

Shapiro, S. S., Wilk, M. B. and Chen, H. J., 1968. A comparative study of various tests for normality. *Journal of the American Statistical Association,* 63(324): 1343-1372.

Venugopalan, R. and Shamasundaran, K.S., 2003. Nonlinear Regression: A realistic modeling approach in Horticultural crops research. *J.Ind. Soc.Ag.Statistics,* 56(1):1-6.

How to cite this article:

Dinesh Kumar, P., Bishvajit Bakshi and Manjunath, V. 2018. Nonlinear Modeling of Area and Production of Sugarcane in Tamil Nadu, India. *Int.J.Curr.Microbiol.App.Sci.* 7(10): 3136-3146. doi: https://doi.org/10.20546/ijcmas.2018.710.363