Indian Journal of Agricultural Sciences **87** (5): 685–9, May 2017/Short Communication

Pre harvest forecasting of crop yield using non-linear regression modelling: A concept

SANJEEV PANWAR[1], K N SINGH[2], ANIL KUMAR[3], BISHAL GURUNG[4], SUSHEEL KUMAR SARKAR[5], SIVARAMANE N[6] and ABHISHEK RATHORE[7]

Indian Council of Agricultural Research, Krishi Bhawan, New Delhi 110 001

Received: 25 October 2016; Accepted: 29 November 2016

Key words: Detrended yield, Forecasting, Nonlinear model, Weather Indices Approach**,** Weather variables

A reliable and timely forecast of crop production helps in planning, formulation and implementation of policies relating to food procurement, its distribution, price, import and export and for exercising several administrative measures for storage and marketing of agricultural commodities. Thus pre-harvest forecasting of production is required when crop is still standing in the field. An efficient forecasting is thus a pre-requisite for food supply information system at district and state level. The final crop production estimates, though based on objective crop-cutting experiments, are of limited utility as these become available much later after the crop harvest. In view of this, there is a need for developing an objective methodology for pre-harvest forecasting of crop yield.

The main factors affecting crop yield are inputs and weather experienced by the crops during growth period. Use of data on these factors forms one approach for forecasting crop yields. The other approach uses plant vigour measured either through plant characters or through remotely sensed data. This approach is based on the fact that various factors affect crop growth through plant processes. These effects are manifested through crop stand, number of tillers, root length, leaf area, number of ear heads etc., which ultimately determine crop yield. A number of techniques based on different types of data have been developed in India and abroad.The efforts are made on forecasting yield rate (productivity) instead of production, since area under crops is available much before actual harvest through timely reporting schemes/remotely sensed data. Several models based on weather indices were attempted at district level and for agro-climatic zones. This approach has been used for one location only and that too at district level. Also, there was a need to study the performance of the models developed through various approaches at district and agro climatic zone level in order to arrive at a proper methodology for forecasting crop productivity at district/ agro climatic/state level.

This paper deals a concept of methodology to develop the weather based forecast models using non-linear regression approach and detrended yield for the important crops, viz rice, wheat, sugarcane for various districts and agro-climatic zones of UP State.

Methodology

Some statistical approaches have been used for development of regressors based on weather variables (subsequently used in forecast models) which are as follows:

Non-linear regression approach

There are several non-linear models which can fit different patterns in the data. The widely used non-linear models such as Logistic, Gompertz, Mono-molecular, Weibull, MMF and Richards which are expected to provide a reasonable representation of crop yield as compared to linear models will be tried for fitting the yield of selected crop in a selected location using the weather data.

Nonlinear models: It is well recognized that any type of statistical inquiry in which principles from some body of knowledge enter seriously into the analysis is likely to lead to a 'Nonlinear model' (Seber and Wild 2003). Such models play a very important role in understanding the complex inter-relationships among variables. A 'nonlinear model' is one in which at least one of the parameters appears nonlinearly. More formally, in a 'nonlinear model', at least one derivative with respect to a parameter should involve that parameter. Examples of a nonlinear model are:

$$Y(t) = exp(at+bt^2) \quad (1a)$$

$$Y(t) = at + exp(-bt). \quad (1b)$$

Note. Some authors use the term 'intrinsically linear' to indicate a nonlinear model which can be transformed to a linear model by means of some transformation. For example, the model given by eq. (1a) is 'intrinsically linear' in view of the transformation $X(t) = log_e Y(t)$.

Important nonlinear growth models: Those models,

[1] e mail: scientist 1775@gmail.com, [7] email: a.rathore@cgiar.org, ICRISAT, Hyderabad.

which describe the growth behaviour over time, are applied in many fields. In the area of population biology, growth occurs in plants, animals, organisms, etc. The type of model needed in a specific situation depends on the type of growth that occurs. In general, growth models are mechanistic in nature, rather than empirical. In the former, the parameters have meaningful biological interpretation; the latter is just like a 'black box' where some input is given and some output is taken out. A mechanistic model usually arises as a result of making assumptions about the type of growth, writing down differential or difference equations that represent these assumptions, and then solving these equations to obtain a growth model. The utility of such models is that, on one hand, they help us to gain insight into the underlying mechanism of the system and on the other hand, they are of immense help in efficient management. We now discuss briefly some well-known nonlinear growth models:

(i) Malthus Model. If X(t) denotes the population size or biomass at time t and r is the intrinsic growth rate, then the rate of growth of population size is given by

$$dX/dt = rX. \qquad (2)$$

Integrating, we get

$$N\ (t) = No\ \exp\ (rt), \qquad (3)$$

Where *Xo* denotes the population size at *t=0*. Thus this law entails an exponential increases for *r>0*. Furthermore, *X(t)* →∞ *as t* →∞, which cannot happen in reality.

Note. The parameter *r* is assumed to be positive in all models.

(ii) Monomolecular Model. This model describes the progress of a growth situation in which it is believed that the rate of growth at any time is proportional to the resources yet to be achieved, i.e.

$$dX/dt = r(K-X), \qquad (4)$$

where *K* is the carrying size of the system. Integrating eq. (4), we get

$$N(t) = K- (K-Xo)\ \exp\ (-rt). \qquad (5)$$

(iii) Logistic Model. This model is represented by the differential equation

$$dX/dt = rX\ (1-X/K). \qquad (6)$$

Integrating, $X(t) = K / [1+(K/Xo-1)\ \exp\ (-rt)]$ (7)

The graph of *X(t)* versus *t* is elongated S-shaped and the curve is symmetrical about its point of inflexion.

(iv) Gompertz Model. This is another model having a sigmoid type of behaviour and is found to be quite useful in biological work. However, unlike the logistic model, this is not symmetric about its point of inflexion. The differential equation for this model is

$$dX/dt = rX\ \log_e\ (K/X). \qquad (8)$$

Integration of this equation yields

$$X(t) = K\ \exp[\log_e\ (X_o/K)\ \exp(-rt)]. \qquad (9)$$

(v) Richards Model. This model is given by

$$dx/d = rX(K^m–X^m)/(mK^m), \qquad (10)$$

Which, on integration, gives

$$X(t) = K\ X_0/[X_0 + (K^m–X^m{}_0)\ \exp(–rt)[^{1/m}, \qquad (11)$$

Evidently, the last three models are particular cases of this model when *m= −1,1,0* respectively. However, unlike the earlier models, this model has four parameters.

There are four major nonlinear estimation procedures, namely (a) Gauss-Newton Method, (b) Steepest-Descent Method, (c) Levenberg-Merquadt Technique and (d) Do Not Use Derivative (DUD) Method. Either Gauss-Newton Method (the most widely used and reliable procedure for computing nonlinear least square estimates) or any method which yields efficient estimates with proper convergence will be applied in the present study.

The residuals from above model will be used in the subsequent linear model for fitting against different forms of the weather variables including their indices. The weather variables within an agricultural year will be aggregated through mean or through indices as mentioned above. Further, indices will be computed based on the influence (positive or negative) of the selected weather variable on the crop yield.

Criteria for model selection

Criteria such as Root Mean Squared Error, Mean Absolute Error, Mean Absolute Percentage Error, Mean Squared Error, Theil statistics and One Step Ahead Forecasting will be used in the study.

Test of randomness of residuals

Randomness is one of the key assumptions in determining if a Univariate statistical process is in control. If the randomness assumption is not valid, then a different model, either a times series model or a non-linear model, needs to be used. Run test is conducted to study whether the successive observations of error term are random or not. The residuals are examined and they were replaced using "+" or "-" sign according as it is positive or negative. Let n_2 the number of pluses and n_2 the number of minuses in the series of residuals. Then, the number of runs, where a run is defined as a sequence of symbols of one kind separated by symbols of another kind (Siegel and Castellan 1988) are counted. The null hypothesis put forth is given below as:

H_0: The set of residuals is random

H_1: The set of residuals is non-random.

The resulting statistical test is the one sample run test. The mean and variance of the sampling distribution of r, the number of runs, is given

$$\text{Mean}(\mu) = \left[\frac{2n_1n_2}{(n_1+n_2)}\right] + 1$$

And

$$\sigma^2 = \frac{2n_1n_2\left(2n_1n_2\text{-}n_1\text{-}n_2\right)}{\left(n_1+n_2\right)^2\left(n_1+n_2\text{-}1\right)}$$

Therefore, for large samples, the required test statistic is

$$Z=\frac{[r+h-(2n_1n_2/n)-1]}{\sqrt{\{[2n_1n_2(2n_1n_2-n_1-n_2)]/[(n_1+n_2)^2(n_1+n_2-1)]\}}}$$

where h = 0.5 if $r<2n_1n_2/(n+1)$ and h= –0.5 if $r>2n_1n_2/(n+1)$, n is number runs.

Since H_1 does not predict the direction of the deviation from randomness, a two-tailed rejection region is used. Thus H_0 is rejected at level of significance α $|Z|>Z_{\alpha/2}$ where $Z_\alpha = P(Z>Z_\alpha) = \alpha$.

Weather indices approach

For each weather variable, two indices will be developed, one as simple total of values of weather variables parameter in different weeks and the other one as weighted total, weights being correlation coefficients between variables to forecast and weather variable in respective weeks. The first index represents the total amount of weather parameter received by the crop during the period under consideration. While the other one takes care of distribution of weather parameters with special reference to its importance in different weeks in relation to the variable to forecast. On similar line, indices were computed with products of weather variables (taken two at a time) for joint effects.

These indices are computed as follows:

$$Y=a_0+\sum_{i=1}^{p}\ \sum_{j=0}^{1}a_{ij}Z_{ij}+\sum_{i\neq i'=1}^{p}\ \sum_{j=0}^{1}b_{ii'}Z_{ii'j}+e$$

$$Z_{ij}=\sum_{w=n_1}^{n_2}r_{iw}^{j}X_{iw}$$

$$Z_{ii'j}=\sum_{w=n_1}^{n_2}r_{ii'w}^{k}X_{iw}X_{i'w}$$

where, Y is variable of forecast, X_{iw} is value of i^{th} weather variable in w^{th} week, r_{iw} is correlation coefficient between Y and i^{th} weather variable in w^{th} week, $r_{ii'w}$ is correlation coefficient between Y and product of X_i and $X_{i'}$ in w^{th} week, p is number of weather variables, n_1 is initial week for which weather data are included in the model, n_2 is final week for which weather data are included in the model and a_{ij} and b_i are parameters to be estimated, r^j_{iw} is appropriate weights (differ as per index approach and nature of variables) to be used in computing index, ε is the random error.

Forecast models

The weather indices (weighted and unweighted) and trend variable Y have been used as regressors for developing forecast models at district level.The form of the model was

$$Y= a_0+\sum_{j=1}^{p}\sum_{k=0}^{1} a_{jk}Z_{jk}+\sum_{j,j'}^{p}\sum_{k=0}^{1} b_{jj'k}Z_{jj'k}+cY_r+e$$

Stepwise regression technique was used for retaining significant variables only in the forecast models in each approach. The variables left in the final models were significant at the 10 % level and no other variable met the 5 % significance level for entry into the model.

Comparison of models

Comparison of forecast models developed through three approaches was done on the basis of root mean square error (RMSE) of the models, and MAPE statistic.

Forecast/validation of models

The performance of forecasts obtained through different approaches was judged on the basis of Root Mean Square Error (RMSE).

Final forecast models (Forecasting model-linear and non-linear model using weather indices)

i) This model has two steps. The first one is a nonlinear model (or best selected models, i.e Logistic, Gompertz, Monomolecular, Richards, Weibull or MMF) of Y on t and the second model is error (obtained from the first step) is a linear function of selected weather index (Max. temp./min temp/ rainfall/relative humidity).

Step 1:

The general form of the model is Y=

$$Y=\frac{a}{\left(1+\exp\left(b-\left(c^*t\right)\right)\right)}+E1$$

Step 2:

The general form of the model is E1 = b1 + b2* X_1 + error

The estimated equation:

$E1 = a + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + c_1X_1X_2 + c_{12}X_1X_2 + c_{13}X_1X_3 + c_{14}X_1X_4 + c_{23}X_2X_3 + c_{24}X_2X_4 + c_{34}X_3X_4$ + error

Note: The parameters which are insignificant where subsequently removed using step-wise method.

ii) Simple linear model(Y=a1+a2*X1+error) where Y stands for the yield of rice/wheat during the year t and X1t= Index developed from the weather variable, max. temp or min. temp/rainfall/relative humidity pertaining to the year t.

$E1 = a + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + c_1X_1X_2 + c1_2X_1X_2 + c_{13}X_1X_3 + c_{14}X_1X_4 + c_{23}X_2X_3 + c_{24}X_2X_4 + c_{34}X_3X_4$ + error.

The parameters which are insignificant where subsequently removed using step-wise method. As a whole summarized as: (i) Trend analysis has been done through linear and non-linear approaches. (ii) Residuals/detrended yield were considered as a character under study. (iii) In which for each weather variable two indices have been developed, one as simple total of values of weather parameter in different weeks and the other one as weighted total, weights being correlation coefficients between detrended yield and weather variable in respective weeks. (iv) Weather indices based regression models were developed using weather indices as independent variables while detrended yield (residuals) was considered as dependent variable.

considered. The two step nonlinear forecasting model was found to be superior to the linear model as its RMSE value (1.40) is much lower as compared that of linear model (1.69). For fitting this residual model the nonlinear model (Logistic) which was found to have better fit was used to output the residuals. The negative values of the coefficients for variables such as X1(Max), X2(Min), X3(RF), X4(RH), X6(max*rf) and X7 (max*rh) showed that increase in the variables results in decrease in the yield. On the other hand variable such as X5(max*min), X8(min*rf), X9(min*rh) and X10(rf*rh) yielded positive coefficients which means that increase in the variables will increase the yield of wheat crop in Bareily district of Uttar Pradesh. Similarly, RMSE values are much lower as compared that of linear model in almost all selected district of UP, thereby showing that two step nonlinear forecasting models (Table 1 to 3) were found to be superior for yield forecasting of wheat crop.At district level, models developed through nonlinear model's residuals approach were found superior in comparison to other approaches.

Table 1 Comparison of fitted linear and different non-linear models for data on Bareily district of UP of wheat yield

Parameters/ Statistics	Bareily (Wheat)				
	Linear	Logistic	Gompertz	Mono-molecular	MMF
A	6.98	32.14	36.97		5.58
B	0.51	1.34	0.57		-1.89
C		0.07	0.04		-1.79
D					0.85
Goodness of fit Statistics					
R2	0.94	0.96	0.95		0.94
MSE	2.69	1.99	2.15		2.62
RMSE	1.64	1.41	1.46		1.62
MAE	1.29	1.08	1.14		1.27
MAPE	7.8	6.84	7.27		8.08
Theil statistics	1.88	0.006	0.05	Not converged	0.91

Table 2 Bareily wheat yield forecast models (Two steps non-linear and linear models) Based on weather indices approach

District	Forecasting Model	Goodness of fit	
		R2	RMSE
Bareily Non-linear	$Yt = -0.55 + \underset{(0.019)}{0.48}\ Z21 - \underset{(0.0014)}{0.009}Z121 - \underset{(0.0009)}{0.0024}\,Z241$	0.85	1.40
Linear	$Yt = -1.8 + \underset{(0.09)}{0.47}\,Z21 + \underset{(0..001)}{0.002}\,Z121 - \underset{(0.01004)}{0.0025}\,Z241$	0.83	1.69

Table 3 MAPE of forecast of wheat yield from models developed at district level through two step nonlinear and linear models based on weather indices approach

District	Two step nonlinear models (MAPE)	Linear models (MAPE)
Bareily	7.11	10.13

Example: For developing a forecast model for Bareily district of Uttar Pradesh is as follows:

Crop yield forecast models based on weather variables (Max. temp., Min. temp., RH (morning), Rainfall) have been developed by using weekly data for wheat crops. These forecast regression models, weather variables based indices were used as regressors. Data from the year 1970-71 to 2007-08 have been utilized for model fitting and two years data for 2008-09 and 2009-10 were set apart for the validation of the model. District level forecasts of crop yields for the subsequent years have been obtained using the forecast models developed. These were averaged to get forecast for state as a whole taking area under the crop in different districts as weights. These forecasts were compared with the actual observed yields at district level.

For developing wheat yield forecast models weekly weather variables data starting from 40th SMW (first week of October) to 03rd SMW (15-21 January) have been

SUMMARY

The concept of pre-harvesting of crop yield using nonlinear growth models and detrended yield for developing yield forecast model is rarely employed in forecasting. A novel approach attempted in this study to use nonlinear models with different weather variables and their indices and compare them to identify a suitable forecasting model. Weather indices based regression models were developed using weather indices as independent variables while detrended yield (residuals) was considered as dependent variable. The approach provided reliable yield forecast about two months before harvest.

REFERENCES

Aditya Kaustav. 2008. Forecasting of crop yield using discriminant function technique. MSc thesis, IARI, New Delhi.

Agrawal, Ranjana, Ramakrishna YS, Rao Kesava AVR, Amrender Kumar, Bhar, Lalmohan, Madan Mohan and Saksena, Asha. 2005. Modeling for forecasting of crop yield using weather parameters and agricultural inputs (Under AP Cess Fund Scheme of ICAR, New Delhi).

Bates D M and Watts D G 1988. *Nonlinear regression analysis and its applications*. John Wiley and Sons, New York.

Baier W. 1973. Crop-weather analysis model. Review and model development. *Journal of Applied Meteorology* ***12(6)***: 937–47.

Baier W. 1977. Crop weather models and their use in yield assessments. Tech. note no. **151**, WMO, Geneva, 48 p.

Fisher R A. 1924. The influence of rainfall on the yield of wheat at Rothamsted. Royal Society London. *Phil. Trans. Ser. B.* **213:** 89–142.

Jain R C, Agrawal, Ranjana and Jha M. 1980. Effects of climatic variables on rice yield and its forecast *Mausam* **31**(4): 591–6.

Johnson D A, Alldredge J R and Vakoch D L. 1996. Potato late blight forecasting models for the semiarid-environment of south-central Washington. *American Phytopathology **86***: 480–4.

Khamis A, Zuhaimy I, Khalid H and Ahmad TM, 2005. Non-linear growth models for modeling oil palm yield growth. *J. Math. Stat.* **1**: 225–33.

Madhav K V. 2003. 'Study of statistical modeling techniques in agriculture'. Ph D thesis, IARI, New Delhi.

Rai T and Chandrahas. 2000. Use of discriminant function of weather parameters for developing forecast model of rice crop (project report). IASRI , New Delhi.

Seber G A F and Wild C J. 1989. *Non-linear Regression.* John Wiley and Sons, New York.

Sanjeev Panwar, Anil Kumar, Susheel Kumar Sarkar, Ranjit Kumar Paul,Bishal Gurung and Abhishek Rathore 2016. Forecasting of common carp fish production from ponds using nonlinear growth models—A modelling approach. *Journal of the Indian Society of Agricultural Statistics* **70**(2) 139–44.

Weibull W. 1951. A statistical distribution function of wide applicability. *Journal of Applied Mechanics* **18**: 293–7.