WEATHER BASED CROP FORECASTING TECHNIQUES

Amrender Kumar
Indian Agricultural Statistics Research Institute
Library Avenue, New Delhi – 110012
akjha@iasri.res.in

Introduction

Reliable and timely forecasts provide important and useful input for proper, foresighted and informed planning, more so, in agriculture which is full of uncertainties. Agriculture now-a-days has become highly input and cost intensive. Without judicious use of fertilizers and plant protection measures, agriculture no longer remains as profitable as before. Uncertainties of weather, production, policies, prices, etc. often lead to mass suicides by farmers. New pests and diseases are emerging as an added threat to the production. Under the changed scenario today, forecasting of various aspects relating to agriculture are becoming essential. But in spite of strong need for reliable and timely forecasts, the current status is far from satisfactory. For most of the sectors, there is no organized system of forecasting.

Crop production forecast system

Crop yield is affected by technological change and weather variability. It can be assumed that the technological factors will increase yield smoothly through time and, therefore, year or some other parameter of time can be used to study the overall effect of technology on yield. Weather variability both within and between seasons is the second and the only uncontrollable source of variability in yields. Weather variables affect the crop differently during different stages of development. Thus extent of weather influence on crop yield depends not only on the magnitude of weather variables but also on the distribution pattern of weather over the crop season which, as such, calls for the necessity of dividing the whole crop season into fine intervals. This will increase number of variables in the model and in turn a large number of constants will have to be evaluated from the data. This will require a long series of data for precise estimation of the constants which may not be available in practice.

Fisher (1924) and Hendricks and Scholl (1943) have done pioneering work in crop weather relationship. They have given models which require small number of parameters to be estimated while taking care of distribution pattern of weather over the crop season.

Fisher assumed that the effect of change in weather variable in successive periods would not be an abrupt or erratic change but an orderly one that follows some mathematical law. He assumed that these effects are composed of the terms of a polynomial function of time. Further, the value of weather variable in w-th week, \( X_w \) was also expressed in terms of orthogonal functions of time.

\[
A_w = a_0[f_0(w)] + a_1[f_1(w)] + \ldots \ldots + a_k[f_k(w)]
\]

\[
X_w = \rho_0 [f_0(w)] + \rho_1 [f_1(w)] + \ldots \ldots + \rho_k [f_k(w)]
\]

where \( \rho_l \)'s are distribution constants.
Substituting these in usual regression equation

\[ Y = A_0 + A_1 X_1 + A_2 X_2 + \ldots + A_n X_n + e \]

(here \(Y\) denoted yield and \(X_w\) rainfall in \(w\)-th week, \(w = 1, 2, \ldots, n\)) and utilising the properties of orthogonal and normalised functions, he obtained

\[ Y = A_0 + a_0 \rho_0 + a_1 \rho_1 + a_2 \rho_2 + \ldots + a_k \rho_k + e \]

where \(A_0, a_0, a_1, a_2, \ldots, a_k\) are constants to be determined and \(\rho_i\) (\(i=1, \ldots, k\)) are distribution constants of \(X_w\). Fisher has suggested to use \(k = 5\) for most of the practical situations. In fitting this equation for \(k = 5\), the number of constants to be evaluated will remain 7, no matter how finely growing season is divided. This model was used by Fisher for studying the influence of rainfall on the yield of wheat.

Hendricks and Scholl (1943) have modified Fisher's technique. They divided the crop season into \(n\) weekly intervals and have assumed that a second degree polynomial in week number would be sufficiently flexible to express the effect of weather on yield in successive periods. Further, they used values of weather variables as such. Mathematically

\[ A_w = a_0 + a_1 w + a_2 w^2 \]

In particular,
\[
\begin{align*}
A_1 & = a_0 + 1.a_1 + 1^2.a_2 \\
A_2 & = a_0 + 2.a_1 + 2^2.a_2 \\
& \ldots \ldots \ldots \ldots \ldots \\
A_n & = a_0 + n.a_1 + n^2.a_2
\end{align*}
\]

Substituting the expression for \(A_w\) in regression equation, the model was obtained as

\[ Y = A_0 + a_0 \sum_w X_w + a_1 \sum_w w X_w + a_2 \sum_w w^2 X_w + e \]

In this model number of constants to be determined reduces to 4, irrespective of \(n\).

This model was extended for two weather variables to study joint effects.

The model obtained was

\[ Y = A_0 + a_0 \sum_w X_{1w} + a_1 \sum_w w X_{1w} + a_2 \sum_w w^2 X_{1w} + \\
\quad b_0 \sum_w X_{2w} + b_1 \sum_w w X_{2w} + b_2 \sum_w w^2 X_{2w} + \\
\quad c_0 \sum_w X_{1w} X_{2w} + c_1 \sum_w w X_{1w} X_{2w} + c_2 \sum_w w^2 X_{1w} X_{2w} + e \]

Since the data for such studies extended over a long period of years, an additional variate \(T\) representing the year was included to make allowance for time trend.
Another important contribution in this field is by Baier (1977). He has classified the crop-weather models in three basic types.

1. Crop growth simulation models
2. Crop-weather Analysis models
3. Empirical statistical models

**Crop-growth simulation models**

A crop growth simulation model may be defined as a simplified representation of the physical, chemical and physiological mechanisms underlying plant growth processes. If the basic plant processes - production and distribution of dry matter and water relations are properly understood and modelled, the entire response of the plant to the environmental conditions can be simulated. Therefore, there is no need to differentiate between climatic regions, since the simulation model itself will show the limiting factors for growth. In humid climates with low temperature and radiation levels, the model will generally show the greatest response of yields to increase in total radiation received. In an arid and hot climate it will show the greatest response to the distribution and total amount of precipitation. Various time intervals can be introduced in simulation models, for example, in view of the daily cycle of many plant processes, hourly intervals are most practical. It is then assumed that the rate calculated for a particular moment does not change appreciably over a period of one hour. It is possible to evaluate thereby specific processes such as photosynthesis, transpiration or respiration for an hour and then accumulate the hourly rates over the day and the daily rates over the growing season in order to arrive at the total seasonal dry matter production or yield of economic products. Simulation programme must be regarded more as a guide to research into the behaviour of biological systems rather than as a final solution. Simulation can be most useful if the model accounts for most relevant phenomena and contains no false assumptions. Simulation provides an insight into crop-weather relationships, explains why some factors are more important for yield than others, suggests factors likely to have statistical significance and provides the basis for new experiments on processes which appear to be important but are not yet sufficiently understood. Thus, the simulation approach does not replace the statistical approach, but is complementary to it.

**Crop-weather analysis models**

Crop-weather analysis models are defined here as the product of two or more factors, each representing the (simplified) functional relationship between a particular plant response (e.g. yield) and the variations in selected variables at different plant developmental phases. The overall effects, as expressed by the numerical values of the factors modify each other but are not additive as in the case of a multivariate linear regression equation. Such models do not require a formulated hypothesis of the basic plant and environmental process; thus, the input requirements are less stringent but the output information is more dependent on the input data and less detailed than in the case of simulation models. Therefore, crop-weather analysis models are a practical research tool for the analysis of crop responses to weather and climate variations when only climatological data are available. Conventional statistical procedures are used in such models to evaluate the coefficients relating crop responses to climatological or derived agrometeorological data. A convenient time interval is one day, but in practice shorter or longer periods can also be used, provided the response characteristics of the crops do not change appreciably over the selected period in relation to the variable taken into consideration.
Empirical Statistical Models

In the empirical approach, one or several variables (representing weather or climate, soil characteristics or a time trend) are related to crop responses such as yield. The weighting coefficients in these equations are by necessity obtained in an empirical manner using standard statistical procedures, such as multivariable regression analysis. This statistical approach does not easily lead to an explanation of the cause and effect relationships but it is a very practical approach for the assessment or prediction of yields. The coefficients in such empirical models and the validity of the estimates depend to a large extent on the design of the model, as well as on the representativeness of the input data. If the soil and climate conditions and the cropping practices are fairly homogeneous over the area represented by the input data, or if soil and geography are properly weighted in the equations, then it can be expected that the coefficients and the estimates have practical significance for the assessment of the crop conditions or prediction of yields for the specific area in question.

Several Empirical Statistical models were developed all over the world. The independent variables included weather variables, agrometeorological variables, soil characteristics or some suitably derived indices of these variables. Water Requirement Satisfaction Index (WRSI), Thermal Interception Rate Index (TIR), Growing Degree Days (GDD) are some agroclimatic indices used in models. Southern Oscillation Index (SOI) has also been used with other weather variables to forecast crop yield (Ramakrishna et al. 2003). To account for the technological changes year variable or some suitable function of time trend was used in the model. Some workers have also used two time trends. Moving averages of yield were also used to depict the technological changes.

In contrast to empirical regression models, the Joint Agricultural Weather Information Centre employs the crop weather analysis models that simulate accumulated crop responses to selected agrometeorological variables as a function of crop phenology. Observed weather data and derived agrometeorological variables are used as input data.

M Frere and G.F. Popov (1979) used the method which utilises actual rainfall and climatological information for the calculation of water requirement of crops and in turn crop water balance. The method is based on a cumulative water balance established over the whole growing season for the given crop and for successive periods of 10 days or a week. The water balance is difference between precipitation received by the crop and the water lost by the crop and the soil. Based on water surplus and deficit they have calculated index. Initially the index is taken to be 100 and is modified in successive decades/weeks depending on the water surplus or deficits. This index has been shown to be directly related to yield and can give a very satisfactory and early qualitative estimation of yields in rainfed crops. It may be possible to derive quantitative estimations of yields also but these estimates will have to be based on the potential yield of crops which will depend on local environmental conditions and will vary from place to place. It may also be mentioned that the method is intended mainly for utilisation in developing countries, where in rainfed agriculture the main constraint is generally inadequate availability of water to the crop. Therefore, the method does not directly involve the temperature which conditions the growth of the crop. However, the temperature intervenes indirectly in three ways in the method of crop water balance assessment. Firstly, the effect of air temperature may be noticed in the length of the growing cycle which is generally directly dependent on temperature. Further air temperature intervenes directly in the calculation of potential evapotranspirations and in this respect influences the whole water balance. Finally the external temperatures may be important in some climatic zones, particularly as regards frosts.
In India, major organisations involved in developing methodology for forecasting crop yield based on weather are IMD and IASRI. The methodology adopted by IMD involves identification of significant correlations between yield and weather factors during successive overlapping periods of 7 to 60 days of the crop growing season. By analysing the correlation coefficients for statistical and phenological significance, the critical periods when the weather variables have significant effect on yield are identified. The weather variables in critical periods are used through multiple regression analysis to obtain forecast equations. Using this methodology models were developed for principal crops on meteorological subdivisions basis. Data from various locations are averaged to get the figures for meteorological sub-divisions and these are utilized along with time trend to develop the forecast model. Monthly forecasts are issued from these models by taking the actual data up to time of forecast and normal for the remaining period. In some models yield Moisture Index, Generalised Monsoon Index, Moisture Stress, aridity anomaly Index are also used (Sarwade, 1988; Sarkar, 2002).

Weather indices based models
At IASRI, the model suggested by Hendricks and Scholl has been modified (Agrawal et al 1980; 1983; Jain et al 1980) by expressing effects of changes in weather variables on yield in the successive periods as second degree polynomial in respective correlation coefficients between yield and weather variables. This will explain the relationship in a better way as it gives appropriate weightage to different periods. Under this assumption, the models were developed for studying the effects of weather variables on yield using complete crop season data whereas forecast model utilised partial crop season data. These models were found to be better than the one suggested by Hendricks and Scholl.

These models were further modified (Agrawal et al 1986) by expressing the effects of changes in weather variables on yield in successive periods as a linear function of respective correlation coefficients between yield and weather variables. As trend effect on yield was found to be significant, its effect was removed from yield while calculating correlation coefficients of yield with weather variables to be used as weights. Effects of second degree terms of weather variables were also studied. The results indicated that (i) the models using correlation based on yield adjusted for trend effect were better than the ones using simple correlations, (ii) inclusion of quadratic terms of weather variables and also the second power of correlation coefficients did not improve the model. This suggests that the following models can be used to study effects of weather on yield and its forecasts.

\[ Y = A_0 + a_0 z_0 + a_1 z_1 + cT + e \]

where \( Z_j = \sum_{w=1}^{n} r_{wj} X_w \); \( j = 0, 1 \)

Here Y is yield, \( r_{wj} \) is correlation coefficient between the weather variable in w-th period(Xw) with yield (adjusted for trend effect), and e is error term. The models were further extended for studying joint effects.

The forecast model has been developed using partial crop season data considering all weather variables simultaneously. The model finally recommended was of the form
\[ Y = A_0 + \sum_{i=1}^{p} \sum_{j=0}^{1} a_{ij} Z_{ij} + \sum_{i \neq i'=1}^{p} \sum_{j=0}^{1} a_{i'i''j} Z_{i'i''j} + cT + e \]

where

\[ Z_{ij} = \sum_{w=1}^{m} r_{iw} X_{iw} \quad \text{and} \quad Z_{i'ij'} = \sum_{w=1}^{m} r_{i'iw} X_{iw} X_{i'w} \]

\( r_{iw}/r_{i'iw} \) is correlation coefficient of \( Y \) with \( i \)-th weather variable/product of \( i \)-th and \( i' \)-th weather variable in \( w \)-th period. \( m \) is period of forecast and \( p \) is number of weather variables used.

In this approach, for each weather variable, two types of indices were developed, one as simple total values of weather variable in different periods [un-weighted index -\( Z_{i0} \)] and the other one as weighted total  [weighted index \( Z_{i1} \)] weights being correlation coefficients between yield/de-trended yield (if trend is present) and weather variable in respective periods. On similar lines, for studying joint effects, un-weighted & weighted indices for interactions were computed with products of weather variables (taken two at a time). Stepwise regression technique was used to select important indices in the model.

These models were used to forecast yield of rice and wheat in different situations, viz (i) rainfed area having deficient rainfall (rice), (ii) rainfed area having adequate rainfall (rice) and (iii) irrigated area (wheat). The results revealed that reliable forecasts can be obtained using this approach when the crops are 10-12 weeks old. This approach was also used to develop forecast model for sugarcane (Mehta, et al. 2000). However, these studies were carried out at district level and required a long series data of 25-30 years which are not available for most of the locations. Therefore, the study has been undertaken to develop the model on agro-climatic zone basis for rice and wheat by combining the data of various districts within the zone so that a long series could be obtained in a relatively shorter time. Previous years yield, moving averages of yield and agricultural inputs were taken as the variables taking care of variation between districts within the zone. Year was included to take care of technological changes. Different strategies for pooling district level data for the zone were adopted. Results revealed that reliable forecasts can be obtained using this methodology at 12 weeks after sowing i.e. about 2 months before harvest. The data requirement reduced to 10-15 years as against 25-30 years for district level models. The study also revealed that forecast model will be appropriate to forecast the yield of zone even if data for some districts within the zone are not available at model development stage or at forecasting stage (Agrawal et al. 2001). The approach has been successfully used for forecasting yields of rice, wheat, sugarcane and potato for Uttar Pradesh. (Agrawal, et al. 2005, Mehta, et al. 2010).

**Complex Polynomial through GMDH technique**

This methodology has been successfully applied by Mustafi and Chaudhuri (1981) for forecasting monthly tea crop production. At IASRI use of this technique was explored for forecasting potato yield in Uttar Pradesh (Mehta, et al. 2010).

The main feature of this technique is that it itself selects the structure of the model without using a prior information about relationship of dependent variable (\( y \)) with independent variables (\( x_1, x_2, \ldots, x_p \)).
The fitted polynomial is of the form

\[ y = a + \sum_{i=1}^{p} b_i x_i + \sum_{j=1}^{p} c_{ij} x_i x_j + \sum_{i=1}^{p} \sum_{j=1}^{p} \sum_{k=1}^{p} d_{ijk} x_i x_j x_k + \ldots \]

The technique involves fitting of quadratic equations for all pairs of independent variables and identifying a few best performers in terms of predictive ability (using appropriate statistic); converting entire set of independent variables (called zero generation variables) to new variables (first generation variables) which are obtained as predicted values from these selected quadratic equations (of zero generation variables). The process of fitting and identifying best quadratic equations is repeated using first generation variables and second generation variables are obtained. The whole process is repeated with every new generation of variables till appropriate model is obtained (using certain criteria). At final stage, one best quadratic equation is selected as the final model.

Two approaches are followed for identifying few best performers. In the first approach, available data set is divided into two non-overlapping sets - training set and checking set. From training set, quadratic equations are fitted and checking set is used to test the predictability of different quadratic equations using root mean square error. In the other approach, PRESS statistic (predicted sum of square) is used wherein the whole data set is used as fit and check set. For evaluation of PRESS, each data point (one by one) is taken as a testing set (of size 1) and is predicted from the quadratic equation fitted from the remaining n-1 data points PRESS is calculated as

\[ PRESS = \sum_{s=1}^{n} [y_s - \hat{y}_s^{(s)}]^2 \]

where \( y_s \) is the value of the dependent variable for \( s \)th observation and \( \hat{y}_s^{(s)} \) is the predicted value of \( y_s \) computed from an equation based on remaining n-1 data points.

Theoretically the generation of variables is continued till the decreasing trend of PRESS statistic ceases i.e. PRESS statistic starts increasing. Some times even after number of generations the decreasing trend continues. This increases complexity of the model. In practical situations when there is abrupt decrease in value of PRESS statistic and at the same time the coefficient of determination is quite high, the procedure is terminated.

This approach was used to obtain district as well as agroclimatic zone level models for potato in Uttar Pradesh using weather indices (unweighted and weighted) as explanatory variables. The performance of this model was found to be better than indices based regression approach for Bareilly district and north eastern zone. For remaining districts and zones the performance was worse or at par with the indices based regression approach.

**Discriminant Function Technique**

Discriminant function technique is a linear/quadratic function that discriminates different groups the best. Use of this technology has been explored for obtaining quantitative forecast of yields for rice in Raipur district. This methodology involved grouping the long series of years into three groups - congenial, normal and adverse with respect to crop yields (adjusted for trend effect, if any). Using weather data of these groups, linear / quadratic discriminant
functions were obtained using phasewise weather data. Weather scores for each year at different phases of crop growth obtained through these discriminant functions were used alongwith inputs and time trend as regressors in model development through stepwise regression. Quadratic discriminant function was found appropriate and the methodology could provide reliable forecast two months before harvest. (Rai and Chandrahas 2000). The methodology was further modified using weekly weather data. Various strategies were proposed to solve the problem of number of variables more than number of observations. The study was carried out to forecast wheat yield in Kanpur district. The finally recommended strategy involved following steps. Discriminant functions were developed using weather variables data of first week. These discriminant functions were used to compute scores for each year. Taking data on weather variables in the second week and discriminant scores computed from the first week, discriminant function analysis was carried out which provided scores for each year based on data upto second week. The process was repeated for successive weeks data till the time of forecast and finally discriminant scores based on entire data were obtained for each year which alongwith trend were used to develop the model through stepwise regression technique. In contrast to earlier model by Rai et al. this model was based on complete data upto the time of forecast and relative importance of weather variables in different weeks (Aditya 2008). A detailed study using the recommended strategy has been carried out forecasting rice, wheat and sugarcane in Uttar Pradesh. Methodology was found to be successful for obtaining district / agroclimatic zone/state level forecasts. Performance was found to be better than weather indices based regression models for some cases whereas in some cases reverse trend was found (Chandrahas et al. 2010). However, this approach requires larger data base as compared to weather indices based regression approach.

**Principal Component Regression**

Principal component regression is a well known technique to reduce number of explanatory variables in the model. The technique involves conversion of explanatory variables into a set of uncorrelated variables with variances in descending order (known as principal components). The whole variation of the system explained by explanatory variables is explained by first few principal components which are used as regressors in the model in place of original variables. Besides solving the problem of number of explanatory variables more than number of observations, the technique also solves the problem of multicollinearity. The approach has been attempted for forecasting yields of rice, wheat and sugarcane in Uttar Pradesh but the approach was not found to be successful (Chandrahas et al. 2010).

**Water Balance Technique**

The concept of potential evapotranspiration (PET) was introduced by Thornthwaite and Penman in 1948 which use of agrometeorological variables in yield assessment models. Performance of models using agrometeorological variables was found to be better than the models using only meteorological variables (Baier and Robertson 1968). This is due to the fact that agrometeorological models use variables like soil moisture, actual and potential evapotranspiration, crop water requirement, effective rainfall etc. which take into account the soil properties and crop characteristics in addition to meteorological variables. These variables are estimated using Water Balance technique. This technique is useful for rainfed crops.
The water balance technique or model is a simple equation keeping an account of receipt and expenditure of water from the soil reserves. Mathematically

\[ S_i = S_{i-1} + R_i - E_i \]

where \( S_i \), \( R_i \) and \( E_i \) represent soil moisture, water received and expenditure respectively at the end of \( i \)-th period. The receipt (R) is in the form of rainfall and expenditure (E) is in the form of evapotranspiration from soil and crop cover. The stress depends on the demand and availability. When demand is more than the availability the stress occurs. Degree of stress depends on the gap between demand and availability. The variation in the model used by different research workers is in the form of variables used for indicating receipt and expenditure and the limitation imposed on these variables. For example in this equation, \( i \) represents the period after which balancing is done. It can be a day, week, fortnight or month. The number of periods depend on the length of the crop season, starting from the time of sowing. The depth of root zone can be kept constant for the entire season or can vary with the age of the crop. Receipt can be in the form of rainfall assuming that it is absorbed by the soil at a constant rate irrespective of the intensity of rain and antecedent moisture condition of soil. It can be made more realistic by using the actual infiltration rate of water for a given soil under a given intensity of rain and antecedent moisture condition. Similarly, expenditure which depends on demand can be estimated in different ways. It can simply be PET, the amount that can potentially evapotranspire under a given climatic condition or a fraction of it i.e. PET/2 or PET/4, irrespective of crop, and its stage of growth. This form of expenditure/demand is used when the objective is crop planning, estimation of drought proneness or agroclimatic classification of an area. For the purpose of monitoring and assessing yield of a crop, actual water requirement (WR) of the crop is taken as demand. The water requirement (at different time points) is estimated by multiplying the crop coefficients \( k \) (representing the crop characteristic and stage of the crop growth) and the value of evaporation (representing the loss of water to climatic condition such as temperature, wind velocity, humidity and sunshine at that stage). Estimates of evaporation are obtained either directly from an evaporimeter or from different formulae developed by Thornthwaite (1948), Penman (1948) and Christiansen (1966). For estimates obtained from different formulae different values of \( k \) are used. The values of \( k \) for different crops or groups of crops under given climatic conditions are developed by agronomists and water technology scientists by conducting field experiments.

If the receipt is more than the expenditure, the excess water is stored in the soil depending upon the water holding capacity and water already stored in the soil. If excess water is more than the retention capacity of the soil, it goes waste as runoff. Thus, in the process of estimating soil moisture an estimate of runoff is also obtained. The amount of rain water used by plants and stored in the soil is taken as effective rainfall (ER), when rainfall is not enough to meet the crop water requirement, the requirement is met from the soil moisture stored in the root-zone. When rain and soil moisture together are not enough to meet the requirement the amount actually available is extracted by the plants and part of requirement remains unfulfilled. In this manner an estimate of actual evapotranspiration (AE) is obtained. When AE is less than WR a stress to the crop occurs. \( 1 - (AE/WR) \) is used as estimate of stress to the crop. The ratio AE/PE is also sometimes used as variable for yield assessment. Since the final yield is the outcome of the aggregate of water/moisture availability or non-availability through its life cycle an accumulated stress or satisfaction index is prepared. As the deficiency at critical stages causes greater damage, weights are assigned to stress at different stages according to their importance. In this manner an accumulated weighted stress index (SI) or water requirement
satisfaction Index (WRSI) is obtained for each season. The index is related to yield through a regression model. The accumulated index helps in crop monitoring right from the date of sowing and provides pre-harvest estimate of yield during any time of the season. Therefore, pre-requisite for a seemingly simple model are estimates of many parameters of soil like depth, water holding capacity, wilting point, field capacity, infiltration rate under different antecedent moisture condition and crop parameters like crop coefficients, rooting pattern, and water extraction pattern of roots. Also a knowledge of critical stages of growth, an insight into effect of stress at different stages and an estimate of initial soil moisture are necessary. The success of the model depends on the accuracy of estimates of these parameters.

In a study conducted at IASRI a water balance model of the following form was used

\[ S_i = S_{i-1} + ER_i - WR_i \]

for estimating moisture stress to the pearl millet crop of IARI, New Delhi. Balancing was done at the end of each day of the crop season. Depth of root zone varied with the age of the crop. A new method was developed to calculate effective rainfall (ER) on the basis of amount of rain and antecedent moisture condition. ER was used in the model in place of rainfall. Estimates of moisture stress and moisture surplus were obtained from the model. Weights were assigned to stress at different stages. Detrimental effect of excess water in the root-zone was also taken into account and weights were assigned to surplus water also depending upon the time of its occurrence. An accumulated stress and surplus moisture index (SI) was prepared. SI was related to yield through a regression equation. The fitted equation explained 91% variation in pearl millet yield. A reduction of 42.7 kg/ha was expected due to per unit of stress in the potential yield of 3000 kg/ha. The error in predicted yield of two years was 3.2% and 0.5% respectively (Saksena and Bhargava 1995).

In another study, models were developed for rainfed sorghum, maize and rice using agrometeorological indices. Water balance was carried out at weekly intervals. Weighted stress index was prepared phase-wise by applying weights to surplus as well as deficit moisture depending upon the stage at which it occurred. Stress index of phase 2 i.e. 4 to 7th weeks after sowing played an important role in determining the yield of sorghum both at Delhi and Parbhani (Maharashtra) district. Models with phase 2 Index and trend variables as regressors could forecast yield 6 weeks before harvest. Deviations in forecast and observed yield varied between 3.5% to 11%. Similarly for maize crop at Delhi model with trend, index of surplus moisture at phase 1 and 2 and deficit moisture at phase 3 and 4 as predictor variables could forecast yield 4 weeks before harvest. Deviation in the forecast and observed yield was only 4.8 %. Model for rice in Raipur district included trend and accumulated index for the five phases up to maturity as explanatory variables.

**Artificial Neural Network Technique**

In contrast to regression approach, Artificial Neural Network (ANN) technique has been explored for forecasting yields of rice, wheat and sugarcane in Uttar Pradesh. (Kumar et al. 2010). This is an attractive tool under machine learning techniques for forecasting and classification purposes. ANNs are data driven self-adaptive methods in that there are few apriori assumptions about the models for problems under study. These learn from examples and capture subtle functional relationships among the data even if the underlying relationships are unknown or hard to describe. After learning from the available data, ANNs can often correctly infer the unseen part of a population even if data contains noisy
information. As forecasting is performed via prediction of future behaviour (unseen part) from examples of past behaviour, it is an ideal application area for ANNs, at least in principle. (Dewolf et al. 1997, 2000). However, the technique requires a large data base.

**Forewarning systems for crop pests and diseases**

Pests and diseases are one of the major causes of reduction in crop yield. Timely application of remedial measures may reduce the yield loss. For application of these measures one must have prior knowledge of the time and severity of the outbreak of these pests and diseases. Forecasting system can help in this direction.

In pests and diseases forewarning system, the variables of interest may be maximum pest population / disease severity, pest population / disease severity at most damaging stage of the crop, pest population / disease severity at different stages of crop growth or at various standard weeks, time of first appearance of pests / diseases, time of maximum pest population / disease severity, time of pest population / disease severity crossing threshold limit, extent of damage, weekly monitoring of pests and diseases progress, occurrence / non-occurrence of pests and diseases. If data are available at periodic interval for 15-20 years, the detailed study can be carried out for different variables of interest. However, depending upon the data availability, different types of models can be utilized for developing forewarning system. The models could be of two types, 'Between year model' and 'Within year model'.

**Between year models**

These models are developed using previous years’ data. An assumption is made that the present year is a part of the composite population of the previous years and accordingly the relationships developed on the basis of previous years’ data will be applicable for the present year. The forecast for pests and diseases are obtained by substituting the current year data into the model developed upon the previous years. Various methods have been attempted when data are available in quantitative form. Some of the important techniques are discussed below :

**Thumb rule**

This approach is the most common and extensively used. It is a simple system which describes the forecasting of the pests and diseases based on past experience. For example for potato late blight, a day is favorable if
- the 5 day temperature average is < 25.5°C
- the total rainfall for the last 10 days is > 3.0 cm
- the minimum temperature on that day is > 7.2°C

When this situation arises, there is a possibility of potato late blight appearance.

**Regression Model**

The regression model taking pest / disease variable as dependent and suitable independent variables such as weather variables, crop stages, population of natural enemies/predators etc. is used. The form of the model is

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + e \]
where $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are regression coefficients, $X_1, X_2, \ldots, X_p$ are independent variables and $e$ is error term. These variables are used in original scale or on a suitable transformed scale such as cos, log, exponential etc. (Coakley et al 1985; Trivedi et al. 1999).

**Fuzzy regression**

In regression analysis, the unfitted errors between a regression model and observed data are generally assumed as observation error that is a random variable having a normal distribution, constant variance, and a zero mean. In fuzzy regression analysis, the same unfitted errors are viewed as the fuzziness. Fuzzy regression can be quite useful in estimating the relationship among variables where the availability data are imprecise and fuzzy.

Fuzzy regression analysis gives a fuzzy functional relationship between dependent and independent variables where vagueness is present in some form. There are three situations where the fuzzy analysis can be viewed viz. Crisp parameters and fuzzy data, Fuzzy parameters and crisp data and Fuzzy parameters and fuzzy data. Fuzzy regression method is based on minimizing fuzziness as an optimal criterion, which can be achieved by linear programming procedures.

**Growing Degree Day Approach**

This method is based on the assumption that the pest becomes inactive below a certain temperature known as base temperature. Growing degree day is worked out as

$$GDD = \sum (\text{mean temp.} - \text{base temp.})$$

GDD is used in the model as explanatory variable. This method requires proper knowledge of base temperature and initial time from which accumulation is to start.

**Model based on weather indices**

In this approach, using weekly and fortnightly weather variables suitable indices are worked out which are used as regressors in the model. The model is of the form

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

where

$$Z_{ij} = \sum_{w=n_1}^{n_2} r_{iw} X_{iw} \quad \text{&} \quad Z_{i'j} = \sum_{w=n_1}^{n_2} r_{i'w} X_{i'w} X_{iw}$$

$Y$ variable to forecast; $X_{iw}$ is value of $i^{th}$ weather variable in $w^{th}$ period; $r_{iw} / r_{i'w}$ is suitable weight given to $i^{th}$ weather variable / product of $i^{th}$ and $i'^{th}$ weather variable in $w^{th}$ period; $p$ is number of weather variables considered; $n_1$ and $n_2$ are the initial & final periods for which weather data were included in the model and $e$ is error term.

If information on favourable weather conditions is known, subjective weights based on this information can be used for constructing weather indices. In absence of such information correlation coefficients between $Y$ and respective weather variable/product of weather
variables can be used [Agrawal et al. (2004), Chattopadhyay et al. (2005-a), Chattopadhyay et al. (2005-b), Desai et al. (2004) and Dhar et al. (2007)]

**Principal component regression**

Forewarning models can be developed using the principal component technique as normally relevant weather variables are large in number and are expected to be highly correlated among themselves. Using the first few principal components of weather variables as independent variables forecast models can be developed.

**Discriminant function analysis**

Forewarning models of pests and diseases based on time series data on weather variables can be developed using the discriminant function analysis. For this analysis, a series of data for 25-30 years are required. Based on the pest and diseases variables, data can be divided into different groups – low, medium and high etc. and using weather data in these groups, linear or quadratic discriminant functions can be fitted which can be used to find discriminant scores. Considering these discriminant scores as independent variables and diseases / pest as a dependent variable, regression analysis can be performed. Johnson et al (1996) used discriminant analysis for forecasting potato late blight.

**Complex polynomial [Group Method of Data Handling (GMDH)]**

It provides complex polynomial in independent variables. It selects the structure of the model itself without prior information about relationship. Form of the model:

\[ Y = a + \sum_{i=1}^{m} b_i X_i + \sum_{i=1}^{m} \sum_{j=1}^{m} c_{ij} X_i X_j + \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{k=1}^{m} d_{ijk} X_i X_j X_k + \ldots \]

The technique involves fitting of quadratic equations for all pairs of independent variables and identifying a few best performers in terms of predictive ability (using appropriate statistics); converting entire set of independent variables (called zero generation variables) to new variables (first generation variables) which are obtained as predicted values from these selected quadratic equations (of zero generation variables). The process of fitting and identifying best quadratic equations is repeated using first generation variables and second generation variables are obtained. The whole process is repeated with every new generation of variables till appropriate model is obtained (using certain criteria). At final stage, one best quadratic equation is selected as the final model. (Bahuguna et al 1992; Trivedi et al 1999).

**Machine Learning Techniques**

Machine learning techniques offer many methodologies like decision tree induction algorithms, genetic algorithms, neural networks, rough sets, fuzzy sets as well as many hybridized strategies for the classification and prediction (Han and Kamber, 2001; Pujari, 2000; Komorowski et al., 1999; Witten and Frank, 1999). Decision tree induction represents a simple and powerful method of classification that generates a tree and a set of rules, representing the model of different classes, from a given dataset. Decision Tree (DT) is a flow chart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test and each leaf node represents the class. The top most node in a tree is the root node. For decision tree ID3 algorithm and its successor C4.5
algorithm by Quinlan (1993) are widely used. One of the strengths of decision trees compared to other methods of induction is the ease with which they can be used for numeric as well as nonnumeric domains. Another advantage of decision tree is that it can be easily mapped to rules. Artificial Neural Networks (ANNs) is another attractive tool under machine learning techniques for forecasting and classification purposes. ANNs are data driven self-adaptive methods in that there are few apriori assumptions about the models for problems under study. These learn from examples and capture subtle functional relationships among the data even if the underlying relationships are unknown or hard to describe. After learning from the available data, ANNs can often correctly infer the unseen part of a population even if data contains noisy information. As forecasting is performed via prediction of future behaviour (unseen part) from examples of past behaviour, it is an ideal application area for ANNs, at least in principle. (Agrawal et al. 2004; Dewolf et al. 1997, 2000; Kumar, et al. 2010). However, the technique requires a large data base.

**Deviation Method**

This method can be utilized when periodical data at different intervals during the crop season are available for only 5-6 years. The pest population at a given point of crop stage is assumed to be due to two reasons – natural cycle of the pest and weather. To identify the natural cycle, data at different intervals is averaged over years and a suitable model is fitted to these averaged data points. Then the entire data is converted as deviations from the predicted natural cycle. Appropriate model is fitted using these deviations as dependent and weather as independent variables. [Mehta et al.(2001)]

**Ordinal logistic model – model for qualitative data**

The timely control measures to prevent pest / disease outbreak can be taken even if the information on the extent of severity is not available but merely the epidemic status is accessible. This information could be obtained through modeling qualitative data. Such models have added advantage that these could be obtained even if the detailed and exact information on pest count / disease severity is not available but only the qualitative status such as epidemic or no epidemic / low, medium or high is known. Such a situation arises quite often in pest / disease data. In such cases, the data are classified as 0/1 (2 categories); 0,1,2 (three categories). The logistic regression is used for obtaining probabilities of different categories. For example, for two categories, the model is of the form:

\[ P(E = 1) = \frac{1}{1 + \exp(-z)} + e \]

where \( z \) is a function of weather variables.

Forecast / Prediction rule:

- If \( P \geq .5 \) more chance of occurrence of epidemic
- If \( P < .5 \) probability of occurrence of epidemic is minimum

Within year model

Sometimes, past data on pests and diseases are not available but the pests and diseases status at different points of time during the crop season are available for the current season only. In such situations, within years growth model can be used for forewarning maximum disease severity / pest population, provided there are 10-12 data points between time of first appearance of pest / disease and maximum or most damaging stage.

The methodology consists in fitting appropriate growth pattern to the pests and diseases data based on partial data and using this growth curve for forecasting the maximum value of variable of interest. A number of growth models such as logistic, Gompertz etc. can be used for this purpose (Agrawal et al. 2004). Prajnesu (1998) developed a non linear statistical model for describing the dynamic population growth.

References


Kumar, Amrender; Ramasubramanian, V. and Agrawal, Ranjana (2010). Neural network based forecast modeling crops. (IASRI Publication).


Sarkar, J. (2002). Forecasting rice and wheat yield over different meteorological sub–divisions of India using statistical models. 56th Annual Conference of ISAS.


