DU Journal of Undergraduate Research and Innovation

# A Study of Crop Yield pattern with Climate Change based on Physical Parameters: Temperature and Rainfall in Western Uttar Pradesh to make future predictions for better Crop Management and Yield

Aditya Saxena*[1], Surendra Pratap Singh[1], Manju Rani[1], Manoj Kumar[2], Vishal Parmar[3], Ayushi Singh[3], Ankita Tiwari[3], Pooja[3], Shubham Yadav[3], Gaurav Kumar[3], Monika[3], Siddhant Shekhar[3], Abhishek Gogna[3]
*adityasaxena123@gmail.com
[1]Department of Physics, Deshbandhu College, Delhi University,
[2] Department of Mathematics, Deshbandhu College, Delhi University
[3]Student, Department of Physics, Deshbandhu College, Delhi University

## ABSTRACT

Study to predict wheat yield of ten districts of western Uttar Pradesh based on a suitable statistical model using rainfall and temperature as input parameters has been done for the first time. Using the Multiple Linear Regression (MLR) model, regression equations have been obtained for all the ten districts to make wheat yield predictions. Correlation analysis between rainfall and temperature shows high correlation ~ 95%. Thus using a single parameter also provides good prediction, the prediction lying mostly below 10% of error and 16% at the maximum

Keywords: Wheat Yield, Regression, correlation, Forecasting Techniques, Weather based crop forecasting.

## INTRODUCTION

Wheat, belonging to the family Tritium is one of the most important cereal grains with respect to humans. It is believed to have been grown as far as back as 7500-7300 BC in regions of what is modern day Jordan and south-eastern Turkey and was one of the first plants to be domesticated and grown primarily for consumption. Wheat has the special ability to self-pollinate, which greatly encouraged its selection over many other plants. Common Wheat, or *Tritium aestivum* is the most widely cultivated species of wheat all over

the world, however a few other species of wheat are also cultivated. The aforementioned ability of wheat to self-pollinate has resulted in difficulty in creating hybrid varieties of wheat despite nearly 90 years of effort [1].

100 g of wheat contains about 12.6 g of protein, 1.5 g of total fat, 71 g of carbohydrate (by difference), and 12.2 g of dietary fiber [2]. Wheat protein, Gluten, is an important economic product. Hence wheat is a significant crop. In order to grow wheat, moderate to low temperature and rainfall are preferred. The crop is planted in the winters during the month of October and harvested around May. It requires lot of sunlight, especially when the grains are being sown and low humidity is preferred as most of the diseases contracted by the crop thrive in warm, moist climates [3].

Forecasting of crop yield is highly dependent on the physical parameters such as temperature, rainfall, greenhouse gas emission, humidity etc. The area of our study are 10 major districts of Western Uttar Pradesh that constitutes 0.0925% of the area in comparison to the country and contribute 6.7% of the total wheat produced in the country making it highly reliable area for this crop to grow. The reason behind this is the availability of the optimum temperature range of 10-25°C and average rainfall of 80 cm between the months of October and March which is ideal for germination of wheat seed.

Correlation & Regression analysis using Multiple Linear Regression (MLR) techniques are employed for forecasting the yield of wheat. It includes weather indices based on the data available in public domain from research institutes such as Indian Agricultural Research Institute (IARI), Indian Meteorological Department (IMD) and Indian Agricultural Statistical Research Institute (IASRI). This data was thus analyzed to study the yield patterns and consequently make predictions based on that. For the first time regression equations have been obtained using MLR model for 10 districts of western Uttar Pradesh, which include Saharanpur, Muzaffarnagar, Meerut, Moradabad, Budaun, Ghaziabad, Bulandshar, Bijnor, Baghpat and Gautam Budh Nagar. This paper discusses detailed study of the district Ghaziabad.

## METHODOLOGY

Forecasting is a technique to predict future trends using statistical modeling. There are several methods of forecasting, like quantitative and qualitative methods, naïve approach, economic forecasting methods etc. Each method is appropriate for certain type of situation only thus making a right decision in selecting a method is most important step in forecasting.

Given a sequence of data $(x_1, x_2 \ldots x_n)$ we define the mean to be $(x_1 + \cdots x_n)/N$ and denote it by

$$\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n \qquad (1)$$

This mean is average value of data. Sometimes several data sets can have same mean but have different variation about mean. This brings the concept of variance. The variance of $(x_1 + \cdots x_n)$ denoted by $\sigma_x^2$ is

$$\sigma_x^2 = \frac{1}{N}\sum_{n=1}^{N}(x_i - \bar{x})^2 \qquad (2)$$

The standard deviation $\sigma_x$is the square root of the variance:

$$\sigma_x = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(x_i - \bar{x})^2} \qquad (3)$$

Note that if the x's have units of meter then the variance $\sigma_x^2$ has units of $\text{meter}^2$, and the standard deviation $\sigma_x$ and the mean x have units of meter. Thus it is the standard deviation that gives a good measure of the deviations of the x's around their mean.
Covariance is a measure of how much two random variables change together. If the greater values of one variable mainly correspond with the greater values of the other variable, and the same holds for the smaller values, i.e., the variables tend to show similar behavior, the covariance is positive. In the opposite case, when the greater values of one variable mainly correspond to the smaller values of the other, i.e., the variables tend to show opposite behavior, the covariance is negative. The sign of the covariance therefore shows the tendency in linear relationship between the variables.
The sample covariance of *N* observations of two variables is a *2*-by-*2* matrix with the entries

$$q_{xy} = \frac{1}{N-1}\sum_{n=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) \quad (4)$$

Thus

$$q_{xx} = \frac{1}{N-1}\sum_{n=1}^{N}(x_i - \bar{x})(x_i - \bar{x}) \quad (5)$$

$$q_{yy} = \frac{1}{N-1}\sum_{n=1}^{N}(y_i - \bar{y})(y_i - \bar{y}) \quad (6)$$

For any forecasting process primary step is to observe the relationship between variables, which is given by correlation analysis.
*Correlation coefficient*: Correlation coefficients (denoted by *R*) quantify the relation between X and Y in unit-free terms. When all points of a scatter plot fall directly on a line with an upward incline, *R* = +1; when all points fall directly on a downward incline, *R* = -1. Such perfect correlation is seldom encountered. We still need to measure *correlational strength*, defined as the *degree* to which data point adhere to an imaginary trend line passing through the "scatter cloud." Strong correlations are associated with scatter clouds that adhere closely to the imaginary trend line. Weak correlations are associated with scatter clouds that adhere marginally to the trend line. The closer *R* is to +1, stronger the positive correlation and the closer *R* is to -1, stronger the negative correlation. Correlation coefficient is given by

$$R_{xy} = \frac{q_{xy}}{\sqrt{q_{xx}}\sqrt{q_{yy}}} \quad (7)$$

After correlation analysis we have to proceed to regression analysis, which helps us to establish a linear relationship between available variables. Usually, the investigator seeks to ascertain the causal effect of one variable upon another for example effect of global warming on crop yield. To explore such issues, the investigator assembles data on the underlying variables of interest and employs regression analysis to estimate the quantitative effect of the causal variables upon the variable that they influence. For estimating wheat yield let us restrict ourselves to one factor only say temperature. Regression analysis with single dependent factor is called simple regression analysis [4]. Relationship between these variable in a mathematical form is given as

$$y = \beta_0 + \beta_1 X + \varepsilon \quad (8)$$

Where $y$ = wheat yield. $\beta_0$= intercept on y-axis
$\beta_1$ = slope, X= temperature
$\varepsilon$ = random error.

Where $\beta_0 \; and \; \beta_1$ are given by using method of least square as

$$\begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix} = \begin{pmatrix}\sum_{n=1}^{N} x_n^2 & \sum_{n=1}^{N} x_n\\ \sum_{n=1}^{N} x_n & \sum_{n=1}^{N} 1\end{pmatrix}^{-1} \begin{pmatrix}\sum_{n=1}^{N} x_n y_n\\ \sum_{n=1}^{N} y_n\end{pmatrix} \quad (9)$$

We have two variables rainfall and temperature, which affect wheat yield. When there are more than one predictor variable available, that leads to the following "multiple regression" mean function:

$$\mathrm{E(Y\,|X)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \ldots\ldots\ldots\ldots \beta_n x_n + \varepsilon \quad (10)$$

Where $\beta_0$ is called the intercept and $\beta_j$ are called coefficients.
One can represent all response values for all observations by n-dimensional vector called the response vector.

$$y = \begin{Bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{Bmatrix} \quad (11)$$

One can denote all predictors by a (n × p + 1) matrix called the design matrix:

$$X = \begin{Bmatrix} 1 & X_{11} & \ldots\ldots\ldots\ldots X_{1p} \\ 1 & X_{12} & \ldots\ldots\ldots\ldots X_{2p} \\ & & \vdots \\ 1 & X_{1n} & \ldots\ldots\ldots\ldots X_{np} \end{Bmatrix} \quad (12)$$

One represents the intercepts and slopes by a p+1-dimensional vector called the slope vector, denoted by

$$\beta = \begin{Bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{Bmatrix} \quad (13)$$

Finally, one denotes all error terms by a n-dimensional vector called the error vector

$$\varepsilon = \begin{Bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{Bmatrix} \quad (14)$$

Using linear algebra notation, the model given by Equation (10) can be rewritten as:

$$Y = X\beta + \varepsilon \quad (15)$$

Where $X\beta$ is the matrix-vector product.

In order to estimate *Y* we take a least squares approach that is analogous to what was done in the simple linear regression case i.e. minimize overall value of intercept and slope given below.

$$\sum_i \left((y_i - \beta_0 + \beta_1 x_i)^2 \dots\dots\dots\dots\dots (y_n - \beta_0 + \beta_p x_{ip})^2\right) \quad (16)$$

It is minimized by setting

$$\hat{\beta} = (x'x)^{-1}x'y \quad (17)$$

And the fitted value comes out to be $\hat{Y} = \hat{\beta}X$ [4].

The above methods are very crude methods of forecasting. Weather indices based Multiple Linear Regression (MLR) model, which creates suitable transformation to construct weather indices used as prediction variable, is of the same form as IARI model [5-7].

$$y = b_0 + \sum_{i=1}^{m} b_i Z_i + \sum_{i=1}^{m} b_{ij} Z_{ij} + \varepsilon \quad (18)$$

*j=0, 1, 2.*

Where $\varepsilon$ denote random error, $y$ denotes yield, $b_0, b_i, b_{ij}$ are the regression coefficients, and $Z_i, Z_{ij}$ are the independent variable which are function of basic weather variables like temperature and rainfall. The index '$m$' represents particular month of the season and 'j' is the degree of equation. Following method was used to calculate the effect of temperature and rainfall for Ghaziabad district of western Uttar Pradesh.

1-To study the individual effect of parameters the generated variables are given by [8]

$$Z_{ij} = \sum_{m=1}^{n} x_{im} R_{im}^{j} / \sum_{m=1}^{n} R_{im}^{j} \quad (19)$$

*j=0,1.*

For j = 0, we have un-weighted generated variable

$$Z_{i0} = \sum_{m=1}^{n} x_{im}/n \quad (20)$$

and weighted generated variables

$$Z_{i1} = \sum_{m=1}^{n} x_{im} R / \sum_{m=1}^{n} R_{im}^{1} \quad (21)$$

And the model becomes-

$$yield = \alpha + \beta_1 Z_{i0} + \beta_2 Z_{i1} + \varepsilon \quad (22)$$

Where $x_{im}$ is the value of $i^{th}$ (i = 1,2,…n) weather variable at $m^{th}$month (m = 1,2,…..n). In this study m is 6 i.e. (October to March). $r_{im}$ is the simple correlation coefficient between weather variable $x_i$ at m$^{th}$ month and crop yield over a period of K years. $\alpha, \beta_1, and\ \beta_2$ are parameters of the model to be evaluated for the effect of variables and ε is error term supposed to be normal distribution with mean zero and variance $\sigma^2$.

2- To study the joint effect of parameters the generated variable is given by [8]

$$Q_{ii',j} = \sum_{m=1}^{n} R_{ii'm}^{j} x_{im} x_{i'm} / \sum_{m=1}^{n} R_{ii'm}^{j} \quad (23)$$

*j=0,1,2*

Where is $R_{ii'm}^{j}$ the correlation coefficient between crop yield $y$ and product of weather variables $x_{im}$and. $x_{i'm}$. Clearly, we have two generated variables (interaction term)

Again un-weighted term will be

$$Q_{ii',0} = x_{i'm} x_{im} / n \quad (24)$$

And the weighted term will be

$$Q_{ii',1} = \sum_{m=1}^{n} R_{ii'm}^{1} x_{im} x_{i'm} / \sum_{m=1}^{n} R_{ii'm}^{1} \quad (25)$$

Including these two interaction terms in the model we finally have an equation of the form

$$Y = a + \sum_{i=1}^{2} \sum_{j=0}^{2} b_{ij} Z_{ij} + \sum_{j=0}^{1} b_{ii',j} Q_{ii',j} + \varepsilon \quad (26)$$

Where $b_{ij}$and $b_{ii',j}$ are parameters (regression coefficients) of the model, and other terms have already been explained in previous model. In simple form it will take the form as following.

$$yield = \alpha + \beta_1 Z_{0r} + \beta_2 Z_{1r} + \beta_3 Z_{0t'} + \beta_4 Z_{1t'} + \beta_5 Q_0 + \beta_6 Q_1 + \beta_7 Q_2 + \varepsilon \quad (27)$$

Where -$Z_{0r}, Z_{1r}$ are generated variable for rainfall
$Z_{0t'}, Z_{1t'}$ are generated variable for average temperature.
$Q_0$, $Q_1$ and $Q_2$ are generated variable for joint effect of both the parameters.
$\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7$ and $\alpha$ are regression coefficients.

## RESULTS

As a very preliminary analysis we tried to use actual values of rainfall and temperature to establish a relationship between these two variables and wheat yield calling it Model 1. Below are the scatter diagram of average value of rainfall and temperature during October-March and wheat yield over 20 years. The graph clearly indicates week negative correlation with r squared value merely 0.1573 and 0.123 respectively for average temperature and rainfall.

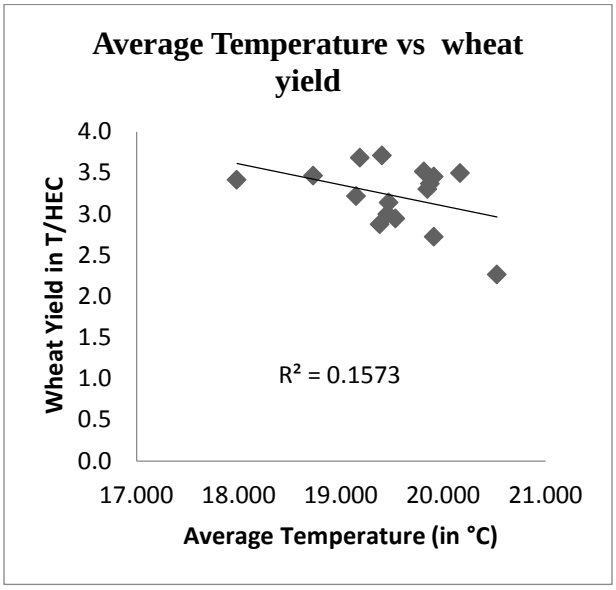


Figure-I Scatter diagram of wheat yield with average temperature

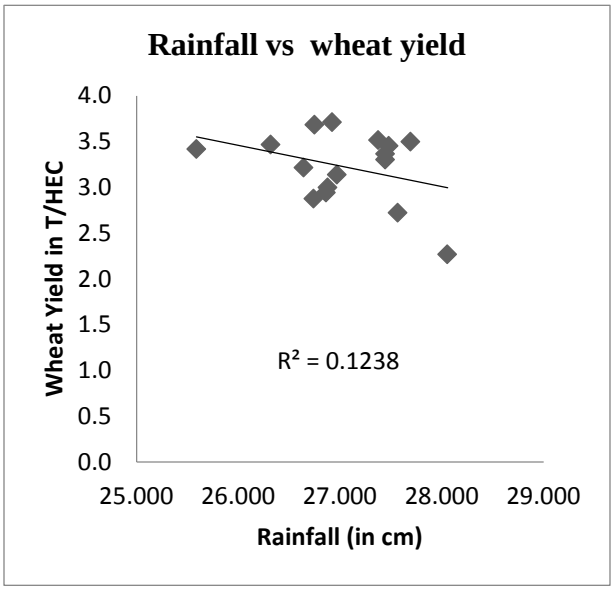


Figure-II Scatter diagram of wheat yield with rainfall

Regression analysis of this model for the year 2001-2002 is as follows.

Table-I

| *Regression Statistics* | |
|---|---|
| $R$ | 0.554236243 |
| $R^2$ | 0.307177813 |
| Adjusted $R^2$ | 0.191707448 |
| Standard Error | 0.350496148 |
| Observations | 15 |

Where $R$ is correlation coefficient.

Table-II

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
|---|---|---|---|---|
| Intercept | 0.490437774 | 8.349219075 | 0.058741 | 0.954126 |
| AVG TEMP | -1.484927196 | 1.029444209 | -1.44246 | 0.174763 |
| RAINFALL | 1.171515218 | 1.020809976 | 1.147633 | 0.273486 |

Table-I shows R squared value 0.30717, which is not a good value though residual graphs of this model are random in nature [9].

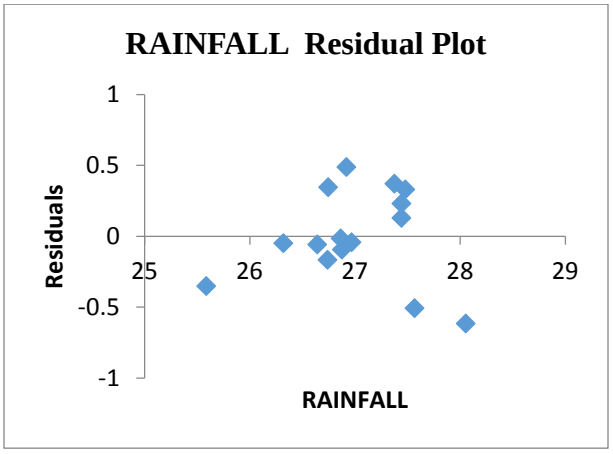


Figure-III Residual plot of rainfall using regression analysis

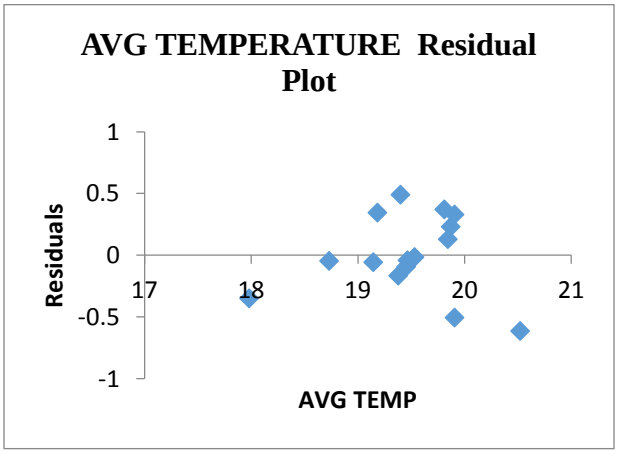


Figure-IV Residual plot of average temperature using regression analysis

Results for the year 2000-2001 and 2001-2002 are

Table-III

| | YEAR | |
|---|---|---|
| | 2001-02 | 2000-01 |
| PREDITED YIELD | 2.987996 | 3.198839 |

| | | |
|---|---|---|
| ACTUAL YIELD | 3.495891 | 3.45197 |
| % ERROR | -14.5283 | -7.33294 |

Yield in tonnes/hectare

During different stages of plant growth, it requires different set of weather condition [10]. Thus during entire cycle of plant growth some months play a major role in final yield. Below are the correlation coefficients of wheat yield from 1986-2002 and rainfall and temperature in the wheat cycle months i.e. from October to March for the same period.

Table-IV

| Correlation coefficients of wheat yield and rainfall, temperature in wheat cycle months | | | | | | |
|---|---|---|---|---|---|---|
| | October | November | December | January | February | March |
| RAINFALL | -0.28299 | -0.325671 | 0.0134623 | -0.520773 | -0.452741 | -0.0858890 |
| TEMPERATURE | 0.026553 | 0.2311165 | 0.0160932 | -0.578450 | -0.437371 | 0.0073769 |

From above table it is evident that months of January and February has greatest correlation coefficient both for rainfall and temperature. Thus we can say that these two months play a major role in total wheat yield.

The second model we used is weather indices based MLR model. Below are the results of this model for the year 2001-2002 and 2000-2001.

Table-V

| REGRESSION STATICS | | |
|---|---|---|
| | YEAR | |
| | 2000-2001 | 2001-2002 |
| $R$ | 0.808211 | 0.801138 |
| $R^2$ | 0.653205 | 0.641822 |
| Adjusted $R^2$ | 0.248611 | 0.283644 |
| Standard Error | 0.345213 | 0.329962 |
| Observations | 14 | 15 |

Where $R$ is correlation coefficient.

Table-VI

| COEFFICIENTS | | |
|---|---|---|
| | 2000-2001 | 2001-2002 |
| Intercept | -76.7176 | -158.557 |

| | | |
|---|---|---|
| $Z_{0r}$ | 5.246067 | 6.348724 |
| $Z_{1r}$ | -0.328 | 1.925953 |
| $Z_{0t}$ | -6.11169 | -10.9337 |
| $Z_{1t}$ | 6.966811 | 15.80177 |
| q0 | -0.06207 | -0.00129 |
| q1 | 0.170723 | 0.003977 |
| q2 | -0.24485 | -0.30992 |

R squared value in Table 5 is quite high for both the years i.e. 0.641 which proves this model to be better and more significant as compared to model 1. This is also visible from the randomness of residual graphs of this model [9].

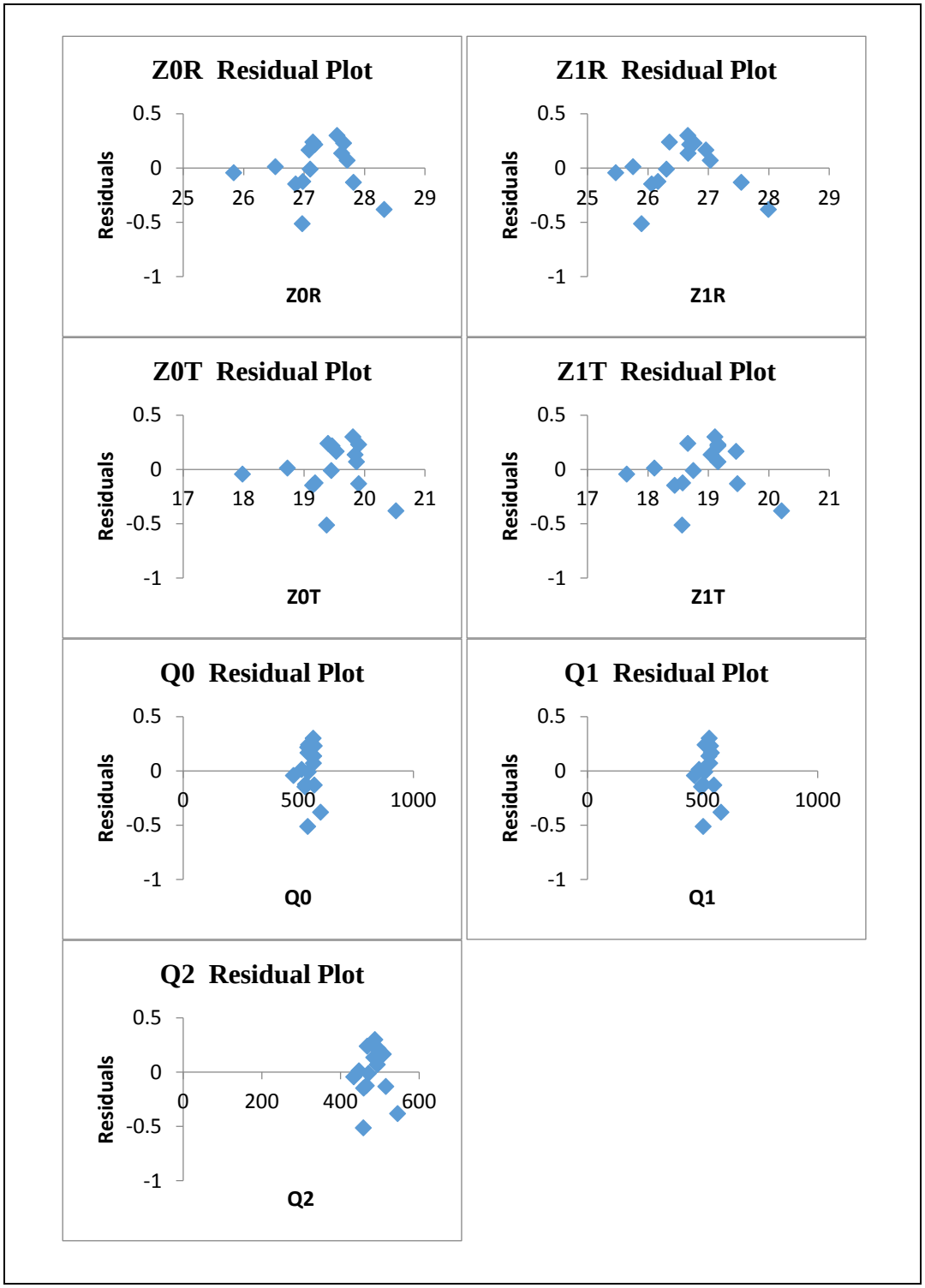


Figure-V Residual plots for the results tabulated in Table-VI

Results for this model are as follows.

Table-VII

| | YEAR | |
|---|---|---|
| | 2000-2001 | 2001-2002 |
| ACTUALL YIELD | 3.45197 | 3.49589099 |
| CALCULATED YIELD | 3.0127294 | 3.3008154 |
| % ERROR | -12.72434 | -5.5801405 |

Yield in tonnes/hectare

## DISCUSSION

From the results it is found that a reasonably good prediction can be made even with very crude and primary methods of forecasting. Data used was monthly average data, which might not be as good for large area prediction but for small area where we can assume that environment condition remain same; these methods can be effectively used. Model 2 (weather indices based MLR model) is quite efficient as it is more significant in terms of R square and more accurate.

## CONCLUSION

Monthly data on rainfall and temperature was received from Indian Meteorological Department for 10 district of western Uttar Pradesh. Computations were done for respective district and it was found that never before have wheat yield calculations been done for these areas. In every district correlation of rainfall (in centimeter) and temperature (in degree centigrade) with wheat yield of a particular year was maximum in the month of January, which clearly indicates that these two factors in the month of January have most significant effect on the total wheat yield of the respective year. Since these values are negative for Ghaziabad (-0.52077 and -0.57845 for rainfall and temperature respectively) they have negative relation with wheat yield. Results were found to have accuracy of 90% thus making it a very good model for crop prediction of small area. Since it is a very crude model requiring only one parameter for of regression analysis, one can easily make computations using this model of regression analysis for making predictions of yield of various crops.

## ACKNOWLEDGEMENT

We are grateful to Dr. Ajay Kumar Arora, Principal, Deshbandhu College for providing us with necessary infrastructure and facilities for carrying out the project work.

We gratefully acknowledge the financial support provided by University of Delhi for carrying out this project. Vishal Parrmar, Ayushi Singh, Ankita Tiwari, Pooja, Ayushi Singh, Shubham Yadav, Gaurav Kumar, Monika, Siddhant Shekhar, Abhishek Gogna, are grateful to University of Delhi for providing Fellowship in the form of student stipend during the course of this project.

References

[1] http://www.hybridwheat.net/anglais/growing-hybrid-wheat-in-europe/history-of-hybrid-wheat/history-of-hybrid-wheat-627.aspx
[2] USDA National Nutrient Database for Standard Reference
[3] Indian Institute of Wheat & Barley Research. http://www.dwr.res.in/node/80
[4] Panchenko, Dmitry. *18.443 Statistics for Applications, Fall 2006*. (MIT Open Course Ware: Massachusetts Institute of Technology), http://ocw.mit.edu/courses/mathematics/18-443-statistics-for-applications-fall-2006 (Accessed 25 Dec, 2014). License: Creative Commons BY-NC-SA)
[5] *Fisher, R. A. "Statistical Method for Research Works", Published by Oliver and Boyd (1925). Fisher, R.A. "The influence of rainfall on the yield of wheat at Rothamsted",Roy. Soc. London, Phil. Trans.,* ***B, 213****, 89-142 (1924).*
[6] Hendricks, W. A. and Scholl,J. C."Techniques in measuring joint relationships-The joint effects of temperature and precipitation on corn yields",N. Carolina. Agric. Exp. Sta. Tech. Bul.**74** (1943).
[7] Amender Kumar and Ramasubramanian V "Crop forecasting based on meteroligical data using SAS". Chapter 7, Data Analysis in Social Sciences Research using SAS-Reference Manual I.A.R.I. Library Avenue, Pusa, New Delhi www.iasri.res.in/sscnars/content_social.htm.
[8] Agrawal, R., Jain, R.C. and Jha, M.P. "Models for studying rice crop weather relationship", *Mausam*, **37 (1)**, 67-70 (1986). Agrawal, R., Jain, R.C. and Jha, M.P. "Models for studying rice crop weather relationship", *Mausam*, **37 (1)**, 67-70 (1986).
[9] Jaime Wisniak and Anna Polishuk "Analysis of residuals — a useful tool for phase equilibrium data analysis" Fluid Phase Equilibria 164 _1999. 61–82
[10] http://www.uky.edu/Ag/GrainCrops/ID125Section2.html