Study of Effect of Climatological Variables on Crop Yield Estimation using Multiple Linear Regression

Dr. T. M. V. Suryanarayana
Associate Professor
Department of Water Resources Engineering
Faculty of Technology and Engineering The Maharaja Sayajirao University of Baroda, Samiala-391410
Vadodara, India

Abstract—An attempt has been made to carry out the study of determining the predominant climatological variables in estimating the crop yield. The climatological data are collected for the period 1981- 2006 and correlated with yield of cotton in Vallabh Vidyanagar using Multiple Linear Regression. The Climatological variables considered are Maximum Temperature, Minimum Temperature, Relative Humidity, Wind Speed and Sunshine Hours. The multiple linear models have been developed, to study their impact in prediction of the crop yield. The study has been carried out with eight different combinations of the five independent variables considered, to correlate with the crop yield. In each combination, i.e 1 to 8, the whole data is divided into proportions for training and Validation, such as 70% and 30% & 60% and 40% respectively. The developed Multiple Linear Regression Models are evaluated based on the performance indices such as Root Mean Squared Error and Correlation Coefficient. Based on the evaluation, the models developed are found to perform better in 60%-40% proportion of the data considered for the Study. Therefore in this considered proportion of the dataset, the models developed are ranked based on the obtained R.M.S.E. and R. The results clearly show that the consideration of all the variables, yield the best model with minimum R.M.S.E. and maximum R, followed by the combinations considering Maximum Temperature, Minimum Temperature, Relative Humidity as dependent variables along with/without Wind Speed/Sunshine hours. Moreover excluding the Relative Humidity, and trying the combinations of Maximum Temperature, Minimum Temperature along with/without Wind Speed/Sunshine Hours yields the poor models with maximum R.M.S.E. amd Minimum R. Hence considering multiple linear regression models and the eight combinations studied, it reveals that the yield of a crop is very much dependent on maximum and minimum temperatures & relative humidity.

Key words: Climatological Data, Crop Yield, Multiple Linear Regression, R.M.S.E., Coefficient of Correlation

I. Introduction

Agrarian sector in India is facing rigorous problem to maximize the crop productivity. More than 60 percent of the crop still depends on monsoon rainfall. Recent developments in Information Technology for agriculture field has become an interesting research area to predict the crop yield. The problem of yield prediction is a major problem that remains to be solved based on available data. Data Mining techniques are the better choices for this purpose. Different Data Mining techniques are used and evaluated in agriculture for estimating the future year's crop production. A brief analysis of crop yield prediction is presented in [2] using Multiple Linear Regression (MLR) technique and Density based clustering technique for the selected region i.e. East Godavari district of Andhra Pradesh in India.

Agriculture is the backbone of Indian Economy. In India, majority of the farmers are not getting the expected crop yield due to several reasons. The agricultural yield is primarily depends on weather conditions. Rainfall conditions also influences the rice cultivation. In this context, the farmers necessarily requires a timely advice to predict the future crop productivity and an analysis is to be made in order to help the farmers to maximize the crop production in their crops.

Yield prediction is an important agricultural problem. Every farmer is interested in knowing, how much yield he is about expect.

Having an estimate of final yield early in the growing season can be a powerful management and economic tool for the farming community. Therefore the possibility of using temporarily high resolution remote sensing data in combination with daily meteorological data for crop yield prediction on a close to field scale has been investigated in [3] for one of the main cropping areas in south−eastern Australia.

Efficient cropping requires yield estimation for each involved crop, where data-driven models are commonly applied. In recent years, some data-driven modeling technique comparisons have been made, looking for the best model to yield prediction. However, attributes are usually selected based on expertise assessment or in dimensionality reduction algorithms. A fairer comparison should include the best subset of features for each regression technique; an evaluation including several crops is preferred. The most common data-driven modeling techniques applied to yield prediction, using a complete method to define the best attribute subset for each model were evaluated in [1].

Neural networks have been gaining a great deal of importance and are used in the areas of prediction and classification; the areas where regression and other statistical models are traditionally being used. In this paper, a comprehensive review of literature comparing feed forward neural networks and traditional statistical methods viz. linear regression with respect to prediction of agricultural crop production has been carried out in [4]. Their study presents an useful insight into the capabilities of neural networks and their statistical counterparts used in the area of prediction of crop yield.

II. Study Area and Data

The entire Gujarat is divided into the various agroclimatic zones. Vallabh Vidyanagar is located in the Anand district and lies in middle Gujarat agro-climatic zone III of Gujarat state. Vallabh Vidyanagar is located at 22˚32’ N latitude, 72˚54’ E longitude at an altitude of 34 m above mean sea level. It is bounded on the north by the Kheda district and south by the Gulf of Khambhat, on the west by Ahmedabad district and, on the east by Vadodara district. The climate of Vallabh Vidyanagar is semi-arid with fairly dry and hot summer. Winter is fairly cold and sets in, in the month of November and continues till the middle of February. Summer is hot and dry which commences from mid of February and ends by the month of June. May is the hottest month with mean maximum temperature around 40.08 ˚C. The average rainfall is 853 mm.

The data required for evaluation in this study are collected from India Meteorological Department, Pune and Krishi Bhavan, Gandhinagar. Long term climatological daily data are collected from IMD (Indian Meteorological Department), Pune. The basic climatological data used comprises of Maximum and minimum temperature(˚C), Relative humidity (%), Wind speed (Kmph) and Sunshine hours (hour). The yield data are collected from the Krishi bhavan, Gandhinagar from year 1981-2006.

III. Methodology

Regression is one of the most powerful tools of statistics. It is used for the estimation of the strength of the relationship between variables. It refers to the method by which estimates are made of the values of one variable from knowledge of the values of one or more other variables.

"Multiple regressions" is a technique that allows additional factors to enter the analysis separately so that the effect of each can be estimated. It is valuable for quantifying the impact of various simultaneous influences upon a single dependent variable. Further, because of omitted variables bias with simple regression, multiple regression are often essential even when the investigator is only interested in the effects of one of the independent variables.

Multiple linear regression analysis is used to find the degree of inter-relationship among three or more variables. The least square regression analysis can be extended to cases where there are more than one independent variables. Let y the dependent variable and x_1, $x_2 \ldots x_p$ be the independent variables. The multiple linear regression equation is then written as

$$y = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + a_5x_5 \qquad [1]$$

Where a_0, $a_1 \ldots$, a_p are the regression constants (coefficients) to be determined.

An attempt has been made to carry out the study of determining the predominant climatological variables in estimating the crop yield. The Climatological variables such as Maximum Temperature, Minimum Temperature, Relative Humidity, Wind Speed and Sunshine Hours are considered as the independent variables and the crop yield is considered as the dependent variable. The multiple linear regression models have been developed, to study their impact in prediction of the crop yield. The study has been carried out with eight different combinations of the five independent variables considered, to correlate with the crop yield. The process of making combinations are as follows: First only two independent variables, i.e. maximum and mimimum temperatures are considered(Combination I), then three variables are considered by including relative humidity/Wind Speed/Sunshine hours to Combination I(Combination II/III/IV respectively) and then Four Variables are considered adding Wind Speed/Sunshine Hours to Combination II(Combination V/VI respectively) and adding Sunshine Hours to the Combination III(Combination VII) and Combination VIII Considering all the five Independent Variables. The same is given in Table 1.

In each combination, the data are divided into 70% & 30% and 60% and 40% of the data length, wherein the first proportion is used for model development and the second proportion is used for model validation. Similarly, eight combinations with two different proportions under each combination, total sixteen models have been developed. Then eight models under 70%-30% proportion of data length and another eight models with another proportion are evaluated based on the performance indices as given in Table 1.

Combination	Maximum Temperature(T1)	Minimum Temperature(T2)	Relative Humidity(R)	Wind Speed(W)	Sunshine Hours(S)
I	√	√			
II	√	√	√		
III	√	√		√	
IV	√	√			√
V	√	√	√	√	
VI	√	√	√		√
VII	√	√		√	√
VIII	√	√	√	√	√

Table 1: Combinations Considered among Independent Variables

IV. Model Evaluation Criteria

The performances of the developed MLR and MNLR models were compared using statistical evaluation performance indices, namely the relative error (RE), the correlation of coefficient (r), root mean square error (RMSE). The values of these performance indices were computed from the observed and model predicted values of the dependent variable. They were calculated for both the development and validation data sets. The values were calculated using the equations given in below table.

Root mean square error (RMSE):	$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y(i) - \hat{y}(i))^2}{n}}$
Correlation coefficient(r)	$\frac{\sum_{i=1}^{n}(y(i) - \bar{y})(\hat{y}(i) - \tilde{y})}{\sqrt{\sum_{i=1}^{n}(y(i) - \bar{y})^2 \sum_{i=1}^{n}(\hat{y}(i) - \tilde{y})^2}}$

Table 2: Performance Indices

V. Results and Analysis

The multiple linear regression models developed for all the eight combinations for 70%-30% and 60%-40% datasets are given in Table 2 and 3.

Combination	Multiple Linear Model
I	y= 21.25276923 x_1 - 71.36695332 x_2 + 2489.066166
II	y= -9.34372425 x1 - 184.2450693 x2 - 19.83068736 x3 + 6337.037524
III	y= 23.94286985 x1 - 106.4382631 x2 + 88.6177832 x3 + 2656.165828

IV	y= 24.14327695 x1 - 73.47970518 x2 - 11.55936938 x3 + 2546.99996
V	y= -50.51233331 x1 - 188.9907215 x2 - 40.81215096 x3 - 289.7791556 x4 + 9861.892926
VI	y= 3.733393732 x1 - 198.4132943 x2 - 20.50662765 x3 - 56.46706645 x4 + 6751.202137
VII	y= 27.93831644 x1 - 109.5617692 x2 + 89.1615189 x3 - 15.91209964 x4 + 2736.940121
VIII	y= -32.35307861 x1 - 212.3082115 x2 - 43.45678073 x3 - 311.1714393 x4 - 91.53494942 x5 + 10793.48161

Table 3: Multiple Linear Regression Models Developed for 70%-30% Dataset

Combination	Multiple Linear Model
I	y= -38.60959463 x1 - 67.58791521 x2 + 4220.040652
II	y= -13.64757734 x1 - 188.1131801 x2 - 23.08784047 x3 + 6705.182476
III	y= 19.23900333 x1 - 108.7358049 x2 + 141.049317 x3 + 2647.455594
IV	y= -52.92151012 x1 - 64.18448255 x2 + 26.46540102 x3 + 4340.374034
V	y= -144.254881 x1 - 191.9254217 x2 - 44.81514657 x3 - 375.7302153 x4 + 13232.96437
VI	y= 95.73234242 x1 - 250.2954181 x2 - 30.37575141 x3 - 187.693181 x4 + 6636.235811
VII	y= 102.4781955 x1 - 141.8452723 x2 + 209.6131572 x3 - 101.9254667 x4 + 1419.587966
VIII	y= -61.73402042 x1 - 224.6785312 x2 - 44.79379074 x3 - 307.5759748 x4 - 100.9504778 x5 + 12011.79783

Table 4: Multiple Linear Regression Models Developed for 60%-40% Dataset

The R.M.S.E and R, obtained for 70%-30% and 60%-40% for the abovesaid eight combinations are as given in Table 5 and Table 6 and the same are illustrated in Fig. 1 and Fig. 2.

Combination	R.M.S.E.		Correlation Coefficient		Parameters Considered
	70%	30%	70%	30%	
I	187.05	174.07	0.22	0.87	**T1T2**
II	134.45	238.80	0.71	0.72	**T1T2R**
III	179.87	267.67	0.35	0.79	**T1T2W**
IV	186.97	176.50	0.22	0.86	**T1T2S**
V	93.86	203.67	0.87	0.63	**T1T2RW**
VI	131.71	252.22	0.73	0.73	**T1T2RS**
VII	179.70	926.57	0.35	0.24	**T1T2WS**
VIII	83.31	199.94	0.90	0.65	**T1T2RWS**

Table 5: Performance Indices for 70%-30%

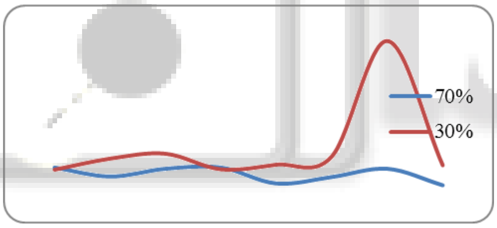

Fig. 1: R.M.S.E. for 70%-30%

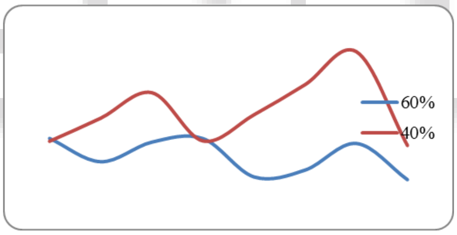

Fig. 2: R.M.S.E. for 60%-40%

Combination	R.M.S.E.		Correlation Coefficient		Parameters Considered
	60%	40%	60%	40%	
I	193.97	186.05	0.21	0.44	**T1T2**
II	130.73	248.62	0.75	0.72	**T1T2R**
III	183.80	317.79	0.38	0.57	**T1T2W**
IV	193.59	187.02	0.22	0.37	**T1T2S**
V	89.47	257.96	0.89	0.65	**T1T2RW**
VI	107.99	339.72	0.84	0.80	**T1T2RS**
VII	180.24	429.17	0.42	0.66	**T1T2WS**
VIII	82.05	174.80	0.91	0.75	**T1T2RWS**

Table 6: Performance Indices for 60%-40%

From the above given Tables 5 & 6 and Fig. 1 & 2, the models developed are found to perform better in 60%-40% proportion of the data considered for the Study. Therefore in this considered proportion of the dataset, the models developed are ranked based on the obtained R.M.S.E. and R and are given in Table 7.

COMBINATION	Parameters Considered	RANK
VIII	T1T2RWS	1
VI	T1T2RS	2
V	T1T2RW	3
II	T1T2R	4
III	T1T2W	5
VII	T1T2WS	6
I	T1T2	7
IV	T1T2S	8

Table 7: Rank based on R.M.S.E. and R

Looking to the Table 7, one can clearly say that the crop yield can be estimated by considering all the five climatological variables. But due to unavailability of some of the variables, if one determines to estimate the crop yield based on the climatological variables, then this table tabulated by the performance indices of the multiple linear regression models developed, gives the predominant variables and their importance. As the rank 1 to Rank 4 is allotted to Combination VIII, Combination VI, Combination V and Combination II, it clearly shows that the maximum and minimum Temperatures and Relative humidity along with/without other variables yields better models. Moreover, comparing the results obtained in Combination I and Combination IV, the R.M.S.E and R are almost same, and hence it may be understood that the inclusion or exclusion of sunshine hours will not have much impact on on estimation of crop yield, when only maximum and minimum temperatures are considered. At the same time, with the data considered for the study, any model developed with exclusion of relative humidity yields poor models. Hence, this shows that the yield of a crop is very much dependent on maximum and minimum temperatures & relative humidity.

VI. Conclusions

The results clearly show that the consideration of all the variables, yield the best model with minimum R.M.S.E. and maximum R, followed by the combinations considering Maximum Temperature, Minimum Temperature, Relative Humidity as dependent variables along with/without Wind Speed/Sunshine hours. Moreover, excluding the Relative Humidity, and trying the combinations of Maximum Temperature, Minimum Temperature along with/without Wind Speed/Sunshine Hours yields the poor models with maximum R.M.S.E. amd Minimum R. Hence, considering multiple linear regression models and the eight combinations studied, it reveals that the yield of a crop is very much dependent on maximum and minimum temperatures & relative humidity.

References

[1] Alberto Gonzalez-Sanchez, Juan Frausto-Solis, and Waldo Ojeda-Bustamante, “Attribute Selection Impact on Linear and Nonlinear Regression Models for Crop Yield Prediction”, The Scientific World Journal, Review Article, Volume 2014 (2014), Article ID 509429, 10 pages, http : // dx . doi . org / 10 . 1155/ 2014 / 509429.

[2] D. Ramesh and B. Vishnu Vardhan, “Analysis of crop yield prediction using data mining techniques”, IJRET: International Journal of Research in Engineering and Technology eISSN: 2319-1163 | pISSN: 2321-7308, Volume: 04 Issue: 01 | Jan-2015.

[3] Edgar AIGNER, Isabel COPPA and Friedrich WIENEKE, “Crop Yield Estimation Using NOAA − AVHRR Data and Meteorological Data in the Eastern Wimmera (South Eastern Australia)”, International Archives of Photogrammetry and Remote Sensing. Vol. XXXIII, Part B7. Amsterdam, 2000.

[4] Raju Prasad Paswan and Shahin Ara Begum, “Regression and Neural Networks Models for Prediction of Crop Production”, International Journal of Scientific & Engineering Research, Volume 4, Issue 9, September 2013 98 ISSN 2229-5518 IJSER © 2013 http://www.ijser.org.